Modélisation stochastique et estimation de la dispersion du pollen de maïs.Estimation dans des modèles à volatilité stochastique.

Agnès Grimaud

▶ **To cite this version:**

Agnès Grimaud. Modélisation stochastique et estimation de la dispersion du pollen de maïs.Estimation dans des modèles à volatilité stochastique.. Mathématiques [math]. Université Paris-Diderot - Paris VII, 2005. Français. NNT : . tel-00011584

HAL Id: tel-00011584
https://theses.hal.science/tel-00011584

Submitted on 10 Feb 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS 7 - DENIS DIDEROT
UFR de Mathématiques

THÈSE

pour l'obtention du Diplôme de

DOCTEUR DE L'UNIVERSITÉ PARIS 7

Spécialité : MATHÉMATIQUES APPLIQUÉES

présentée par
Agnès GRIMAUD

Titre :

MODÉLISATION STOCHASTIQUE ET ESTIMATION DE LA DISPERSION DU POLLEN DE MAÏS. ESTIMATION DANS DES MODÈLES À VOLATILITÉ STOCHASTIQUE.

Directrice de thèse : **Catherine LAREDO**

Soutenue publiquement le **05 Décembre 2005**, devant le jury composé de
Mme Laure ELIE, Université Paris 7
Mme Elisabeth GASSIAT, Université Paris 11
Mme Catherine LAREDO, INRA Unité MIA
Mme Dominique PICARD, Université Paris 7
M. Mats RUDEMO, Université de Göteborg
M. Jian-Feng YAO, Université Rennes 1

au vu des rapports de **M. Jean-Marc AZAÏS** (Université Toulouse 3) et de **M. Mats RUDEMO** (Université de Göteborg).

Remerciements

Je tiens tout d'abord à remercier Catherine Larédo qui a eu confiance en moi et m'a permis de mener à bien ce travail. Ses connaissances et ses remarques m'ont beaucoup apporté tout au long de ma thèse.

Je remercie Jean-Marc Azaïs et Mats Rudemo d'avoir accepté de consacrer de leur temps à rapporter ma thèse.

Je tiens aussi à remercier chaleureusement Laure Elie, Elisabeth Gassiat, Dominique Picard et Jian-Feng Yao d'avoir accepté d'être membre du jury. Je suis très touchée de leur présence aujourd'hui.

Merci à l'unité MIA de l'INRA de Jouy-en-Josas de m'avoir accueillie tout au long de ces années de thèse dans une ambiance chaleureuse. En particulier merci à Sylvie Huet et Hervé Monod qui m'ont permis de faire mes premiers pas dans le domaine de la recherche lors de mon stage de DEA. Et merci à Suzanne T. également.
Merci à mes informaticiens préférés : Eric M., Valérie R. et Annie B. sans qui je n'aurais pas pu mener à bien toute la partie applicative de mon travail.
Et merci aux thésards de l'unité, en particulier David L., Fanny V. et Valérie A. avec qui j'ai partagé de très bons moments.

Je remercie également Florent A. et Vincent R. pour leur oreille attentive et pour m'avoir soutenue dans les moments de doute.

Je voudrais également remercier les enseignants de Paris 7, avec qui j'ai effectué mon monitorat, en particulier Jacqueline Mac Aleese qui m'a fait partagé son goût pour l'enseignement. Merci à l'équipe de Probabilités, Statistique et Modélisation de Paris 11, Orsay, de m'avoir accueillie en tant qu'ATER.

Et merci à Patricia P. et à Michèle Wasse pour leur aide dans les démarches administratives.

Pour finir, je n'oublie pas ma famille, en particulier mes parents et Pascal, qui m'ont toujours soutenue et encouragée au cours de ces années.

Table des matières

This Thesis aims to study stochastic processes, applied on one hand to modelling and estimation of the corn pollen dispersion and on the other hand to estimation in stochastic volatility models. It is organized as following.

The first chapter is an introduction presenting the subject of this thesis and a synthesis of the obtained results.

The second chapter aim is to model and to estimate the corn pollen dispersion in an homogenous environment, i.e. two corn fields have no separation between them. The pollen grain path is modeled using diffusion processes and this leads to different individual parametric dispersion functions. Then the parameters associated to these functions are estimated and results are validated.
This chapter is also an english article which is to be submitted, and given in appendix at the end of the chapter.

The third chapter aim is to model and to estimate the corn pollen dispersion in an heterogeneous environment, i.e. two corn fields are separated by another culture or a nude ground. Different individual dispersion functions are proposed based on the obtained results in chapter II.

The chapter IV aims to estimate parameters in mean-reverting stochastic volatility models. For this, Whittle and Tapered Whittle estimators are used. Furthermore, a two step statistical method is proposed in which one of the parameters is estimated using an empirical moment method. Some simulations are done to compare performances of those estimators.

The chapter V aim is estimation for mean-reverting stochastic volatility models with a leverage effect. The parameters are then estimated using an indirect inference method. Simulations are also done in this case.

The Sixth and last chapter is composed of two appeared articles (one in french and one in english), resulting of the work during my DEA stage and begining of my thesis. It studies the control of homeologous pairing in Oilseed Rape Haploids using the mixture of two mixed model.

Cette thèse est consacrée à l'étude de processus stochastiques, appliqués d'une part à la modélisation et l'estimation de la dispersion du pollen de maïs et d'autre part à l'estimation dans des modèles à volatilité stochastique. Elle est organisée de la façon suivante.

Le chapitre I est un chapitre d'introduction présentant le sujet de cette thèse et une synthèse des résultats obtenus.

Le chapitre II est consacrée à la modélisation et l'estimation de la dispersion du pollen de maïs en milieu dit homogène, c'est-à-dire lorsque deux champs sont l'un à côté de l'autre. La trajectoire d'un grain de pollen est modélisée à l'aide de processus de diffusion et conduit à l'obtention de différentes fonctions de dispersion individuelles, paramétriques. Ensuite les paramètres de ces fonctions sont estimés et les résultats obtenus validés. Ce travail fait l'objet d'un article à soumettre qui se trouve en annexe à la fin de ce chapitre.

Le chapitre III porte sur la modélisation et l'estimation de la dispersion du pollen de maïs en milieu dit hétérogène, c'est-à-dire lorsqu'il y a une discontinuité dans le couvert végétal. Deux champs de maïs sont séparés par une autre culture ou un sol nu. Différentes fonctions de dispersion individuelles sont proposées basées sur les résultats obtenus au chapitre II.

Le chapitre IV est consacré à l'estimation paramétrique pour des modèles à volatilité stochastique "mean-reverting" à l'aide des estimateurs de Whittle et de Whittle raboté. De plus, une méthode statistique à deux pas est proposée où l'un des paramètres est estimé à l'aide d'une méthode de moments. Des simulations sont réalisées pour comparer les performances des estimateurs.

Le chapitre V porte sur l'estimation dans des modèles à volatilité stochastique "mean-reverting" avec un effet leverage. Les paramètres sont alors estimés à l'aide d'une méthode d'inférence indirecte. Des simulations sont à nouveau faites.

Le dernier chapitre est constitué de deux articles parus (l'un en français, l'autre en anglais) sur un travail commencé lors de mon stage de DEA. Il porte sur l'étude du déterminisme génétique des appariements de chromosomes chez des haploïdes de colza à l'aide d'un mélange de modèles mixtes.

Chapitre 1

Introduction

1.1 Partie I : Modélisation stochastique et estimation de la dispersion du pollen de maïs

Les trente dernières années ont vu se développer des techniques modernes de "génie génétique", appelées transgénèse, consistant à introduire un ou plusieurs gènes dans le patrimoine génétique d'un organisme, afin de lui conférer une caractéristique nouvelle, et à construire ainsi des organismes dits "génétiquement modifiés", notés OGM. Le premier transfert artificiel de gènes entre deux bactéries fut réalisé en 1973. Ces techniques peuvent être appliquées aussi bien sur des organismes animaux ou végétaux que sur des micro-organismes. La transformation permet ainsi, selon les cas, de modifier, supprimer ou introduire certains caractères. De plus, elle peut s'effectuer sur des individus dont toutes les cellules (reproductrices ou non) contiennent l'ADN étranger, ou transgène, et qui sont donc capables de le transmettre à leur descendance.

Les progrès apportés par l'utilisation des OGM

Le génie génétique a permis et permet de progresser dans différents domaines de la recherche :

* Dans le domaine médical, la thérapie génique a déjà été expérimentée pour des pathologies très diverses, du cancer aux maladies cardio-vasculaires, de la myopathie à la mucoviscidose. Le génie génétique permet également un usage nouveau des plantes afin d'obtenir des molécules à usage thérapeutique se substituant aux synthèses chimiques ou à l'extraction de substances issues d'organes humains ou animaux (insuline, hormone de croissance par exemple).

* Dans le domaine agricole, les modifications actuelles du génome des plantes

visent à protéger les cultures et à améliorer leurs caractéristiques agronomiques (rendement, adaptation à différentes conditions climatiques). En 2003, 67.7 millions d'hectares de plantes transgéniques étaient cultivés contre 40 millions en 1999. Les principales améliorations sont :

- La résistance des plantes aux insectes nuisibles aux cultures. Ceci permet d'éviter des traitements insecticides qui peuvent occasionner des pertes importantes de rendement et être nuisibles à l'environnement. Par exemple, un transgène a été développé pour combattre la pyrale, chenille destructrice des plants de maïs. En chine, après dix de recherche, la commercialisation du riz transgénique Bt est sur le point d'être autorisée. Ce riz contient le gène d'une bactérie du sol qui fabrique naturellement un poison contre plusieurs prédateurs de la céréale.

- La tolérance des plantes aux herbicides. Cela permet de répandre des herbicides sur les cultures pour agir sur les plantes sauvages indésirables, tout en étant assuré que la plante cultivée est protégée contre l'action de l'herbicide par l'introduction d'un "gène de tolérance" dans son génome, dont l'expression empêche la substance active de détruire la plante. Ce procédé permet d'utiliser moins d'herbicides ou des produits plus respectueux de l'environnement et est appliqué avec succès à de nombreuses espèces végétales telles que le colza, le blé ou la pomme de terre.

- La résistance aux maladies. Les virus, par exemple, provoquent également des dégâts dans les cultures. La résistance (ou la tolérance aux maladies) est donc une voie essentielle afin d'éviter des pertes de cultures.

- Enfin, la résistance aux conditions climatiques extrêmes. La recherche travaille sur la création de plantes capables de s'adapter à des conditions telles que la sécheresse, la salinité des sols ou le froid. Cela représente un grand intérêt pour les pays en développement tout comme pour les pays industrialisés.

* Dans le domaine de l'alimentation, des perspectives de développement sont attendues avec de nouveaux aliments possédant des caractéristiques telles que l'enrichissement du riz en vitamine A ou en fer, une diminution de la quantité de nitrates dans les salades, une modification en acides gras des huiles, afin de limiter les risques de maladies cardio-vasculaires ou encore d'améliorer la conservation des fruits en retardant leur flétrissement (tomate, melon). Il est cependant important de souligner qu'aucun aliment génétiquement modifié n'est aujourd'hui autorisé en Europe.

* Dans le domaine environnemental, on pourra envisager d'utiliser des micro-organismes permettant de dépolluer les sols contaminés et plus généralement d'éliminer les contaminants de l'environnement. Les biotechnologies employant aujourd'hui des enzymes permettent de traiter les eaux usées industrielles.

Les risques liés à l'utilisation des OGM

En opposition aux améliorations et aux éventuels avantages économiques apportés

par les OGM , se pose le problème des risques possibles sur la santé et l'environnement. En effet, la consommation de produits contenant des OGM ou issus d'OGM ne risque-t-elle pas d'être toxique pour l'homme, de causer des réactions allergiques ou de résistance aux antibiotiques ? Les OGM n'étant cultivés que depuis 1995 dans un nombre limité de pays, les données et le recul nécessaire manquent pour évaluer ces risques. En Europe, un seuil de tolérance qui s'applique ingrédient par ingrédient a été fixé à 1 % et doit prochainement passer à 0.9%. Ainsi, tous les produits contenant des protéines ou de l'ADN résultant d'une transformation génétique, ayant un taux d'OGM supérieur à 1%, portent, sur leur emballage, la mention "ingrédient issu d'OGM".

Du point de vue environnemental, les plantes génétiquement modifiées pour s'autoprotéger contre un insecte (comme le maïs résistant à la pyrale) pourraient susciter l'apparition d'insectes résistants à ces plantes transgéniques, à la suite d'une mutation génétique "naturelle" de ces insectes.
Mais le principal travail est l'étude de la dissémination de gènes des plantes génétiquement modifiées. Il est important, par sécurité notamment, de gérer les cultures transgéniques en assurant leur isolement. Or la culture de plantes transgéniques entraîne, par la dispersion du pollen, la possibilité de diffusion des gènes introduits par transgénèse aux variétés non modifiées de la même espèce ou aux espèces sauvages apparentées (les rendant alors par exemple résistantes à un herbicide). Il est donc nécessaire d'évaluer la dispersion des flux de pollen. C'est dans ce cadre que se situe la première partie du travail de cette thèse.

L'unité MIA de Jouy-en-Josas étudie essentiellement la dissémination du pollen de maïs et de colza. Le maïs est une espèce anémophile, c'est-à-dire que la dispersion du pollen a lieu uniquement à l'aide du vent contrairement au colza qui est une espèce anémophile et entomophile, c'est-à-dire que son pollen est dispersé également par les insectes. Une caractéristique du colza est de pouvoir se croiser avec des espèces sauvages apparentées. Ainsi, il a été constaté l'apparition de plantes sauvages résistantes aux herbicides. D'autre part, la dissémination est étudiée à deux échelles : au niveau d'un paysage agricole (expériences dans la région de Selommes) pour le colza et au niveau de deux champs contigus ou séparés par une autre culture (colza et maïs).

Au moment de l'apparition des premières céréales transgéniques, pour le maïs en 1988, l'objectif des études sur la dissémination du pollen était de déterminer une distance "minimale" au-delà de laquelle aucun échange de gènes (modifiés) n'aurait lieu. Mais, il est rapidement apparu qu'il est impossible de confiner un transgène strictement dans une parcelle cultivée, même si certaines hybridations ne se produisent qu'à une fréquence très faible. En effet, pour le maïs par exemple, des études suggèrent qu'il existe du pollen viable susceptible d'être transporté pendant la journée à des distances de plusieurs kilomètres. A l'heure actuelle, la séparation entre les différents champs est le principal moyen employé pour minimiser la pollution génétique d'un champ à un autre. Il est donc nécessaire, si un seuil de tolérance est donné, de déterminer une distance minimale entre deux champs au-delà de laquelle l'échange de gènes sera minime et inférieur à ce seuil.

La finalité de ces études, à partir des expériences réalisées, est donc de pouvoir faire de la prédiction, c'est-à-dire de pouvoir, à partir des résultats obtenus, élaborer des modèles valables dans différents paysages agricoles et différentes conditions climatiques afin de prévoir l'évolution du système écologique après l'introduction de plantes transgéniques dans une zone de culture : pollution des récoltes, évolution des populations sauvages, influence sur le comportement des insectes ... De plus, l'objectif est d'essayer de gérer l'occupation des sols au sein d'une zone de production.

1.1.1 Cadre de ce travail

La première partie de cette thèse est donc consacrée à la modélisation et l'estimation de la dispersion du flux de pollen de maïs à l'aide de modèles dits "mécanistes". En effet, la trajectoire d'un grain de pollen est modélisée à l'aide de processus de diffusion. Dans les précédents travaux, la position d'un grain de pollen au cours du temps était modélisée. Pour les modèles proposés, on s'est intéressé à modéliser les composantes du vecteur vitesse du grain de pollen soit dans le plan vertical, soit dans le plan horizontal.
De plus, la dispersion est étudiée au niveau de deux champs de maïs contigus (milieu dit homogène) et au niveau de deux champs de maïs séparés par un champ d'une autre culture (milieu dit hétérogène). Une méthode statistique est développée pour pouvoir traiter les données issues des expériences réalisées.

Il est à noter que ces modèles sont applicables à toute autre espèce anémophile dans des conditions de milieu homogène ou hétérogène. Le principe est également applicable pour l'étude de la dispersion par le vent de particules telles que les spores (Aylor 1990, McCartney *et al* 1999) ou les graines (Portnoy et Willson 1993).

1.1.2 Evaluation et modélisation de la dispersion du pollen

a) Les types d'expériences

Il existe principalement deux types d'expériences pour l'étude de la dissémination des graines ou du pollen (une description détaillée peut se trouver dans la thèse d'E. Klein, 2000) :

– "Expériences dites de type I" : elles consistent à planter un certain nombre de plantes sources puis à placer des capteurs physiques dans le voisinage de cette source. Pour observer la dispersion du pollen, les capteurs physiques peuvent être des lames enduites d'une substance collante. Pour la dispersion de graines, les capteurs peuvent être simplement des pots ou des zones délimitées sur le sol. On observe alors le nombre de graines ou de grains de pollen sur chaque capteur.
Dans ce type d'expériences, on observe un nombre de grains de pollen à différentes distances, mais on n'a pas idée de leur pouvoir fécondant. De plus, en général, le nombre total de grains émis par la source n'est pas connu. On ne peut donc pas donner la proportion de pollen attrapé par chaque capteur par rapport au pollen total émis par la source. Pour pallier ce problème, il est possible de placer

un capteur dans la source pour déterminer le nombre total de grains de pollen. Cependant, cette solution n'est sans doute pas la plus adaptée.

– "Expériences dites de type II" : elles utilisent des marqueurs génétiques.
On utilise comme première source de pollen des plantes homozygotes possédant un marqueur génétique dominant et comme deuxième source des plantes homozygotes ne possédant pas ce marqueur. Des plantes de ce second génotype sont alors utilisées comme capteurs et chaque descendant de celles-ci possédant le marqueur est de ce fait issu d'une fécondation par un grain de pollen provenant de la source marquée. Il existe différentes sortes de marqueurs : des marqueurs codant la couleur des graines (par exemple pour le maïs, il existe un gène non transgénique colorant les grains en bleu), la couleur des plantes ou encore une résistance à un herbicide (marqueur transgénique pour le colza par exemple).
Pour ces expériences, la dispersion observée est la dispersion efficace et on observe directement des proportions. Les expériences utilisées dans ce travail seront de ce type.

b) Dispersion globale et individuelle

A partir des expériences de type II, deux approches existent pour étudier les flux de pollen. Une première façon de travailler, appelée approche backward, est d'ajuster des modèles directement sur les observations représentant la dispersion bruitée de l'ensemble des plantes du champ (Morris *et al*, 1994). Cette dispersion globale du pollen, observée au point (x, y), est donc la probabilité pour qu'une graine localisée en ce point possède le marqueur et est notée $\mu(x, y)$. Cependant, cette fonction dépend beaucoup des dispositifs expérimentaux (essentiellement les tailles, formes et positions des sources) et donc ne permet pas d'approche prédictive.

Dans ce travail, c'est la seconde approche, dite forward, qui est adoptée, tout comme Nurminiemi *et al* (1998), Klein (2000) et Klein *et al* (2003). Il s'agit de modéliser et d'estimer, à partir des observations bruitées décrites ci-dessus, les paramètres de la fonction de dispersion efficace individuelle du pollen, notée γ. Cette fonction est une densité sur $\mathbb{R}^2$ et $\gamma(x, y)$ représente la probabilité qu'un grain de pollen émis au point $(0, 0)$ tombe et féconde un ovule situé en (x, y). La fonction de dispersion individuelle présente l'avantage de ne pas dépendre (théoriquement) du dispositif et fournit ainsi une quantification "robuste" de la dispersion du pollen.
La fonction de dispersion globale s'exprime comme un quotient de produits de convolution intégrant le dispositif expérimental et la fonction de dispersion individuelle. Plus précisément, le numérateur correspond aux contributions apportées par la source marquée au capteur situé en (x, y) et le dénominateur correspond aux contributions apportées par l'ensemble des plantes émettrices de pollen à ce capteur, permettant ainsi de considérer la compétition pollinique entre les plantes. Cette relation permet également de prendre en compte des phénomènes de décalage de floraison des fleurs ou de perte de viabilité du pollen au cours de leur trajectoire (Klein 2000). La relation exacte reliant les fonctions μ et γ se trouve au paragraphe suivant.

Ainsi, à partir des observations, différents modèles pour la fonction de dispersion individuelle peuvent être tester. Cependant, le fait d'avoir un quotient implique un calcul de déconvolution non linéaire, ce qui n'est pas un problème classique en statistiques. C'est pourquoi, par la suite, on étudiera des familles de fonctions de dispersion individuelles paramétriques de la forme $\{\gamma_\theta(x,y), (x,y) \in \mathbb{R}^2 \text{ et } \theta \in \Theta\}$.

1.1.3 Modélisation de la dispersion du pollen en milieu homogène à l'aide de modèles mécanistes

a) L'expérience

Les données utilisées proviennent d'une expérience réalisée par l'AGPM près de Montargis durant l'été 1998. Le champ est approximativement un carré de 120m sur 120m. Il a été cultivé suivant le modèle : 155 lignes distantes de 0.8m avec 800 plantes espacées de 0.2m par ligne. Au centre du champ, on a planté un carré de 20m sur 20m de maïs possédant un marqueur dominant non transgénique colorant les grains en bleu. En dehors de ce carré central, du maïs jaune a été semé (voir Figure 1.1). On appelle milieu homogène ce type de géométrie où des champs d'une même culture sont côte à côte.

Le début de l'étape de pollinisation a lieu quand le pollen est libéré de la panicule mâle. Ensuite a lieu une phase de transport puis de dépôt (fécondation). Un grain de pollen est viable en moyenne une heure (au maximum une heure et demie). Cela dépend des conditions climatiques : température, humidité de l'air par exemple. Le marqueur utilisé étant dominant, lorsqu'un grain de pollen provenant de maïs à grains bleus féconde une fleur de maïs à grains jaunes, on obtient un grain bleu. Lorsque l'on observe sur les épis le nombre de grains bleus, cela donne donc une image de toutes les pollinisations ayant eu lieu pendant la période de floraison, qui s'étale sur deux semaines environ.

Au total, 2937 épis ont été récoltés selon un maillage non régulier. En effet, l'échantillonnage est plus dense à proximité de la parcelle centrale afin d'avoir une estimation plus fine de la fonction de dispersion individuelle à courtes distances. Puis, le nombre de grains bleus sur chaque épi échantillonné a été compté. Dans ce travail, on fera l'hypothèse que le nombre total de grains sur un épi de maïs est constant et égal à 394 (nombre moyen).

b) Modèles mécanistes

Dans la littérature, il existe principalement deux façons pour modéliser la fonction de dispersion individuelle. La première, dite empirique, consiste à considérer des fonctions ayant une expression mathématique "simple" : fonctions exponentielles négatives, fonctions puissances décroissantes ou un compromis entre ces deux fonctions, par exemple une loi Gamma (Klein 2000, Nurminiemi *et al* 1998). Ces fonctions sont isotropes donc adaptées à des dispersions possédant cette propriété. Cependant, le maïs étant une espèce anémophile, la dispersion du pollen n'a pas lieu de manière isotrope (effet d'une direction dominante du vent).

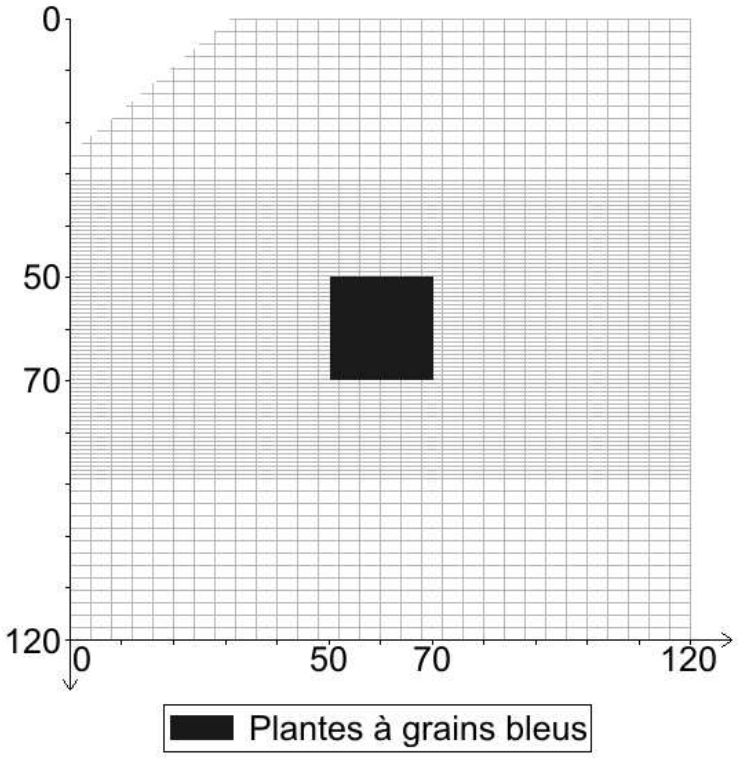


FIG. 1.1 – Dispositif expérimental avec maillage de l'échantillonnage

La seconde façon, adoptée dans ce travail, est de considérer des modèles dits "mécanistes". Le principe est de considérer le grain de pollen comme une particule soumise à un champ de forces lors de sa trajectoire (modèles dits de type Lagrangien). Ces modèles de transports de particules sont issus de la mécanique des fluides. Ils conduisent à l'utilisation d'équations différentielles stochastiques. Cette méthode a déjà été utilisée par Klein *et al* (2003) et Tufto *et al* (1997).
Il existe également des modèles dit Eulériens basés sur l'étude d'équations aux dérivées partielles (par exemple pour la dispersion du maïs, Yamamura, (2004) ; Loubet *et al* (2004) en milieu dit hétérogène).

Dans la suite, on note $P_t = (X_t, Y_t, Z_t)$ la position à l'instant t d'un grain de pollen. On modélise $(P_t)_{t\geq 0}$ par un processus de diffusion.
On introduit également T_F le temps de fécondation, c'est-à-dire le temps d'arrêt de la trajectoire sur une plante femelle.

A partir de la modélisation de la trajectoire, en faisant des hypothèses assurant que le couple (X_{T_F}, Y_{T_F}) admet une fonction de densité, on obtient alors une fonction de dispersion individuelle γ en calculant cette densité . Ainsi, γ est obtenue en cherchant la densité de probabilité du point de l'espace où la trajectoire rencontre un ovule et le féconde.

Une fois modélisée la trajectoire d'un grain de pollen, le but est d'estimer les pa-

ramètres des fonctions de dispersion individuelles à l'aide des données. Pour cela, il est donc nécessaire de définir un modèle mathématique.

c) Modèle statistique

On considère une source S_A de maïs à grains bleus localisés en $(x_k, y_k)_{k=1,\dots,S_A}$ et S_B une source composée de maïs à grains jaunes localisés en $(x_k, y_k)_{k=1,\dots,S_B}$.
Dans cette partie, on fait les hypothèses suivantes :

- (H1) : *Toutes les plantes dispersent leur pollen suivant la même fonction de dispersion individuelle γ.*
- (H2) : *Toutes les plantes produisent la même quantité de grains de pollen, quelque soit leur génotype.*
- (H3) : *Il n'y a pas de différence génétique entre les deux sortes de maïs (même durée de viabilité des grains de pollen, même taux de fécondation, même date de floraison).*

L'hypothèse (H1) ne tient pas compte des effets de bord sur les pollinisations aux extrémités du champ. Ceci est justifié pour de grands champs où peu de plantes sont soumises à cet effet. Elle est valable uniquement dans le cas du milieu homogène. Ainsi, la fonction de dispersion individuelle est indépendante de la position initiale de la plante émettrice.

Sous les hypothèses (H1)-(H3), il existe une relation liant les fonctions μ et γ (cadre discret) :

$$\mu(x,y) = \frac{\sum_{k=1}^{S_A} \gamma(x-x_k, y-y_k)}{\sum_{k=1}^{S_A} \gamma(x-x_k, y-y_k) + \sum_{k=1}^{S_B} \gamma(x-x_k, y-y_k)} \tag{1.1}$$

Lorsque la densité de plantes est suffisamment grande, la somme discrète peut être remplacée par une intégrale.

Lors de l'expérience pour le maïs, n descendants (c'est-à-dire le nombre total de grains sur un épi) ont été échantillonnés sur chaque capteur localisés en (x_k, y_k). On note N_k le nombre de grains bleus sur le capteur situé en (x_k, y_k).
Les données étudiées étant des données de comptage, l'hypothèse la plus naturelle est de considérer que les variables aléatoires N_k suivent une loi binomiale de paramètres $(n, \mu(x_k, y_k))$ (Klein *et al*, 2003). Cependant, on ne peut pas négliger les effets liés à la corrélation des génotypes des grains échantillonnés sur un même capteur ou les effets dus au fait que l'ensemble des observations n'ont pas été prises dans les mêmes conditions expérimentales (Collett, 1991).

C'est pourquoi, dans ce travail, on considère un modèle statistique de régression plus général de la forme :

$$N_k = n\mu(\theta; x_k, y_k) + \varepsilon_k \text{ avec } E(\varepsilon_k) = 0 \text{ et } Var(\varepsilon_k) = \sigma^2 nv(\theta, b; (x_k, y_k))$$

où les $(\varepsilon_k)_k$ sont supposées indépendantes.
Le paramètre σ^2 est appelé paramètre de dispersion.

Il est à noter que ce modèle statistique n'est pas spécifique à l'expérience utilisée. Il peut être utilisé pour traiter d'autres données sur une autre espèce de plantes ou pour un autre dispositif expérimental (cf le paragraphe 1.1.4 en milieu hétérogène).

d) Modélisations de la trajectoire

Modèles existants :

Les trois premières fonctions de dispersion étudiées sont basées sur une modélisation de la trajectoire d'un grain de pollen, $(P_t)_{t\geq 0}$, par trois mouvements browniens avec drift indépendants.
Ainsi, la trajectoire $(P_t)_{t\geq 0}$ s'écrit sous la forme

$$\begin{cases} dX_t = f_x dt + \tau_x dB_t^1 \\ dY_t = f_y dt + \tau_y dB_t^2 \\ dZ_t = f_z dt + \tau_z dB_t^3 \end{cases}$$

où les $(B_t^i)_{i=1,2,3}$ sont trois mouvements browniens indépendants ;
τ_x, τ_y, τ_z sont positifs et f_z est supposé négatif.
f_x et f_y représentent les composantes horizontales du vent moyen.
f_z représente la vitesse de chute due à la gravité, appelée aussi vitesse de sédimentation.
La matrice de variance-covariance représente les turbulences des écoulements.

Klein *et al* (2003) ont fait différentes hypothèses sur la loi du temps d'arrêt de la trajectoire sur une plante femelle. Elles conduisent à trois familles de fonctions de dispersion individuelles.

1. Première hypothèse : prédominance de la végétation
   La trajectoire s'arrête au bout d'un temps aléatoire indépendant de la trajectoire suivant une loi exponentielle, modélisant le temps d'atteinte sur n'importe quelle partie végétale. Ensuite, le temps de fécondation, T_F, est obtenu en conditionnant cette variable par l'événement : " le grain de pollen atteint la hauteur h des fleurs femelles".
   Dans ce cas, T_F suit une GIG (Generalized Inverse Gaussian) de paramètres $\frac{1}{2}$, $\lambda + \frac{f_z^2}{2\tau_z^2}$ et $\frac{h^2}{2\tau_z^2}$.
   La fonction de dispersion individuelle γ associée est alors la densité d'une GHD (Generalized Hyperbolic Gaussian, Barndorff-Nielsen 1997) avec le paramètre $\alpha = \frac{1}{2}$. Cette loi est appelée GTM (Generalized Tufto Model). (Ces fonctions sont définies au chapitre 2.)

2. Deuxième hypothèse : prédominance du sol
   Dans ce cas, la végétation n'arrête pas les trajectoires. Le temps de fécondation est défini comme le temps de premier passage de la trajectoire (du grain de pollen) au niveau $z = h$, hauteur des plantes femelles. D'après les propriétés du mouvement brownien, T_h suit une GIG de paramètres $\frac{3}{2}$, $\frac{f_z^2}{2\tau_z^2}$ et $\frac{h^2}{2\tau_z^2}$.

La fonction de dispersion individuelle γ est alors la densité d'une GHD avec le paramètre $\alpha = \frac{3}{2}$. Cette loi est appelée NIG (Normal inverse Gaussian).

3. Troisième hypothèse : généralisation
   Le temps de fécondation est une généralisation des deux cas précédents : il suit une GIG de paramètre α quelconque. Ce paramètre est lié à l'influence de la trajectoire sur la végétation. (Pour le premier modèle, on avait $\alpha = 1/2$ et pour le second $\alpha = 3/2$.)
   Alors , la fonction de dispersion individuelle est une GHD avec α quelconque. Ce modèle englobe les deux précédents.

Modèles proposés :

Deux nouvelles fonctions de dispersion individuelles ont été considérées dans ce travail. Les modèles précédents étaient uniquement basés sur la modélisation des composantes de la trajectoire. Ici, les modèles introduits sont basés sur une modélisation plus précise : on prend également en compte les composantes du vecteur vitesse de la trajectoire du grain de pollen.

– Dans un premier temps, dans le plan horizontal, la trajectoire est toujours modélisée par deux mouvements browniens indépendants avec drift. Par contre, en se basant sur les équations de Langevin (provenant de l'équation des forces en physique), la composante verticale est modélisée par un processus d'Ornstein-Uhlenbeck intégré indépendant des mouvements browniens. Ceci permet de tenir compte du fait que le grain de pollen est soumis à une force due à la gravité mais également à une force due au vent. La trajectoire pour Z_t est donc de la forme :

$$\begin{cases} dZ_t = V_t dt \\ dV_t = c_z dt - \beta V_t dt + \sigma_z dB_t \end{cases}$$

avec fc_z négatif et β, σ_z positifs.

V_t représente la vitesse verticale du grain de pollen. c_z est une force résultant de la gravité et $\sigma_z dB_t$ représente les contributions de la force exercée par le fluide (ici le vent) sur la particule durant son trajet, qui n'ont pas déjà été mises dans le terme linéaire de friction $-\beta V_t$.

Le temps de fécondation est toujours pris comme le temps de premier passage de la trajectoire au niveau h, mais cette fois-ci, c'est pour un processus O-U intégré et non plus un mouvement brownien avec drift.
Sous l'hypothèse $f_z < 0$, T_h est fini presque surement. Cependant il n'existe pas d'expression analytique exacte de la densité de ce temps de passage. Un changement de temps permet alors de se ramener au temps de premier passage d'une courbe dépendant du temps pour un mouvement brownien. Une approximation de la densité de cette variable aléatoire est alors faite en utilisant un théorème d'approximation (Durbin, 1992).
On obtient une nouvelle fonction de dispersion individuelle en calculant la loi du couple (X_{T_F}, Y_{T_F}).

– Dans un deuxième temps, les composantes de la vitesse dans le plan horizontal sont modélisées par des processus d'Ornstein-Uhlenbeck, noté O-U, indépendants et pour la composante verticale de la trajectoire, on conserve un mouvement brownien avec drift. La trajectoire $(P_t, t \geq 0)$ s'écrit donc sous la forme :

$$\begin{cases} dX_t = V_t^x dt \\ dY_t = V_t^y dt \\ dZ_t = f_z dt + \tau_z dB_t^3 \end{cases}$$

avec $\begin{cases} dV_t^x = -c_x V_t^x dt + \tau_x dB_t^1 \\ V_0^x = \eta_x \end{cases}$ et $\begin{cases} dV_t^y = -c_y V_t^y dt + \tau_y dB_t^2 \\ V_0^y = \eta_y \end{cases}$

$(X_0, Y_0, Z_0) = (0, 0, 0)$.

où les $(B_t^i)_{i=1,2,3}$ sont 3 mouvements browniens indépendants ;
τ_x, τ_y, τ_z sont positifs, c_x, c_y sont supposés positifs et f_z est supposé négatif, ce qui signifie que $T_h = \inf\{t > 0, Z_t = h\}$ est fini presque sûrement $(h < 0)$.
Cette modélisation permet ainsi de ne plus considérer constants les paramètres décrivant la vitesse et la direction du vent à la différence des modèles existants décrits ci-dessus. Tufto *et al* (1997) avait envisagé un modèle semblable. Cependant contrairement à leurs hypothèses, on ne suppose pas les processus stationnaires. Cela permet d'introduire une vitesse minimale d'émission des grains de pollen.
Les composantes X_t et Y_t obtenues sont des processus d'O-U intégrés.
Le temps de fécondation est défini comme le temps de premier passage de la trajectoire au niveau h.

Cela conduit à une fonction de dispersion individuelle sous forme d'une intégrale, qui numériquement pose des problèmes de convergence lors de l'utilisation de l'algorithme servant à estimer les paramètres.
Ainsi une approximation de ce modèle est alors étudiée, basée sur l'approximation des fonctions d'espérance et de variance des processus X_t et Y_t. De plus, les calculs de la fonction de dispersion globale se font dans un cadre continu contrairement aux autres modèles étudiés.

e) Validation des résultats

Pour estimer les paramètres des différentes familles paramétriques de fonctions de dispersion individuelles proposées au paragraphe précédent (modèles existants et proposés), on a employé une méthode semblable à la méthode de quasi-vraisemblance introduite par Wedderburn (1974) et utilisé l'algorithme de Gauss-Marquardt implémenté dans le logiciel Nls2 [1], avec lequel les calculs ont été effectués (en langage C).

Trois fonctions de variance v ont été envisagées :

[1] pour plus de renseignements : http ://www.inra.fr/bia/produits/logiciels

1. Les variables étudiées étant des données de comptage, il est naturel de considérer dans un premier temps une fonction de type binomiale de la forme : $v_1(\theta;(x_k,y_k)) = \mu(\theta;(x_k,y_k))(1-\mu(\theta;(x_k,y_k)))$.
2. Dans un second temps, la plupart des comptages étant nuls ou très proches de zéro (tous inférieurs à 0.2), une fonction de variance non nulle en 0, de type linéaire, est considérée afin de pondérer de façon correcte les observations significatives : $v_2(\theta,b;(x_k,y_k)) = b + \mu(\theta;(x_k,y_k))$ pour $0 \leq \mu \leq 0.5$.
3. Enfin, l'étude du graphique des variances empiriques en fonction de l'espérance a conduit à essayer de modéliser la fonction de variance par une fonction de type exponentielle : $v_3(\theta,b;(x_k,y_k)) = \exp(b\,\mu(\theta;(x_k,y_k))$.

Une fois les paramètres des différentes familles de fonctions de dispersion individuelles estimés, pour les trois fonctions de variance proposées, l'objectif est de déterminer quelle modélisation est la mieux adaptée aux données.
N'ayant fait des hypothèses que sur l'espérance et la variance des variables aléatoires N_k, l'utilisation d'un critère de sélection classique (type Akaïke ou C_p Mallows par exemple) n'est pas possible. Cependant, il existe un critère statistique de sélection de modèle de type Akaïke (Hurvich et Tsai, 1995) dans le cas d'une méthode de quasi-vraisemblance.

D'autre part, des méthodes graphiques sont disponibles :
– L'étude des résidus réduits en fonction des ajustés et sur le champ. Si le modèle est assez proche de la réalité, la plupart des résidus doit être compris entre -2 et 2 et le graphique ne doit pas présenter de structure particulière.
– L'étude des ajustés. Elle permet de comparer les valeurs observées aux valeurs estimées.
– La comparaison des différentes courbes des fonctions de dispersion individuelles.
Enfin, pour les modèles mécanistes, les paramètres estimés peuvent être reliés à certains paramètres biologiques et physiques, en particulier, des données météorologiques relevées lors de l'expérience.

f) Interprétation des résultats

Les résultats montrent que la modélisation ajustant le mieux les données de l'expérience est la fonction de dispersion individuelle γ de type NIG avec une fonction de variance de type linéaire et un paramètre de dispersion dû à la corrélation des génotypes des grains d'un même capteur. Les composantes de la trajectoire d'un grain de pollen correspondant à cette modélisation sont alors trois mouvements browniens avec drift indépendants et le temps de fécondation est défini comme le temps de premier passage au niveau des fleurs femelles. De plus, les variables aléatoires, associées en coordonnées polaires, (R,Θ), ne sont pas indépendantes comme l'avaient supposé Tufto *et al* (1997).
Les modélisations plus fines proposées de la trajectoire basées sur les composantes de la vitesse dans le plan vertical ou horizontal ne se sont pas révélés être meilleures contrairement à ce qu'on aurait pu penser. Sans doute est-ce dû au fait que des approximations ont été faites pour obtenir des résultats lors de l'estimation des

paramètres. D'autre part, on a dû utiliser la formule reliant les fonctions de dispersion individuelle et globale, donnée en (1.1), dans un cas continu. On peut tout de même noter que la fonction de dispersion obtenue pour le modèle 4 se rapproche de celle du modèle 2, NIG.

Un travail en perspective serait d'intégrer les nouveaux résultats obtenus au logiciel MAPOD. Ce logiciel a été développé par l'unité Eco-Innov de l'INRA Grignon et le laboratoire ESE d'Orsay. Il permet de simuler, dans des conditions physiques proches de la réalité (prise en compte des caractéristiques physiques des grains de pollen (poids, hauteur des fleurs femelles) et de paramètres liés au vent (direction, intensité)), des périodes de pollinisation du maïs et par suite d'étudier la dispersion du pollen, par exemple en faisant varier la taille et la forme des champs.

En effet, un inconvéniant rencontré ici est le manque d'expériences disponibles. Actuellement a lieu une mise en commun, dans le cadre du projet européen SIGMEA, d'expériences sur la dispersion de flux de pollen.

Le but, à la base, est de pouvoir prédire quelle distance doit séparer deux champs pour qu'il n'y ait pas de pollution génétique d'une culture par l'autre. Cependant les précédents modèles ne permettent pas de décrire de façon satisfaisante la dispersion du pollen au niveau d'une discontinuité dans un paysage, c'est-à-dire quand deux champs de maïs sont séparés par un sol nu ou une autre culture. Klein (2000) avait simulé des trajectoires à partir de la fonction de dispersion individuelle trouvée pour le milieu homogène pour une expérience décrite au paragraphe a) ci dessous. La comparaison entre les valeurs simulées et les données n'était pas très satisfaisante.
La suite du travail présenté ici est donc d'essayer de modéliser le flux de pollen en présence d'une discontinuité.

1.1.4 Modélisation de l'effet d'une discontinuité du couvert végétal

Une discontinuité du couvert végétal correspond à séparer deux champs, ici de maïs, par un autre type de culture ou un sol nu.
Des équipes de bioclimatologie, de l'INRA-Grignon et Bordeaux, ont développé deux modèles dits de type Eulérien, basés sur l'étude de la turbulence et du transport de pollen de maïs (Loubet *et al*, 2004). Ils permettent d'estimer la concentration et le dépôt en aval d'une source donnée. Cependant, ils présentent le désavantage d'avoir des temps de calculs très longs.

Dans cette partie, à partir des résultats obtenus précédemment, on va donc essayer d'étudier la dispersion du pollen.
Il existe deux modes de dispersion du pollen : une grande partie reste à petite altitude et a une durée de vie de une à deux heures. Une petite partie est emportée par des vents ascensionnels à haute altitude et vit plus longtemps (les basses températures et le taux d'humidité le rendent plus viable). Ainsi, pour le maïs, des études suggèrent qu'il existe du pollen viable susceptible d'être transporté à des distances de plusieurs kilomètres (Loubet *et al*, 2004). La question de savoir si ce pollen peut effectivement

polliniser est ouverte.

Les modèles présentés dans la suite considèrent uniquement le premier mode de dispersion. La difficulté essentielle est que nous ne pouvons plus considérer que toutes les plantes dispersent leur pollen suivant la même fonction de dispersion individuelle. En effet, pour chaque plante du champ, la discontinuité se trouve à une distance différente. Il faut donc tenir compte de la position initiale du grain de pollen l'effet de la discontinuité sur la fonction de dispersion individuelle.
Dans le cas du colza, la méthode proposée par Poilleux-Milhem (2002) consistait en partie à introduire un paramètre de translation. Cependant ici, dans le cas du maïs, la fonction de dispersion individuelle n'est pas isotrope.

a) Les expériences

Deux expériences ont eu lieu dans la région de Montargis en 1999 et 2000 sur deux champs d'environ 200 $m \times 160$ m. Pour chacun des champs, une parcelle de 20 $m \times$ 20 m de maïs homozygote pour la coloration du grain en bleu a été semée. Autour des parcelles de maïs à grains bleus, une autre culture a été plantée d'une largeur d'environ 50 m. Pour la première expérience, il s'agit de trèfle (considéré comme un sol nu) et pour la seconde, il s'agit de tournesol (culture d'une hauteur élevée). Dans le reste du champ, du maïs à grains jaunes (homozygote pour l'absence du marqueur) de la même variété a été semé. Un échantillonnage d'épis de maïs a été fait sur un maillage régulier du champ de maïs jaune au moment de la récolte.
Une difficulté rencontrée ici, comme dans le cas du milieu homogène, est le manque d'expérience. En effet, par exemple, il n'y a pas d'expérience "témoin" en milieu homogène qui aurait permis, dans un premier temps, d'estimer les paramètres de la fonction de dispersion individuelle . Il faut donc estimer simultanément ces paramètres et les paramètres liés à l'effet de la discontinuité du couvert végétal.

b) Introduction de l'effet d'une discontinuité du couvert végétal

Dans un premier temps, la trajectoire du grain de pollen a été modélisée par trois mouvements browniens avec drift indépendants dans les deux champs de maïs et par une trajectoire déterministe dans la partie cultivée avec du trèfle ou du tournesol. Le temps de fécondation introduit tient alors compte du fait qu'un grain de pollen peut se fixer et féconder une fleur dans le premier champ ou mourir dans la discontinuité ou encore se fixer sur une fleur du deuxième champ. Cela a conduit au calcul d'une fonction de dispersion individuelle malheureusement non exploitable numériquement étant composée de plusieurs intégrales doubles.

Pour résoudre le problème, il a fallu trouver une autre méthode. Partant de la fonction de dispersion individuelle de type NIG, notée $f_{NIG}(x, y; \theta)$ qui sert de modèle de base, noté modèle A, l'effet de la discontinuité a été modélisé en introduisant des paramètres de "translation" (Dans le cas de l'étude de la dispersion du pollen de colza, Poilleux-Milhem (2002) avait également introduit un tel paramètre, fonction de la largeur de la discontinuité.)

On remplace l'hypothèse (H1) par
(H1') : *Une plante émettrice située en* (x_0, y_0) *disperse son pollen suivant la fonction de dispersion individuelle* $\gamma_{(x_0,y_0)}$*, qui dépend de sa position.*

On suppose donc dans la suite (H1'), (H2) et (H3).

Les deux expériences disponibles se prêtant mal à une analyse en coordonnées polaires, une méthode mixte est proposée.

Si $P_0 = (x_0, y_0)$ est la position de la plante émettrice et $P_1 = (x_1, y_1)$ la position de la plante réceptrice, se situant dans les champs de maïs, notés D_1 ou D_3 (voir Figure 1.2), le grain de pollen est supposé parcourir une distance D_{x_0,x_1} suivant l'axe des x et une distance D_{y_0,y_1} suivant l'axe des y dans la discontinuité, notée D_2.
On modélise alors l'effet de la discontinuité du couvert végétal dans le domaine D_2 en introduisant des paramètres de translation $\alpha_{1,i}$ suivant l'axe des x et $\alpha_{2,i}$ suivant l'axe des y de la façon suivante :
Pour $(x, y) \in \mathbb{R}^2$, on considère ses coordonnées polaires associées (r, θ). Alors pour $1 \leq i \leq m$, on définit l'ensemble $S_i = \{(x, y) \in \mathbb{R}^2, u_i \leq \theta \leq u_{i+1}\}$. Les (u_i) sont tels que $\cup_i S_i = \mathbb{R}^2$.
Pour $(x, y) \in D_1$ ou D_3 et $(x, y) \in S_i$, on remplace donc $(x - x_0)$ par $(x - x_0 - \alpha_{1,i} D_{x_0,x})$ et $(y - y_0)$ par $(y - y_0 - \alpha_{2,i} D_{y_0,y})$.

Dans la pratique les ensembles S_i sont choisis par rapport aux directions du vent et m varie entre 2 et 4.

Alors, la fonction de dispersion individuelle pour une plante émettrice située en (x_0, y_0) est définie par

$$\gamma_{(x_0,y_0)}(x, y) = \begin{cases} f_{NIG}(x - x_0 - \alpha_1 D_{x_0,x}, y - y_0 - \alpha_2 D_{y_0,y}; \theta) & \\ 0 & \text{si } (x, y) \in D_2 \\ f_{NIG}(x - x_0 - \alpha_{1,i} D_{x_0,x}, y - y_0 - \alpha_{2,i} D_{y_0,y}; \theta) & \text{si } (x, y) \in D_3 \cap S_i \end{cases}$$

On remarque que $\alpha_i > 0$ signifie qu'il y a un effet d'accélération sur la dispersion du pollen et inversement, si $\alpha_i < 0$, il y a un effet de ralentissement.

Le second modèle considéré tient compte du fait de la forte pollinisation en bordure de la discontinuité comme on peut le voir sur les représentations graphiques des données (voir Chapitre 3).
On note $\tilde{D}_3$ l'ensemble constitué des points de D_3 situés à au plus 1 mètre de D_2 dans la ou les direction(s) du vent.
On va considérer que si l'épi de maïs appartient à l'ensemble $\tilde{D}_3$ et si le grain de pollen passe dans la discontinuité, alors la fonction de dispersion individuelle est égale à la somme d'un paramètre $q \in]0, 1[$ et du terme de la fonction individuelle proposée pour le précédent modèle. Sinon on prend l'expression de la fonction individuelle proposée pour le précédent modèle.

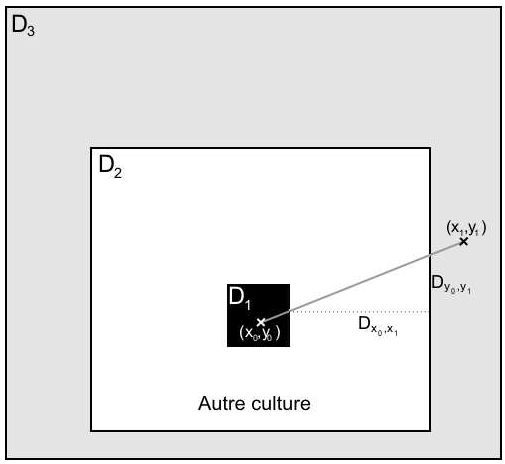


FIG. 1.2 – Modélisation de la discontinuité à l'aide de paramètres de translation

c) Résultats et perspectives

Le modèle statistique utilisé pour estimer les paramètres des modèles proposés est celui décrit en 1.1.2 c).

Les résultats des estimations des paramètres pour les différents modèles montrent qu'une discontinuité du couvert végétal a un effet d'accélération sur la dispersion du pollen. Dans le cas du trèfle, il y a une plus grande pollution à la bordure de la discontinuité par rapport au cas du tournesol où la pollution a tendance à se répartir plus sur l'ensemble du champ.

Le modèle le plus adapté aux données est le premier modèle décrit ci-dessus. Cependant, pour le tournesol en particulier, il faudrait envisager un paramètre q qui soit une fonction dépendant de θ afin de mieux prendre en compte les différentes directions de vent.

Il serait également souhaitable de réaliser de nouvelles expériences, si possible couplées avec des expériences en milieu homogène, afin de comprendre de façon précise quel est l'effet d'une discontinuité sur les paramètres des modèles et de quoi dépendent les paramètres de translation introduits ci-dessus (par exemple, on peut penser qu'ils vont dépendre de l'intensité du vent). Ceci dans le but de pouvoir effectuer des prédictions.
Une première perspective est de traiter des données similaires, disponibles depuis peu dans le projet SIGMEA.
D'autre part, il pourrait être intéressant d'intégrer ces nouveaux résultats au logiciel MAPOD qui actuellement ne prend pas en compte les effets du paysage (discontinuité due à l'isolement ou à une autre culture ou encore à la présence de haies). Un travail de simulations pour introduire des distances d'isolement croissantes (de quelques mètres à quelques centaines de mètres) et différentes conditions de vent pourrait également être mis en place dans le but final d'effectuer des prédictions et donc de déterminer une distance d'isolement.

1.2 Partie II : Estimation dans des modèles à volatilité stochastique

Dans ce travail, on considère un processus stochastique bi-dimensionnel (Y_t, V_t) défini par :

$$\begin{cases} dY_t = \sigma_t dB_t, \quad Y_0 = 0 \\ dV_t = b(\theta, V_t)dt + a(\theta, V_t)dW_t \\ V_t = \sigma_t^2, \quad V_0 = \eta \end{cases} \tag{1.3}$$

où (B_t, W_t) est un mouvement Brownien standard de $\mathbb{R}^2$, η une variable aléatoire indépendante de $(B_t, W_t)_{t\geq 0}$.

Ce modèle est appelé modèle à volatilité stochastique et σ_t^2 représente la volatilité instantanée.

Ces modèles ont été proposés par Hull et White (1987) et sont souvent utilisés en finance. Pour une revue sur les modèles à volatilité stochastique, on pourra se reporter à Ghysels *et al* (1996).

Le modèle de Black et Scholes (1973) en est un exemple simple. Le prix S_t d'un actif financier est ainsi modélisé par un mouvement brownien géométrique où le processus Y_t, défini par $Y_t = \ln(S_t)$ vérifie l'équation : $dY_t = (\mu - \sigma^2/2)dt + \sigma dB_t$ avec B un mouvement brownien , μ et σ des constantes. (σ est appelée la volatilité de l'actif)

Le but est d'estimer le vecteur de paramètres θ du processus (V_t), θ étant inconnu.

L'estimation des paramètres lorsque la diffusion (V_t) est observée a déja été beaucoup étudiée avec différentes hypothèses faites sur les observations. La diffusion peut être observée de façon continue sur un intervalle de temps $[0, T]$ (Kutoyants, 1984). Elle peut également être observée de façon discrétisée avec ou bien Δ_n tend vers 0 avec $n\Delta_n$ tend vers l'infini ou $n\Delta_n = T$ (Kessler, 1997 ; Genon-Catalot et Jacod, 1993) ; ou bien Δ est fixé (Kessler et Sørensen, 1999 ; Kessler, 2000).

Dans ce travail, la diffusion (V_t) n'est pas observée. On a des observations discrétisées du processus intégré (Y_t). Plus précisement, pour Δ fixé et positif, on observe le processus $(Z_i)_{1\leq i\leq n}$ défini par

$$Z_i = \frac{1}{\sqrt{\Delta}} \int_{(i-1)\Delta}^{i\Delta} \sigma_s \, dB_s = \frac{1}{\sqrt{\Delta}} (Y_{i\Delta} - Y_{(i-1)\Delta})$$

On introduit également les variables $(\bar{V}_i, i \geq 1)$ définies par

$$\bar{V}_i = \frac{1}{\Delta} \int_{(i-1)\Delta}^{i\Delta} V_s \, ds$$

On peut constater que la variance du processus (Z_i), conditionnellement au passé, est une fonction dépendant de ce passé et est hétéroscédastique. Un exemple plus simple est le modèle à temps discret GARCH (Bollerslev, 1986) dont la variance conditionnelle dépend des observations passées et de la variance précédente.

Dans la littérature, une première approche consiste à étudier le processus $(Z_i, i \geq 1)$ en faisant tendre n vers l'infini avec $\Delta = \Delta_n$ dépendant de n et tendant vers 0, la longueur du temps de l'observation $n\Delta_n$ tendant vers l'infini. Genon-Catalot *et al* (1999) proposent une méthode explicite basée sur des fonctions des observations $(Y_{i\Delta_n}, 1 \leq i \leq n)$. (Cette approche inclut la méthode de moments empiriques proposée dans Genon-Catalot *et al*, 1998.) Gloter (2000) propose un contraste dans le cas où (Y_t) est observé avec un double pas de discrétisation.

Dans la suite, Δ est considéré fixé. Une difficulté est que la loi exacte du processus intégré $(Z_i, i \geq 1)$ n'est en général pas explicite, excepté pour quelques modèles. Ainsi, la vraisemblance du modèle est difficile à calculer explicitement. D'autre part, le processus $(\bar{V}_i, i \geq 1)$ n'est pas markovien. Cependant, $(Z_i, i \geq 1)$ peut être vu comme une chaîne de Markov cachée avec pour chaîne cachée $(U_i, i \geq 1)$ où $U_i = (\bar{V}_i, V_{i\Delta})$. Les méthodes statistiques pour les chaînes de Markov cachées ont été essentiellement développées dans le cas où l'espace d'états de la chaîne était fini ou compact. Ici, ce n'est plus le cas puisque l'espace de temps est infini. Genon-catalot *et al* (2000) ont étudiées les propriétés de ces chaînes de Markov cachées.

Différentes méthodes d'estimation ont déjà été étudiées. Par exemple : l'utilisation du filtre de Kalman (Harvey, 1989) ; une fonction de contraste basée sur une méthode d'espérance conditionnelle (Genon-Catalot *et al*, 2003), ou des méthodes de Monte-Carlo bayésiennes (Kim *et al*, 1998).

Dans toute la suite de ce travail, on s'intéressera aux processus de diffusion cachée appelés "mean-reverting" et très souvent utilisés en finance. Pour ces modèles, la fonction de dérive b est de la forme :

$$b(\theta, V_t) = \alpha(\beta - V_t)$$

De plus, les hypothèses sur la fonction a sont telles que le processus (V_t) est stationnaire strict et ergodique, et donc η admet pour loi la loi stationnaire.

1.2.1 Estimation pour des modèles à volatilité stochastique mean-reverting à l'aide de variantes de la méthode de Whittle

Les modèles mean-reverting possèdent certaines propriétés propres dues à la forme de la fonction de dérive (Genon-Catalot *et al*, 2003 ; Sørensen, 2000).

Pour le modèle étudié, on a la propriété suivante sous des hypothèses assurant l'ergodicité et la stricte stationnarité de (V_t) :

Proposition 1.1 *:*
Le processus $((Z_i^2 - \beta), i \geq 1)$ est centré et a une structure d' ARMA(1,1). Plus précisément, on a

$$Z_i^2 - \beta - e^{-\alpha\Delta}(Z_{i-1}^2 - \beta) = \varepsilon_i - \psi(\theta)\varepsilon_{i-1}$$

où (ε_i) *est un bruit blanc centré avec* $\mathrm{Var}(\varepsilon_i) = \sigma^2(\theta)$ *et* $|\psi(\theta)| < 1$.
Si $e^{-\alpha\Delta} \neq \psi(\theta)$ *alors le processus est de plus causal et inversible.*
La densité spectrale est alors donnée, pour $\lambda \in [0, 2\pi]$, *par :*

$$f(\lambda, \theta) = \frac{\sigma^2(\theta)}{2\pi} \frac{1 + \psi(\theta)^2 - 2\psi(\theta)\cos\lambda}{1 + e^{-2b} - 2e^{-b}\cos\lambda} \tag{1.4}$$

(Les expressions exactes de $\psi(\theta)$ et $\sigma^2(\theta)$ sont données au Chapitre 4, Proposition 1).

1) Estimation des paramètres par la méthode de Whittle et Whittle raboté :

Pour un processus ARMA causal et inversible, une méthode usuelle pour estimer les paramètres est de maximiser la fonction de vraisemblance gaussienne. L'estimateur du maximum de vraisemblance possède de bonnes propriétés asymptotiques : il est consistant asymptotiquement et converge en loi à la vitesse $\sqrt{n}$ (cf Brockwell et Davis 1991). (Dans le cas où le processus est un AR, l'algorithme de Durbin-Levinson peut également être utilisé. De même dans le cas d'un MA, on peut utiliser l'algorithme des innovations).

Ici, le processus considéré n'est pas gaussien. Par conséquent, on utilise la fonction de minimum de contraste de Whittle (provenant de l'approximation de la vraisemblance, Whittle 1953) définie par :

$$U_n(\theta) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left[\ln(2\pi f(\lambda, \theta)) + \frac{I_n(\theta)}{f(\lambda, \theta)} \right] d\lambda \tag{1.5}$$

où $I_n(\theta)$ représente le périodogramme associé au vecteur $(Z_i^2, i \geq 1)$.
L'estimateur de Whittle est alors défini par $\hat{\theta_n} = arg \inf_{\theta \in \Theta} U_n(\theta)$.

Sous certaines hypothèses, on peut calculer explicitement la fonction de minimum de contraste de Whittle pour un modèle à volatilité stochastique mean-reverting (Proposition 2 du Chapitre 4).

L'estimateur de Whittle est consistant et converge asymptotiquement en loi vers une loi normale quand n tend vers l'infini à la vitesse $\sqrt{n}$. Plus précisément, on a sous l'hypothèse d'inversibilité de la matrice d'information $\Gamma(\theta_0)$:

$$\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow[n \to +\infty]{\mathcal{D}} \mathcal{N}\left(0, \Gamma^{-1}(\theta_0)(\Gamma(\theta_0) + B(\theta_0))\Gamma^{-1}(\theta_0)\right)$$

où $B(\theta_0)$ une matrice faisant intervenir la fonction spectrale d'ordre 4, représentant un degré de non Gaussianité du processus étudié et définie au Chapitre 4.

Quand l'échantillon est de taille petite, l'estimateur de Whittle peut être mauvais. Pour palier à ce problème éventuel, le processus original est alors multiplié par une fonction dite de rabotage, notée $(h_i, i \geq 1)$. Le périodogramme est remplacé par une version rabotée (Dahlhaus, 1988) obtenue en remplaçant les $(Z_i^2 - \bar{Z}, i \geq 1)$ par $(h_i(Z_i^2 - \bar{Z}), i \geq 1)$ et en le renormalisant, avec $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i^2$. Enfin, la fonction

de minimum de contraste est obtenue en remplaçant le périodogramme initial par celui "raboté" dans l'expression calculée.

L'estimateur de Whittle raboté est également consistant et converge asymptotiquement en loi vers une loi normale quand n tend vers l'infini à la vitesse $\sqrt{n}$ (Dahlhaus, 1988).

$$\sqrt{n}(\tilde{\theta_n} - \theta_0) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, C\Gamma^{-1}(\theta_0)(\Gamma(\theta_0) + B(\theta_0))\Gamma^{-1}(\theta_0)\right)$$

avec $C = \lim_{n\to+\infty} \frac{nH_{4,n}}{H_{2,n}^2}$ lorsque cette limite existe (les termes $H_{4,n}$ et $H_{2,n}$ étant définis au Chapitre 4, Paragraphe 4.3.3.

2) Estimation des paramètres par une méthode statistique à deux pas :

Sous des hypothèses supplémentaires pour le processus (V_t),

Les résultats des estimations des paramètres par les méthodes précédentes, lors de simulations, montrent que le paramètre β, représentant l'espérance de Z_1^2, est mal estimé. On propose donc d'estimer ce paramètre par un estimateur de moments empiriques en utilisant un résultat de Genon-Catalot *et al* (2000) qui ont montré que l'estimateur $\beta_n = \bar{Z}^2$ est fortement consistant asymptotiquement et converge en loi vers une loi normale $\mathcal{N}(0, \sigma^2)$ à la vitesse $\sqrt{n}$.
Les autres paramètres du modèle, notés ω, sont alors estimés à l'aide de la méthode de Whittle raboté (et Whittle) en remplaçant dans la fonction de minimum de contraste obtenue précédemment β par β_n. On démontre alors le théorème suivant sous certaines hypothèses :

Théorème 1.1 *L'estimateur $(\hat{\omega}_n)_n$ est consistant en ω_0.*

Un résultat de convergence asymptotique en loi à la vitesse $\sqrt{n}$ n'a pas été démontré. Cependant, l'étude sur les simulations laisse à penser que c'est le cas.

Simulations

Des simulations ont été réalisées pour comparer la performance des différents estimateurs proposés sur deux exemples pour $\Delta = 0.1$ et 1 (estimateurs de Whittle et Whittle raboté, estimateurs des moments empiriques et estimateurs de la méthode mixte Whittle et Whittle raboté). Il est à noter que, quand $\Delta = 1$, les observations (Z_i) sont presque non corrélées.

Les deux modèles étudiés sont :

- Modèle 1 : $a(V_t) = cV_t$. Ce modèle apparaît comme une approximation d'un modèle GARCH(1, 1) (Nelson, 1990).
- Modèle 2 : $a(V_t) = c\sqrt{V_t}$ avec $c \in \mathbb{R}$. Ce modèle a été proposé par Heston (1993). La fonction a est la fonction racine carrée, utilisée par Cox *et al* (1985) pour modéliser les taux d'intérêt.

Dans ces cas, le vecteur des paramètres est de dimension 3 : $\theta = (b, \beta, c)$. Une fonction de rabotage de type Tukey-Hanning a été utilisée dans le cas de l'estimateur de Whittle raboté (Dahlhaus, 1988).

Tout d'abord, dans l'ensemble, les résultats pour le modèle 2 sont meilleurs que ceux pour le modèle 1. Une explication possible, suggérée par Genon-Catalot *et al* (1999), peut être que le paramètre lié au nombre de degrés de liberté d'une loi de Student est mal estimé même dans le cas d'observations indépendantes. Ceci a déja été remarqué dans des applications en finance.

Dans l'ensemble, les résultats avec la méthode à deux pas rabotée donnent de meilleurs résultats que ceux obtenus avec les autres méthodes étudiées. De plus cette méthode permet un gain de temps lors du calcul numérique de l'estimateur (étape de minimisation de la fonction de minimum de contraste de Whittle). Le paramètre β est toujours estimé de façon satisfaisante. Cependant, ce n'est pas le cas pour les paramètres b et c suivant le modèle et la valeur de Δ choisie.
Pour finir, on peut remarquer qu'à Δ fixé, plus n est grand, plus les estimations sont meilleures et les écart-types petits.

Finalement, ces résultats montrent la nécessité pour l'estimation des modèles à volatilité stochastique (en temps continu) d'avoir un grand nombre d' observations sur un long intervalle de temps (ce qui est souvent le cas dans le domaine de la finance).

1.2.2 Estimation par méthode d'inférence indirecte pour des modèles mean-reverting avec un effet leverage

Les modèles à volatilité stochastique décrits précédemment peuvent être trop restrictifs pour modéliser les données financières. Par conséquent, il est possible d'introduire un effet dit de leverage de la façon suivante :

$$\begin{cases} dY_t = \sigma_t dB_t + \rho dW_t, \quad Y_0 = 0 \\ dV_t = b(\theta, V_t)dt + a(\theta, V_t)dW_t \\ V_t = \sigma_t^2, \quad V_0 = \eta \end{cases} \tag{1.6}$$

où (B_t, W_t) est un mouvement brownien standard de $\mathbb{R}^2$, η une variable indépendante de $(B_t, W_t)_{t\geq 0}$, θ un paramètre inconnu à estimer et ρ l'effet leverage supposé négatif. (Ainsi quand ρ est égal à 0 on retrouve le modèle classique défini au paragraphe précédent.)
Il est à noter, qu'ici, l'effet leverage est introduit comme dans Barndorff-Nielsen et Shephard (2001). Cependant il existe d'autres façons de le faire en supposant par exemple que les deux mouvements browniens (B_t) et (W_t) sont corrélés entre eux.
Ce modèle fut introduit par Black (1976). En finance, ce paramètre ρ signifie qu'un accroissement positif dans la volatilité aura un effet négatif sur le prix des stocks.

Comme pour les modèles à volatilité stochastique classiques, il est difficile de calculer explicitement la fonction de vraisemblance. De plus, la structure du modèle ne permet pas d'étendre facilement les résultats statistiques pour l'estimation des paramètres obtenus dans le cas des modèles à volatilité stochastique. Par exemple les méthodes proposées par Genon-Catalot *et al* (2003) ou Sørensen (2000) nécessitent le calcul de moments d'ordre un ou plus pour les variables (Z_i). Or peu de calculs explicites sont faisables, excepté dans quelques cas. Par exemple, cela est

possible lorsque le processus (Z_i) est un processus de Lévy d'Ornstein-Uhlenbeck (Barndorff-Nielsen et Shepard, 2001).

Pour faciliter l'estimation des paramètres, on suppose que le processus non observé est une discrétisation (V_i) de la diffusion continue (V_t). Les observations (Z_i) sont alors de la forme $Z_i = F(V_i, r_i)$ où (r_i) est une séquence de dimension 2 de variables i.i.d. de loi normale.
Il est à noter que les modèles à volatilité stochastique mean-reverting en temps continu peuvent être obtenus comme approximation de diffusions de modèles discrets de type ARCH (Nelson, 1990). C'est le cas en particulier du modèle étudié lors des simulations dans le paragraphe précédent, qui peut être vu comme l'approximation d'un modèle GARCH(1, 1).
Il semble que peu de travaux aient été faits sur les modèles leverage. Jacquier *et al* (2004) ont utilisé des méthodes de Monte-Carlo bayésiennes dans le cas particulier de modèles log-normals, c'est-à-dire que $\ln \sigma_i^2 = a + b \ln \sigma_{i-1}^2 + c\varepsilon_i$. (Il est à noter qu'ils considèrent que les deux séquences de bruit blanc sont corrélées.)

Le modèle étudié :

On s'intéresse toujours au cas des modèles mean-reverting. Le modèle considéré s'écrit, pour $i \geq 1$,

$$\begin{cases} Z_i = \sigma_i r_i + \rho\varepsilon_i \\ V_i = \alpha\beta\Delta + (1 - \alpha\Delta)V_{i-1} + \sqrt{\Delta}a(V_{i-1})\varepsilon_i, \quad V_0 = \eta \\ V_i = \sigma_i^2 \end{cases} \tag{1.7}$$

où (r_i) et (ε_i) sont deux séquences de variables aléatoires de loi $\mathcal{N}(0, 1)$ indépendantes et indépendantes de V_0. a est une fonction réelle positive pouvant dépendre d'un paramètre inconnu noté c. α, β et Δ sont supposés positifs ; et en finance ρ est supposé négatif (pour avoir un effet leverage).

Seul le processus (Z_i) est observé.
Sous de certaines hypothèses, précisées au Chapitre 5, les processus (V_i) et (Z_i) satisfont de bonnes propriétés.

Propriété 1.1 *(V_i) est l'unique chaîne de Markov strictement stationnaire solution de (1.7) admettant un moment d'ordre 2. De plus, (V_i) est ergodique et α-mélangeant avec des coefficients décroissant exponentiellement vers 0.*
Le processus (Z_i) est une chaîne de Marov cachée et est également strictement stationnaire, ergodique et α-mélangeant avec des coefficients décroissant exponentiellement vers 0.

Les modèles mean-reverting sans effet leverage en temps continu peuvent être vu comme une approximation de diffusions de modèles discrets de type ARCH ou GARCH. De plus, les modèles étudiés sont strictement stationnaires et faciles à simuler. La méthode choisie ici pour estimer les paramètres est une méthode d'inférence indirecte (Gourieroux *et al* (1993) et Gallant et Tauchen (1996)).

Ainsi l'estimation des paramètres est basée sur l'introduction d'un modèle auxiliaire pour lequel la fonction de vraisemblance est facilement calculable contrairement au modèle étudié.

La méthode statistique et le modèle auxiliaire proposé :

De façon naturelle, on a envie de choisir un modèle auxiliaire assez simple et ayant des similitudes avec le modèle étudié. Ici, pour le modèle défini en (1.7), on remarque que lorsque ρ est égal à 0, Z_i et le début du processus (V_i) sont le début d'un processus GARCH. On a donc choisi comme modèle auxiliaire un GARCH(2, 1), c'est-à-dire

$$\begin{cases} y_i = \sqrt{h_i}\nu_i \\ h_i = \omega_0 + \omega_1 y_{i-1}^2 + \omega_2 h_{i-1} + \omega_3 h_{i-2} \end{cases} \tag{1.8}$$

avec (ν_i) i.i.d. $\mathcal{N}(0,1)$; $\omega_0 > 0$ et $0 \leq \omega_1 + \omega_2 + \omega_3 < 1$ (pour avoir l'existence d'une solution admettant un moment d'ordre 2).

On prend alors comme critère auxiliaire la fonction de log-vraisemblance conditionnelle associée au GARCH(2, 1) :

$$Q_n(\mathbf{y_n}, \omega) = \frac{1}{n-1} \sum_{i=2}^{n} \left(\ln(h_i) + \frac{y_i^2}{h_i} \right) \tag{1.9}$$

En minimisant le critère 1.9, on obtient un estimateur $\hat{\omega}_n = \arg\max_{\omega \in \Omega} Q_n(\mathbf{Z}_n, \omega)$.

Comme le gradient de la fonction Q_n est calculable de façon explicite (par itération), on a choisi d'utiliser l'estimateur d'inférence indirect proposé par Gallant et Tauchen (1996) et qui est asymptotiquement équivalent à celui proposé par Gourieroux *et al* (1993). En effet, cela permet d'avoir une seule étape de minimisation au lieu de deux. Il est basé sur la fonction score et défini comme suit :

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{\partial Q_{nH}}{\partial \omega'}(\tilde{\mathbf{z}}_{nH}(\theta), \hat{\omega}_n) \Sigma_n \frac{\partial Q_{nH}}{\partial \omega}(\tilde{\mathbf{z}}_{nH}(\theta), \hat{\omega}_n)$$

où Σ_n est une matrice aléatoire définie positive et convergeant vers une matrice déterministe Σ également définie positive.
De plus, $\tilde{\mathbf{z}}_{nH}(\theta) = \{z_i(\theta), i = 1, ..., nH\}$ est une trajectoire simulée du processus étudié de dimension nH $(H \geq 1)$.

Sous certaines hypothèses (précisées à la section 3 du Chapitre 5), $\hat{\theta}_n$ est un estimateur consistant de θ_0 quand n tend vers $+\infty$ et converge asymptotiquement en loi vers une loi normale à la vitesse $\sqrt{n}$.

D'après la Propriété 1.1, dans le cas du modèle étudié, on peut appliquer le théorème ergodique et un théorème centrale limite pour les processus α-mélangeant (Hall et Heyde, 1980), permettant ainsi de vérifier toutes les hypothèses nécessaires pour avoir la consistance et la convergence en loi de l'estimateur d'inférence indirect.

Simulations :

Finalement des simulations ont été faites pour étudier la performance de cet estimateur avec les deux modèles définis précédemment. Un troisième modèle avec $a(x) = cx^{0.7}$ a également été étudié.

Les paramètres β, ρ et c sont bien estimés. Pour les modèles 1 et 3, le paramètre α est difficile à estimer (tout comme à la section précédente pour l'estimation par une méthode de Whittle). La valeur obtenue devient correcte quand $n = 3000$.

Pour conclure, en prenant comme modèle auxiliaire un GARCH d'ordre supérieur, on pourrait étudier les modèles avec une fonction a de la forme $a(x) = cx^{\gamma}$ où $\gamma \in (\frac{1}{2}, 1)$ serait un paramètre inconnu à estimer.

D'autre part, il serait intéressant, pour mieux prendre en compte la structure du modèle leverage, de choisir comme modèle auxiliaire un modèle GARCH non paramétrique cette fois-ci, de définir un critère auxiliaire et d'étudier le comportement asymptotique des estimateurs ainsi obtenus.

1.3 Partie III : Mélange de modèles mixtes pour l'analyse des appariements de chromosomes chez le colza

Le dernier chapitre est consacré à l'analyse du déterminisme génétique des appariements de chromosomes chez des haploïdes de colza. Ce travail a été commencé lors de mon stage de DEA avec Sylvie Huet et Hervé Monod dans l'unité MIA de l'INRA de Jouy-en-Josas.

Lors de la méiose de colzas haploïdes (ne contenant qu'une copie du génôme), un certain nombre de chromosomes homologuent s'apparient et les autres, dits univalents, restent non appariés. Le nombre d'univalents est variable et dépend en particulier de la variété de colza. L'objectif est de savoir si le contrôle des appariements de chromosomes, lors de la méiose, est dû à l'action d'un gène unique ou non.

Le modèle statistique utilisé est un modèle de mélange, dont chacune des deux composantes suit un modèle mixte. Les paramètres sont estimés par maximum de vraisemblance à l'aide d'un algorithme ECM. Le test de rapport de vraisemblance est présenté pour deux hypothèses sur les paramètres, dont l'une inclut la nullité d'un paramètre de variance.

Deux articles sont parus et font l'objet de ce chapitre.

1.4 Références

Aylor, D.E. (1990). The role of intermittent wind in the dispersal of fungal pathogens. *Annual Review of Phytopathology*, **28**, 73-92.

Black, F. (1976). Studies of stock price volatility changes. *Proceesinds of the Business and Economic Statistics section, American Statistical Association*, 177-181.

Black, F. et Scholes, M. (1973). The valuation of option and corporate liabilities. *Journal of political economy* **81**, 637-284.

Barndorff-Nielsen, O.E. (1997). Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling. *Scandinavian Journal of Statistics* **24**, 1-13.

Barndorff-Nielsen, O.E. et Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion). *J. Roy. Statist. Soc. Ser. B* **63**, 167-241.

Bollersler, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **51**, 307-327.

Brockwell, P.J. et Davis, R.A. (1991). *Time Series : Theory and Methods*. Springer-Verlag.

Collett, D. (1991). *Modelling binary data* Chapman and hall, London.

Cox, J. C., Ingersoll, J. E. et Ross, S. A. (1985). A theory of term structure of interest rates. *Econometrica* **53**, 385-407.

Dahlhaus, R. (1988). Small sample effects in time series analysis : a new asymptotic theory and a new estimate. *Annals of Statistics* **16**, 808-841.

Durbin, J. (1992) The first-passage density of the brownian motion process to a curved boundary. *J. Appl. Prob.* **29**, 291-304.

Gallant, A.R. et Tauchen, G. (1996) Which moments to match. *Econometric theory* **12**, 657-681.

Genon-Catalot, V. et Jacod, J. 1993 On the estimation of the diffusion coefficient for multidimensional diffusion processes. *Ann. Inst. Henri Poincaré, Probab-Statist* **29**, 119-151.

Genon-Catalot, V., Jeantheau, T., Laredo, C. (1998) Limit theorems for discretly observed stochastic volatility models. *Bernouilli* **4**, 283-303.

Genon-Catalot, V., Jeantheau, T., Laredo, C. (1999) Parameter estimation for discretely observed stochastic volatility models. *Bernouilli* **5**, 855-872.

Genon-Catalot, V., Jeantheau, T., Laredo, C. (2000) Stochastic volatility models as hidden Markov models and statistical applications. *Bernouilli* **6**, 1051-1079.

Genon-Catalot, V., Jeantheau, T., Laredo, C. (2003) Conditional Likelihood Estimators for Hidden Markov Models and Stochastic Volatility Models. *Scandinavian Journal of Statistics* **30**, 297-316.

Ghysels, E., Harvey, A., Renault, E. (1996) Stochastic volatility. *Handbook of Statistics* **14** , 119-192.

Gloter, A. (2000) *Estimation des paramètres d'une diffusion cachée.* Thèse de doctorat, Université de Marne-La-Vallée.

Gourieroux, C., Monfort, A. et Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics* **8**, 85-118.

Hall, P. et Heyde, C.C (1980) *Martingale Limit Theory and its Application.* New-York : Academic Press.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge : Cambridge University Press.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility, with applications to bond and currency options. *Review of Financial Studies* **6**, 327-343.

Hull, J. et White, A. (1987). The pricing of options on assets with stochastic volatilities. *J. Finance* **42**, 281-300.

Hurvich, C.M. et Tsai, C.L. (1995). Model selection for extended quasi-likelihood in small samples. *Biometrics* **51**,1077-1084.

Jacquier, E. , Polson, N.G. et Rossi, P.E. (2004) Bayesian analysis of stochastic volatility models with a fat-tails and correlated errors. *Journal of Econometrics* **122**, 185-212.

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics.* **24**(2), 211-229.

Kessler, M. (2000). Simple and explicit estimating functions for a discretely observed diffusion process *Scandinavian Journal of Statistics* **27**, 65-82.

Kessler, M. et Sorensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion processes. *Bernouilli* **5**, 299-314.

Kim, S., Shephard, N., Chib, S. (1998). Stochastic volatility : likelihood inference and comparaison with ARCH models. *Review of Economic Studies* **65**, 361-393.

Klein, E. (2000) *Estimation de la fonction de dispersion du pollen. Application à la dissémination de transgènes dans l'environnement.* Thèse, Université Paris XI, Orsay.

Klein, E.K., Lavigne, C., Foueillassar, X., Gouyon, P.H., Laredo, C. (2003) Corn pollen dispersal : quasi-mechanistic models and field experiments. *Ecological Monographs* **73**,131-150.

Kutoyants, Y. (1984). Parameter estimation for stochastic processes. *Heldermann*, Berlin.

Loubet, B , Brunet, Y;, Fouieillassar, X., Caltagirone, J.P. *et al.*, (2004). Etude mécaniste du transportet du dépôt de pollen de maïs dans un paysage hétérogène. Rapport de fin de projet.

McCartney, H.A. et Fitt, B.D.L. (1998). Dispersal of foliar fungal plant pathogens : mechanisms, gradients and spatial patterns. Pages 138-160 dans *The epidemiology of plant diseases*. Kluwer, Dordrecht, The Netherlands.

Morris, W. F., Kareiva, P.M. et Raymer P.L. (1994). Do barren zones and pollen traps reduce gene escape from transgenic crops ? *Ecological Applications* **4**, 157-165.

Nelson, D.B. (1990). ARCH models as diffusion approximations. *J. Econometrics* **45**, 7-38.

Nurminiemi M., Tufto J., Nilsson O., Rognli O.A. (1998). Spatial models of pollen dispersal in the forage grass meadow fescue. *Evolutionary Ecology* **12**, 487-502.

Poilleux-Milhem, H. (2002) *1) Test de validation adaptatif dans un modèle de régression. 2) Modélisation et estimation de l'effet d'une discontinuité du couvert végétal sur la dispersion du pollen de colza.* Thèse, Université Paris XI, Orsay.

Portnoy, S. et Willson, M.F. (1993). Seed dispersal curves : behaviour of the tail of the distribution. *Evolutionary Ecology* **7**, 25-44.

Sørensen, M. (2000). Prediction-based estimating functions. *Econom. J.* **3**, 121-147.

Tufto, J., Engen, S., Hindar, K. (1997). Stochastic Dispersal Processes in Plant Populations. *Theoretical Population Biology* **52**, 16-26.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

Whittle, P. (1953). Estimation and information in stationary time series. *Ark. Mat.***2**, 423-434.

Yamamura, K. (2004). Dispersal distance of corn pollen under fluctuating diffusion coefficient. *Popul Ecol* **46**, 87-101.

Première partie

Modélisation stochastique et estimation de la dispersion du pollen de maïs

Chapitre 2

Modélisation stochastique et estimation de la dispersion du flux de pollen du maïs en milieu homogène

Sommaire

2.1 Introduction et motivations

Depuis quelques années maintenant, l'apparition et l'utilisation de plantes génétiquement modifiées ont amené un certain nombre de questions au sein de notre société, en particulier sur les risques de transfert de gènes dans l'environnement et sur les conséquences pour l'homme.
Les enjeux économiques et les débats sociaux autour des organismes génétiquement modifiés (OGM) ont donc conduit à étudier les flux de gènes. L'étude de la dissémination du pollen à partir d'expériences et d'observations permet d'essayer de modéliser les flux de pollen et de développer des modèles permettant de prédire la dispersion de ces gènes à différentes échelles : au niveau d'un champ, au niveau d'un paysage.
Dans ce contexte, il est nécessaire, pour un seuil de tolérance est donné, de déterminer une distance minimale entre deux champs au-delà de laquelle l'échange de gènes sera minime et inférieur au seuil. En effet, pour le maïs par exemple, des études suggèrent qu'il existe du pollen viable susceptible d'être transporté pendant la journée à des distances de quelques kilomètres.
Mais quel taux de "pollution" de transgènes est acceptable pour les consommateurs et l'environnement ?

Dans ce chapitre, nous nous intéressons à la modélisation du flux de pollen de maïs en milieu homogène, c'est à dire à l'échelle d'une parcelle.
Notre objectif est d'estimer une quantification "robuste" de la dispersion du pollen, c'est-à-dire une quantification ne dépendant pas du dispositif expérimental utilisé. Cela nous permettra alors d'utiliser les résultats obtenus pour un champ ayant une configuration différente et donc de pouvoir effectuer des prédictions.
Le plus souvent, les expérimentateurs travaillent avec la fonction de dispersion globale du pollen qui représente la probabilité qu'une plante possède le transgène. Cette fonction est mesurable sur le terrain de manière simple à l'aide de capteurs. Cependant, cette fonction dépend beaucoup du dispositif expérimental utilisé. C'est pourquoi nous allons travailler sur la fonction de dispersion individuelle, représentant la dispersion efficace du pollen et ayant l'avantage de ne pas dépendre du dispositif (ou très peu).
Il existe principalement deux façons pour modéliser la fonction de dispersion individuelle. La première, dite empirique, consiste à considérer des fonctions ayant une expression mathématique "simple" : fonctions exponentielles négatives, fonctions puissances décroissantes ou compromis entre ces deux fonctions (Klein 2000, Nurminiemi *et al* 1998). Ces fonctions sont isotropes donc adaptées à des dispersions possédant cette propriété.
Ici, nous nous intéressons uniquement à la modélisation dite "mécaniste" : on considère que le grain de pollen est une particule soumise à un champ de force lors de sa trajectoire et un temps de fécondation est introduit, représentant le temps d'arrêt de la trajectoire lors de la pollinisation. Cela permet de prendre en compte le phénomène de dispersion dans une direction dominante (celle du vent) pour le maïs. A partir d'une expérience réalisée par l'AGPM (Association Générale des Producteurs

de Maïs), nous avons pu essayer plusieurs modèles mécanistes et paramétriques.

Dans un premier temps, nous avons fait des hypothèses sur le modèle mathématique. En effet, les observations résultant de l'expérience représentent une proportion de grains de maïs bleus pour les épis échantillonnés. Cette proportion est reliée à la dispersion efficace à l'aide d'un quotient de produits de convolution par les formules définies en (3.1) et (2.2) et introduites par Klein (2000), Tufto *et al* (1996).
Klein *et al* (2003) avaient fait l'hypothèse que ces proportions suivaient une loi binomiale. Mais la comparaison avec les données météorologiques n'était pas concluante. Dans ce travail, nous introduisons donc un paramètre de dispersion afin de prendre en compte la corrélation des génotypes des grains de pollen (Collett 1991, Huet *et al* 1996), et nous proposons de modéliser la fonction de variance de trois façons différentes : une fonction de type binomiale, une de type linéaire et une de type exponentielle.

Dans la suite du chapitre, se trouve la description des modèles mécanistes proposés. Tout d'abord, nous reprenons les modèles proposés par Klein *et al* (2003) où les trois composantes de la trajectoire sont supposées être des mouvements browniens avec drift indépendants. Trois cas sont envisagés pour le temps de fécondation. Trois fonctions de dispersion individuelles sont obtenues, notées : GTM (Generalized Tufto Model) (Tufto *et al* 1997), NIG (Normal Inverse Gaussian) et GHD (Generalized Hyperbolic Distribution) (Barndorff-Nielsen 1997 par exemple).

Puis nous proposons deux nouvelles modélisations de la trajectoire, plus fines, basées sur la modélisation des composantes du vecteur vitesse, et non les composantes de la trajectoire.

Pour la première modélisation, les composantes de la trajectoire dans le plan horizontal restent modélisées par deux mouvements browniens avec drift mais nous modélisons la composante de la vitesse dans le plan vertical à l'aide d'un processus d'Ornstein-Uhlenbeck. Le temps de fécondation est le temps de premier passage de la trajectoire au niveau des plantes femelles. Nous approchons la fonction de densité de ce temps de fécondation à l'aide d'un théorème d'approximation, dû à Durbin (1992), du temps de premier passage d'une courbe dépendant du temps pour un mouvement brownien standard. Et nous obtenons une nouvelle fonction de dispersion individuelle.
Pour le second modèle proposé, les composantes du vecteur vitesse dans le plan horizontal sont modélisées par des processus d'Ornstein-Uhlenbeck indépendants. Ainsi, on ne considère plus constante la vitesse du vent tout au long de la trajectoire et cela permet d'introduire une vitesse minimale d'émission des grains de pollen, hypothèse souvent envisagée par les biologistes. Pour la composante verticale, nous conservons un mouvement brownien avec drift, indépendant des composantes dans le plan horizontal. Le temps de fécondation considéré est encore le temps de premier passage de la trajectoire au niveau des plantes femelles. En effectuant certaines approximations, on obtient une nouvelle fonction de dispersion individuelle.

Ces différents modèles mécanistes conduisent à différentes familles paramétriques

de fonctions de dispersion individuelles. Ayant considéré un modèle de régression, nous avons ensuite estimé tous ces paramètres à l'aide d'une méthode de quasi-vraisemblance (Wedderburn, 1974).

Pour finir, nous avons analysé et validé les résultats obtenus afin de choisir le modèle mécaniste et la fonction de variance les mieux adaptés aux données. Pour cela, nous avons à notre disposition d'une part un critère statistique de sélection de modèle pour une estimation par quasi-vraisemblance (MacQuarrie and Tsai 1998). D'autre part, il existe également des méthodes graphiques : étude des résidus réduits par rapport aux ajustés, sur le champ ; étude des courbes de dispersion obtenues ; étude des ajustés. Enfin, nous comparons les paramètres physiques calculés à partir des mesures réalisées au cours de l'expérience et des données météorologiques.

2.2 Description de l'expérience

2.2.1 Dispersion du pollen de maïs

Le maïs est une espèce anémophile, c'est-à-dire que la pollinisation n'a lieu que grâce au vent, contrairement au colza par exemple où les insectes interviennent dans cette étape.
Le pollen part d'une hauteur H située au niveau des étamines sur la panicule mâle et la pollinisation a lieu au niveau des soies sur la fleur femelle à une hauteur h', avec $h' < H$.
Pour le maïs, on dispose d'un marqueur dominant, non transgénique, qui colorie les grains de maïs en bleu. A chaque fois qu'une fleur est pollinisée par du pollen provenant d'un maïs à grains bleus, on observe dans l'épi récolté un grain bleu. Ainsi, le nombre de grains bleus parmi tous les grains d'un épi reflète exactement l'intensité de la pollinisation sur un épi donné.

2.2.2 Description de l'expérience

L'expérience a été réalisée par l'AGPM près de Montargis durant l'été 1998. Le champ est approximativement un carré de 120m sur 120m. Il a été cultivé suivant le modèle : 155 lignes distantes de 0.8m avec 600 plantes espacées de 0.2m par ligne. Au centre du champ, on a planté un carré de 20m sur 20m de maïs bleu. En dehors de ce carré central, du maïs jaune a été semé.
(Le coin en haut à gauche est en fait une route donc il n'y a pas eu de culture à cet endroit.)

La dispersion du pollen a commencé en Juillet et a duré deux semaines. Les épis ont été récoltés et analysés en Octobre.
Au total, 2937 épis ont été récoltés selon un maillage non régulier. En effet, on a échantillonné 101 lignes (1 ligne sur 3 ou chaque ligne suivant la distance au point central) et on a prélevé 31 épis sur chaque ligne (un tous les 4m). L'échantillonnage est plus dense à proximité de la parcelle centrale afin d'avoir une estimation plus fine de la fonction de dispersion individuelle à courtes distances.

A la Figure (2.1) se trouve une représentation du dispositif.
Puis, on a compté le nombre de grains bleus sur chaque épi échantillonné. Ici, on fera l'hypothèse que le nombre total de grains sur un épi de maïs est constant et égal à 394 (nombre moyen estimé de grains sur un épi).
La représentation des proportions observées du nombre de grains bleus sur chaque épi échantillonné se trouve à la Figure (2.26).
Pour finir, des données météorologiques effectuées pendant l'étape de pollinisation sont disponibles. Elles permettront de valider les différents modèles proposés.

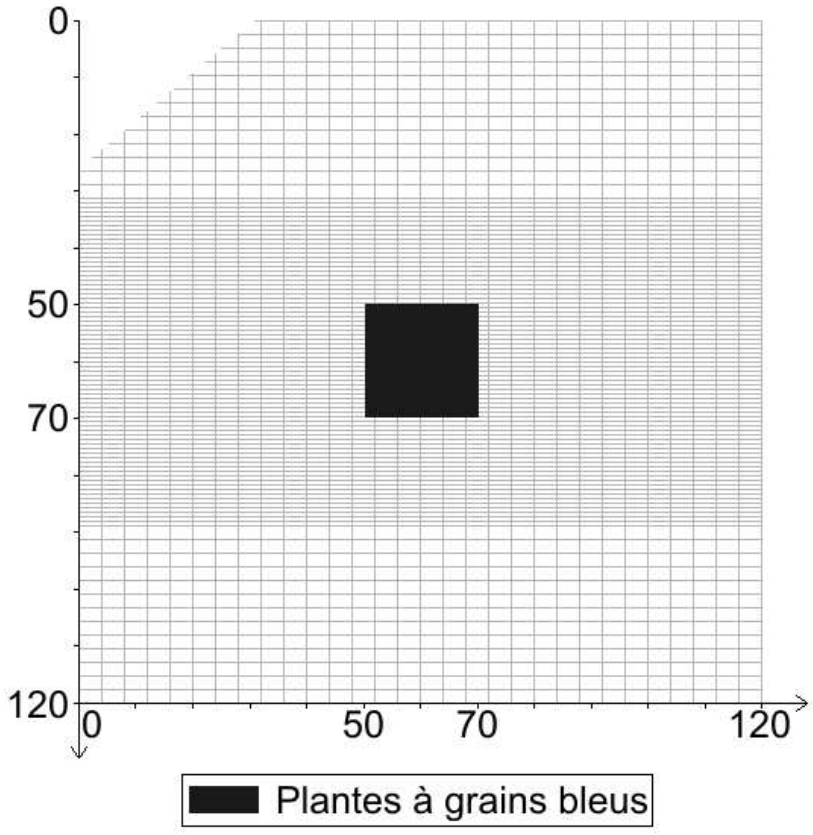


FIG. 2.1 – Dispositif expérimental avec maillage de l'échantillonnage

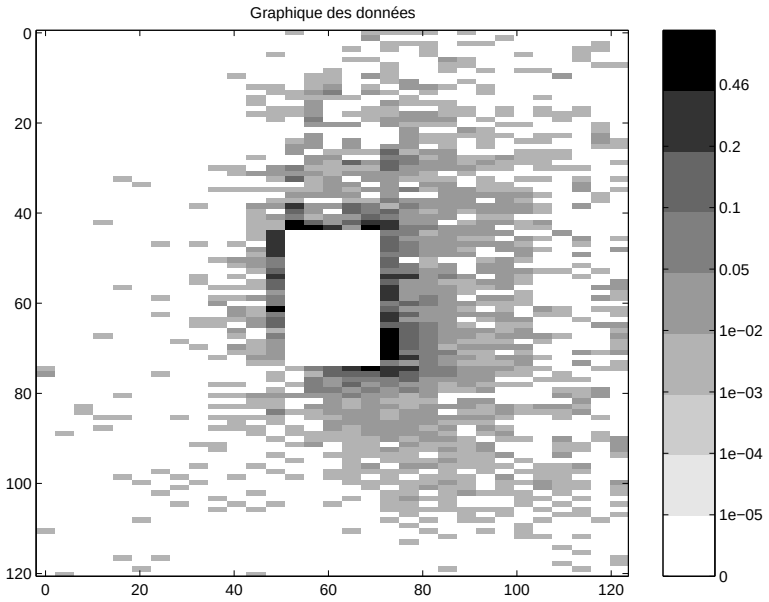


FIG. 2.2 – Proportions observées des grains bleus sur les épis échantillonnés

2.3 Modélisation de la dispersion du pollen

2.3.1 Fonctions de dispersion

On souhaite pouvoir faire des prévisions de la dispersion dans différents dispositifs de culture. Une expérience doit donc être capable de fournir une quantification "robuste" de la dispersion du pollen, c'est-à-dire une dispersion de pollen ne dépendant pas du dispositif expérimental (taille, forme de la source émettrice et de la partie réceptrice). Deux fonctions sont utilisées dans la littérature pour décrire cette dispersion.

Tout d'abord il y a la fonction de dispersion efficace individuelle du pollen, notée par la suite γ.

Définition 2.1 *$\gamma(x,y)dxdy$ représente la probabilité qu'un grain de pollen émis en $(0,0)$ tombe et féconde une plante dans le rectangle $((x,y),(x+dx,y+dy))$. γ est une densité de probabilité sur $\mathbb{R}^2$.*

La seconde fonction utilisée est la fonction de dispersion globale.

Définition 2.2 *La dispersion globale du pollen est la probabilité pour qu'un grain de maïs situé au point (x,y) soit bleu. Elle est notée $\mu(x,y)$.*

La dispersion globale du pollen est le résultat des dispersions individuelles de plusieurs plantes jaunes et bleues. De plus, le pollen émis étant surabondant, les fleurs d'une plante réceptrice sont fécondées suivant les différentes proportions de pollen en présence dans le nuage pollinique au dessus de cette plante.

La fonction μ dépend beaucoup des dispositifs expérimentaux utilisés (surtout les tailles et formes des sources) tandis que la fonction de dispersion individuelle γ est plus robuste et presque indépendante du dispositif. C'est pourquoi c'est sur cette fonction que l'on va travailler, comme l'ont fait précédemment Nurminiemi *et al* (1998), E. Klein (2000) et en particulier Klein *et al* (2003).
Cette approche est appelée approche "forward". En opposition avec l'approche "backward" plus souvent utilisée par les expérimentateurs et qui consiste à travailler avec la fonction de dispersion globale μ, qui résulte directement des observations expérimentales.

Dans toute la suite de ce chapitre, on fait les hypothèses suivantes :
(H1) *Toutes les plantes dispersent leur pollen suivant la même fonction de dispersion individuelle γ.*
(H2) *Toutes les plantes produisent la même quantité de grains de pollen, quel que soit leur génotype.*
(H3) *Il n'y a pas de différences intrinsèques entre les plantes marquées et non marquées (même viabilité, même taux de fécondation, même période de floraison).*

Remarque : l'hypothèse (H1) est essentielle dans le travail qui suit. Elle est raisonnable pour des champs d'expérience de grande taille où la densité de plantes est assez élevée et où l'on suppose que seules quelques plantes du bord de champ subissent des effets de bords.
Cette hypothèse est liée à l'"homogénéité" de la parcelle. Elle ne tient plus dans le cas d'une discontinuité du couvert végétal.

Sous ces hypothèses, il existe une relation liant la fonction de dispersion individuelle et la fonction de dispersion globale.

On considère une source composée de S_A plantes de maïs à grains bleus et localisées en $(x_k, y_k)_{k=1,\dots,S_A}$ et une source composée de S_B plantes de maïs à grains jaunes et localisées en $(x_k, y_k)_{k=1,\dots,S_B}$. Alors, on a

$$\mu(x,y) = \frac{\sum_{k=1}^{S_A} \gamma(x - x_k, y - y_k)}{\sum_{k=1}^{S_A} \gamma(x - x_k, y - y_k) + \sum_{k=1}^{S_B} \gamma(x - x_k, y - y_k)} \tag{2.1}$$

Dans cette expression, le numérateur correspond aux contributions apportées par la source marquée au capteur situé en (x, y) et le dénominateur correspond aux contributions apportés par l'ensemble des plantes émettrices de pollen à ce capteur, permettant ainsi de considérer la compétition pollinique entre les plantes.

Cette approche a été utilisée par Tufto *et al* (1997) et généralisée par Klein (2000) afin de prendre en compte des phénomènes de décalage de floraison des fleurs ou de perte de viabilité du pollen au cours de leur trajectoire.

Ainsi, à partir des données du nombre de grains bleus sur un épi, qui constituent des observations bruitées de la fonction μ dont nous disposons, le travail consistera à estimer la fonction de dispersion individuelle γ.

Si la densité des plantes est suffisamment grande, on peut utiliser une version continue de la relation 3.1. Soit $A \subset \mathbb{R}^2$ le sous-ensemble couvert par la source des plantes de maïs à grains bleus et B le sous-ensemble couvert par les plantes de maïs à grains jaunes. Alors si on suppose que la densité est la même dans les deux sources , on a

$$\mu(x,y) = \frac{\int\int_{(x',y')\in A} \gamma(x-x',y-y')dx'dy'}{\int\int_{(x',y')\in A} \gamma(x-x',y-y')dx'dy' + \int\int_{(x',y')\in B} \gamma(x-x',y-y')dx'dy'} \tag{2.2}$$

$$\mu(x,y) = \frac{\mu * \mathbb{I}_A(x,y)}{\mu * \mathbb{I}_{A\bigcup B}(x,y)}$$

On remarque, qu'en fait, les fonctions μ et γ sont liées par un produit de convolution. Ainsi, l'estimation de la fonction de dispersion individuelle à partir des observations, en utilisant les égalités (3.1) et (2.2), nécessite des calculs effectués par les ordinateurs qui peuvent être coûteux en temps, en particulier pour la définition dans le cadre discret.

2.3.2 Définition du modèle statistique

On note n_k le nombre total de grains sur l'épi localisé en (x_k, y_k) et on observe N_k le nombre de grains bleus sur cet épi.

Comme les observations sont des données de comptage, les variables aléatoires N_k sont souvent supposées être des lois binomiales de paramètres $(n_k, \mu(x_k, y_k))$ (c'est l'hypothèse qu'ont fait Klein *et al*, 2003). Cette hypothèse est basée sur le fait que si les génotypes des grains de pollen qui fécondent les ovules d'une même plante sont indépendants, alors N_k suit une loi binomiale (comme somme de n_k variables aléatoires de Bernoulli indépendantes de paramètre $\mu(x_k, y_k)$).

Cependant, on ne peut sans doute pas négliger les effets liés à la corrélation des génotypes des grains échantillonnés sur un même épi (Collett 1991 et McCullagh et Nelder 1989). Cela peut être dû à une hétérogénéité temporelle du vent accompagnée d'une variabilité de la période de fertilité des ovules, ce qui fait que les graines d'une même plante sont fécondées dans les mêmes conditions de vent.
C'est pourquoi un modèle statistique de régression plus général est considéré avec l'introduction d' un paramètre de dispersion. On suppose donc que

$$N_k = n\mu(\theta; x_k, y_k) + \varepsilon_k \text{ avec } E(\varepsilon_k) = 0 \text{ et } Var(\varepsilon_k) = \sigma^2 n v(\theta, b; \mu(x_k, y_k))$$

où les $(\varepsilon_k)_k$ sont supposées indépendantes.
Différentes fonctions de variance v sont envisagées par la suite et décrites au paragraphe 2.6.

L hypothèse d'indépendance des observations entre différents capteurs est réaliste : elle provient du fait que les grains de pollen dans le nuage pollinique sont très nombreux et non limitants. Le fait d'observer beaucoup de grains dans un capteur n'affecte donc pas le nombre de grains trouvé dans les autres capteurs.

Le paramètre σ^2 doit être strictement positif. De plus, si l'on note $\sigma_k^2 = (1+d(n_k-1))$, le paramètre d représente la corrélation entre les génotypes de deux descendants échantillonnés sur un même capteur (Collett, 1991 ; Huet *et al*, 1996).
Si le paramètre d est positif, il est appelé paramètre de sur-dispersion (et on a $\sigma_k^2 > 1$). Dans le cas contraire, le paramètre d est appelé paramètre de sous-dispersion (et on a $0 < \sigma_k^2 \leq 1$).

Enfin, d'après la relation (3.1), on peut remarquer que l'on est face à un problème statistique de déconvolution non linéaire qui n'est pas classique. C'est pourquoi, dans la suite, on s'intéressera à l'étude de fonctions de dispersion individuelles paramétriques.

2.4 Modèles mécanistes pour le maïs

Les modèles mécanistes sont des modèles utilisés pour la dispersion anémophile, donc quand la pollinisation a lieu seulement avec l'aide du vent. Ils ne sont pas adaptés au cas de la dispersion entomophile, voir par exemple pour le colza Klein (2000).
Le principe général est de considérer le grain de pollen comme une particule soumise à un champ de forces lors de sa trajectoire (modèles dits de type Lagrangien). Ces modèles de transports de particules sont issus de la mécanique des fluides. Ils conduisent à l'utilisation d'équations différentielles stochastiques et permettent des simulations numériques, par exemple à l'aide de schémas d'Euler. (Kloeden et Platen 1992)

En physique, il existe également des modèles dits de type Eulérien, basés sur l'étude de la turbulence et du transport de pollen. Ils conduisent à résoudre des équations aux dérivées partielles. En particulier, Loubet *et al* (2004) ont développé deux modèles pour l'étude du pollen de maïs : un modèle eulérien (AQUILON) et un modèle (SMOP) qui simule la trajectoire d'un ensemble de grains de pollen . Leur approche est basée sur la modélisation des processus de libération de transport et de dépôt du pollen. Elle prend en compte les caractéristiques de la structure du paysage (sol nu, autres cultures), la turbulence atmosphérique ainsi que les caractéristiques du grain de pollen de maïs. Les deux modèles permettent d'estimer la concentration et le dépôt en aval d'une source donnée.

L'utilisation des fonctions de dispersion individuelles efficaces pour l'étude du flux de pollen présente deux avantages sur les modèles physiques décrits ci-dessus :
- Les modèles physiques nécessitent des temps de calculs excessifs, dus à la simulation d'un grand nombre de trajectoires, contrairement à l'utilisation des fonctions de dispersion individuelles efficaces.
- Les fonctions de dispersion individuelles représentent la pollinisation efficace, comme leur nom l'indique, du pollen, alors que les modèles physiques ne prennent pas en compte des phénomènes de perte de viabilité du pollen au cours de leur trajectoire et des phénomènes de compétition. Ainsi, ils ne prédisent qu'une proportion de pollen issue d'une source et atteignant le champ étudié.

Dans toute la suite, on s'intéresse donc aux modèles dit Lagrangiens. On note $P_t = (X_t, Y_t, Z_t)$ la position à l'instant t d'un grain de pollen. On modélise $(P_t)_{t\geq 0}$ par un processus stochastique, en particulier par un processus de diffusion. C'est-à-dire que l'on peut écrire la trajectoire du grain de pollen sous la forme :

$$dP_t = b(P_t)dt + a(P_t)dW_t$$

où (W_t) est un mouvement brownien standard de dimension trois ici.

On note T_F le temps de fécondation, c'est-à-dire le temps d'arrêt de la trajectoire sur une plante femelle.

A partir de la modélisation de la trajectoire, il est possible d'obtenir une fonction de dispersion individuelle en cherchant la densité de probabilité du point de l'espace où la trajectoire rencontre un ovule qui est fécondé . Plus précisément, on a la propriété suivante :

Propriété 2.1 *On suppose que*
(i) T_F *est une variable aléatoire positive finie presque sûrement admettant une densité* f *sur* $\mathbb{R}^{+*}$.
(ii) le couple (X_t, Y_t) *admet une densité sur* $\mathbb{R}^2$ *pour* $t > 0$, *notée* $g_t(x, y)$.
(iii) Les variables T_F *et* (X_t, Y_t) *sont indépendantes.*
Alors le couple (X_{T_F}, Y_{T_F}) *est une variable aléatoire sur* $\mathbb{R}^2$ *admettant une densité* γ *égale à*

$$\gamma(x, y) = \int_0^{+\infty} g_t(x, y) f(t)\, dt \tag{2.3}$$

Démonstration :

Soit ϕ une fonction mesurable positive. D'après une formule de Bayes et le fait que les variables T_F et (X_t, Y_t) sont indépendantes, on a :

$$\begin{aligned} \mathrm{E}[\phi(X_{T_F}, Y_{T_F})] &= \int\int_{(x,y)\in\mathbb{R}^2} \int_{t=0}^{+\infty} \phi(x, y) g_T(x, y|T = t) f(t)\, dt\, dx\, dy \\ &= \int\int_{(x,y)\in\mathbb{R}^2} \phi(x, y) \left\{ \int_{t=0}^{+\infty} g_t(x, y) f(t)\, dt \right\} dx\, dy \end{aligned}$$

D'où le résultat.

On s'intéressera donc par la suite d'une part à la loi du couple (X_t, Y_t) et d'autre part à celle de la composante verticale de la trajectoire (Z_t), supposée indépendante des composantes horizontales .

Dans toute la suite de ce travail, on supposera :
- Les coordonnées de la trajectoire au temps 0 sont $P_0 = (0, 0, 0)$.
- La hauteur des fleurs mâles est prise comme origine de l'axe des z.
- Les fleurs femelles sont alors supposées être à une hauteur h avec $h < 0$.

Les modèles existants utilisés dans ce travail :

Ces modèles sont basés sur la modélisation de la trajectoire d'un grain de pollen à l'aide d'un mouvement brownien avec drift dans $\mathbb{R}^3$.

$$\begin{cases} dX_t = f_x dt + \tau_x dB_t^1 \\ dY_t = f_y dt + \tau_y dB_t^2 \\ dZ_t = f_z dt + \tau_z dB_t^3 \end{cases} \tag{2.4}$$

où les $(B_t^i)_{i=1,2,3}$ sont trois mouvements browniens indépendants ;
τ_x, τ_y, τ_z sont positifs.
f_x et f_y représentent les vitesses moyennes du vent dans le plan horizontal.
f_z, supposé négatif, représente la vitesse de chute due à la gravité, appelée aussi vitesse de sédimentation.
La matrice de variance-covariance représente les turbulences dues au vent.
La Figure (2.3) ci-après représente la trajectoire d'un grain de pollen partant du point $(0,0,0)$ et s'arrêtant sur le plan d'équation $\{z = h\}$, où h est la hauteur des plantes femelles.

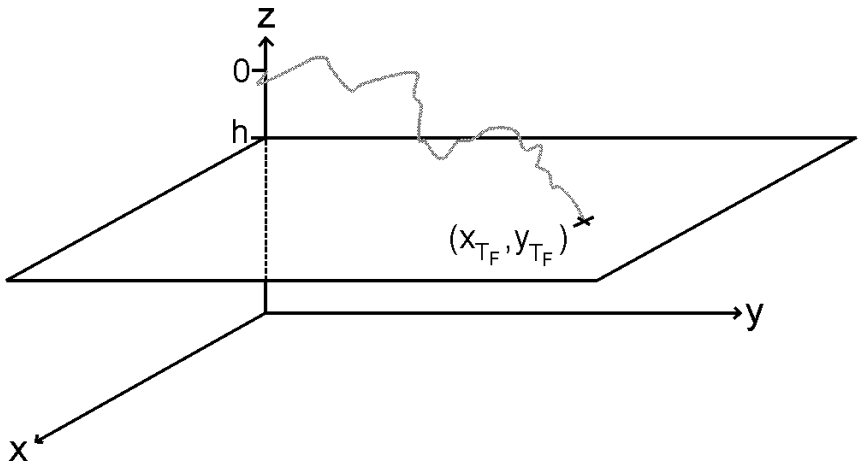


FIG. 2.3 – Une trajectoire d'un grain de pollen

Dans ce cas, le couple (X_t, Y_t) admet pour loi une loi normale de moyenne $(f_x t, f_y t)$ et de matrice de covariance $\begin{pmatrix} \tau_x^2 t & 0 \\ 0 & \tau_y^2 t \end{pmatrix}$.

Klein *et al* (2003) ont faits trois hypothèses différentes pour la loi du temps de fécondation T_F, conduisant à trois familles de fonctions de dispersion individuelles détaillées au paragraphe 2.4.2.

1. Un temps aléatoire indépendant de la trajectoire et de loi exponentielle est introduit pour modéliser le temps d'atteinte sur n'importe quelle partie végétale.

Ensuite, le temps de fécondation est obtenu en conditionnant cette variable par l'événement : " le grain de pollen atteint la hauteur h des fleurs femelles".

2. Le temps de fécondation est défini comme le temps de premier passage de la trajectoire (du grain de pollen) au niveau $z = h$. Sa loi est connue d'après les propriétés du mouvement brownien.
3. Le temps de fécondation est une généralisation des deux cas précédents.

Dans la prochaine section, on s'intéresse à des lois de probabilités utilisées par la suite : les lois hyperboliques généralisées (GHD) dans $\mathbb{R}^2$. Les résultats énoncés seront utilisés ensuite.

2.4.1 Loi hyperbolique généralisée (GHD)

La définition de la loi hyperbolique généralisée en dimension 1 est décrite en détails à l'annexe A. De plus, certaines propriétés caractéristiques y sont présentées. Ici, on s' intéresse au cas de la dimension 2.

Définition 2.3 *Une variable aléatoire suit une loi Inverse Gaussienne Généralisée, notée GIG, si elle admet pour fonction de densité*

$$f_{GIG}(\alpha, \rho, \eta; t) = \frac{1}{I(\alpha, \rho, \eta)} t^{-\alpha} e^{-\rho t - \frac{\eta}{t}} \mathbb{I}_{t \geq 0} \tag{2.5}$$

où

$$I(\alpha, \rho, \eta) = 2 \left(\frac{\eta}{\rho}\right)^{(1-\alpha)/2} \mathcal{K}_{1-\alpha}(2\sqrt{\rho\eta})$$

α est un réel et η et ρ sont positifs.

$K_\nu(x)$ représente la fonction modifiée de Bessel de troisième ordre (pour une définition détaillée, par exemple : Abramovitz et Stegun 1972).

Définition 2.4 *Une variable aléatoire suit une loi Hyperbolique Généralisée de dimension 2 (GHD) si elle admet pour fonction de densité sur $\mathbb{R}^2$:*

$$f_{GHD}(\alpha, \lambda_z, \lambda_x, \lambda_y, \delta_x, \delta_y; x, y) = \frac{\lambda_z^{1-\alpha} \delta_x \delta_y (p/q(x,y))^{\frac{\alpha}{2}}}{2\pi} \frac{\mathcal{K}_\alpha(\sqrt{pq(x,y)})}{\mathcal{K}_{1-\alpha}(\lambda_z)} e^{\delta_x \lambda_x x + \delta_y \lambda_y y} \tag{2.6}$$

où $p = \lambda_z^2 + \lambda_x^2 + \lambda_y^2$ et $q(x, y) = 1 + \delta_x^2 x^2 + \delta_y^2 y^2$.

Deux résultats essentiels utilisés par la suite dans les calculs sont (par exemple Prudnikov *et al* 1986) :

Propriété 2.2 *: Pour tout réel ν et pour tous réels positifs p, q, on a*

$$\int_0^{+\infty} t^{\nu-1} \exp\left(-pt - \frac{q}{t}\right) dt = 2\left(\frac{q}{p}\right)^{\nu/2} \mathcal{K}_\nu(2\sqrt{pq}) \tag{2.7}$$

Propriété 2.3 *: Pour tout entier n positif, on a*

$$\mathcal{K}_{n+1/2}(s) = \sqrt{\frac{\pi}{2}} s^{-1/2} e^{-s} \left[1 + \sum_{i=1}^{n} \frac{(n+i)!}{(n-i)!i!}(2s)^{-i}\right]$$

En particulier, on a :

$$\mathcal{K}_{1/2}(s) = \sqrt{\frac{\pi}{2}} s^{-1/2} e^{-s} \text{ (cas particulier) et } \mathcal{K}_{3/2}(s) = \sqrt{\frac{\pi}{2}} s^{-1/2} e^{-s}(1 + s^{-1}) \tag{2.8}$$

Les résultats énoncés dans la prochaine section (2.4.2) découlent de la propriété suivante :

Propriété 2.4 *On suppose :*

- *la variable aléatoire T admet pour loi une GIG de paramètres (α, ρ, η).*
- *(X_t, Y_t) est un vecteur gaussien de moyenne $(f_x t, f_y t)$ et de matrice de variance*

$$\begin{pmatrix} \tau_x^2 t & 0 \\ 0 & \tau_y^2 t \end{pmatrix}$$

- *Les variables aléatoires T et (X_t, Y_t) sont indépendantes.*

Alors le couple (X_T, Y_T) admet pour densité une GHD de vecteur paramètre $\theta = (\alpha, \lambda_x, \lambda_y, \lambda_z, \delta_x, \delta_y)$ avec $\lambda_z = \sqrt{\eta'\rho'}$, $\lambda_x = \dfrac{f_x\sqrt{\eta'}}{\tau_x}$, $\lambda_y = \dfrac{f_y\sqrt{\eta'}}{\tau_y}$,
$\delta_x = \dfrac{1}{\tau_x\sqrt{\eta'}}$ $\delta_y = \dfrac{1}{\tau_y\sqrt{\eta'}}$, *où $\eta' = 2\eta$ et $\rho' = 2\rho$.*

Démonstration :

Les variables aléatoires T et (X_t, Y_t) étant indépendantes, on peut utiliser une formule de Bayes et on a

$$f_{(X_T,Y_T)}(x, y) = f(x, y) = \int_0^{+\infty} f_{(X_T,Y_T)|T=t}(x, y) f_T(t) dt$$

où f_T est la densité de T définie en (2.5).

$$f(x,y) = \int_0^{+\infty} \frac{1}{2\pi\tau_x\tau_y t} \exp\left(-\frac{(x-f_x t)^2}{2\tau_x^2 t} - \frac{(y-f_y t)^2}{2\tau_y^2 t}\right) \times \frac{1}{I(\alpha,\rho,\eta)} t^{-\alpha} e^{-\rho t - \frac{\eta}{t}} dt$$

En développant les termes situés dans l'exponentielle et en les regroupant, on obtient :

$$f(x,y) = A(x,y) \int_0^{+\infty} t^{(-\alpha-1)} \exp\left(-tp' - \frac{q'}{t}\right) dt \tag{2.9}$$

avec

$$p' = \rho + \frac{f_x^2}{2\tau_x^2} + \frac{f_x^2}{2\tau_x^2} \text{ et } q'(x,y) = \eta + \frac{x^2}{2\tau_x^2} + \frac{y^2}{2\tau_y^2}$$

$$\begin{aligned} A(x,y) &= \frac{1}{2\pi I(\alpha,\rho,\eta)\tau_x\tau_y} \exp\left(\frac{f_x}{\tau_x^2}x + \frac{f_y}{\tau_y^2}y\right) \\ &= \frac{1}{4\pi\tau_x\tau_y} \frac{(\eta/\rho)^{(\alpha-1)/2}}{\mathcal{K}_{1-\alpha}(2\sqrt{\rho\eta})} \exp\left(\frac{f_x}{\tau_x^2}x + \frac{f_y}{\tau_y^2}y\right) \end{aligned}$$

On écrit alors $\eta = \frac{\eta'}{2}$ et $\rho = \frac{\rho'}{2}$. D'où

$$p' = \frac{1}{2\eta'}\left(\rho'\eta' + \frac{f_x^2\eta'}{2\tau_x^2} + \frac{f_y^2\eta'}{2\tau_y^2}\right) = \frac{1}{2\eta'}p$$

$$\text{et } q'(x,y) = \frac{\eta'}{2}\left(1 + \frac{x^2}{2\tau_x^2\eta'} + \frac{y^2}{2\tau_y^2\eta'}\right) = \frac{\eta'}{2}q$$

En effectuant dans l'intégrale le changement de variable $u = \frac{t}{\eta'}$, en utilisant la formule (2.27) et le fait que pour tout $x \in \mathbb{R}$, $\mathcal{K}_\alpha(x) = \mathcal{K}_{-\alpha}(x)$, on obtient :

$$f(x,y) = \frac{(\eta'/\rho')^{(\alpha-1)/2}\,\eta'^{(-\alpha)}}{2\pi\tau_x\tau_y\mathcal{K}_{1-\alpha}(\sqrt{\rho'\eta'})} \times \exp\left(\frac{f_x}{\tau_x^2}x + \frac{f_y}{\tau_y^2}y\right) \times \left(\frac{p}{q}\right)^{\alpha/2} \mathcal{K}_\alpha(\sqrt{pq})$$

On définit alors $\lambda_z = \sqrt{\eta'\rho'}$, $\lambda_x = \frac{f_x\sqrt{\eta'}}{\tau_x}$, $\lambda_y = \frac{f_y\sqrt{\eta'}}{\tau_y}$, $\delta_x = \frac{1}{\tau_x\sqrt{\eta'}}$ et $\delta_y = \frac{1}{\tau_y\sqrt{\eta'}}$.

D'où $p = \lambda_z^2 + \lambda_x^2 + \lambda_y^2$ et $q(x,y) = 1 + \delta_x^2x^2 + \delta_y^2y^2$.

Finalement,

$$f(x,y;\theta) = \frac{\lambda_z^{1-\alpha}\delta_x\delta_y(p/q(x,y))^{\frac{\alpha}{2}}}{2\pi} \frac{\mathcal{K}_\alpha(\sqrt{pq(x,y)})}{\mathcal{K}_{1-\alpha}(\lambda_z)} e^{\delta_x\lambda_x x + \delta_y\lambda_y y}$$

et on obtient le résultat.

2.4.2 Mouvement brownien avec drift dans $\mathbb{R}^3$

Dans cette section, on rappelle en détail les modèles proposés par Klein *et al* (2003).

a) Modèle 1, GTM : prédominance de la végétation

La trajectoire s'arrête au bout d'un temps aléatoire T_e suivant une loi exponentielle de paramètre λ (positif) et indépendant de la trajectoire.
Cette hypothèse correspond à une végétation dense qui limite les trajectoires. Le grain de pollen a une probabilité infinitésimale de s'arrêter sur une partie végétale. Les trajectoires qui fécondent effectivement un ovule sont celles qui s'arrêtent à la hauteur des fleurs femelles.
Ainsi, on prend pour T_F la loi conditionnelle du temps d'arrêt T_e conditionnellement à l'événement $\{Z_{T_e} = h\}$.
T_F suit alors une GIG de paramètres $\frac{1}{2}$, $\lambda + \frac{f_z^2}{2\tau_z^2}$ et $\frac{h^2}{2\tau_z^2}$ (Klein *et al*, 2003).

Proposition 2.1 *La fonction de dispersion individuelle γ est alors la densité d'une GHD avec les paramètres*

$$\alpha = \frac{1}{2},\ \delta_x = \frac{\tau_z}{\tau_x|h|},\ \delta_y = \frac{\tau_z}{\tau_y|h|},\ \lambda_x = \frac{f_x h}{\tau_x \tau_z},\ \lambda_y = \frac{f_y h}{\tau_y \tau_z},\ \lambda_z^2 = \left(2\lambda + \frac{f_z^2}{\tau_z^2}\right)\frac{h^2}{\tau_z^2}$$

Démonstration : Il suffit d'appliquer la propriété (2.4).

Définition 2.5 *Cette loi est appelée GTM (Generalized Tufto Model) et une écriture simplifiée, obtenue en utilisant (2.8) est*

$$f_{GTM}(\lambda_z, \lambda_x, \lambda_y, \delta_x, \delta_y; x, y) = \frac{\delta_x \delta_y \lambda_z e^{\lambda_z}}{2\pi} \frac{e^{-\sqrt{pq(x,y)}}}{\sqrt{q(x,y)}} e^{\delta_x \lambda_x x + \delta_y \lambda_y y}$$

avec $p = \lambda_z^2 + \lambda_x^2 + \lambda_y^2$ *et* $q(x,y) = 1 + \delta_x^2 x^2 + \delta_y^2 y^2$

Le vecteur paramètre $\theta = (\delta_x, \delta_y, \lambda_x, \lambda_y, \lambda_z)$ appartient à $\Theta = (\mathbb{R}^+)^2 \times (\mathbb{R})^2 \times \mathbb{R}^+$.

b) Modèle 2, NIG : prédominance du sol

On considère que la végétation n'arrête pas les trajectoires et on définit le temps de fécondation comme le temps de premier passage de la trajectoire à la hauteur des fleurs femelles. On a donc $T_F = T_h = inf\{t > 0, Z_t = h\}$.
Dans ce cas, il est connu que T_h suit une GIG de paramètres $\frac{3}{2}$, $\frac{f_z^2}{2\tau_z^2}$ et $\frac{h^2}{2\tau_z^2}$ (voir par exemple Karatzas et Shreeve (1991) ou la démonstration donnée à l'annexe A).

Proposition 2.2 *La fonction de dispersion individuelle γ est alors la densité d'une GHD avec les paramètres*

$$\alpha = \frac{3}{2}, \delta_x = \frac{\tau_z}{\tau_x|h|},\ \delta_y = \frac{\tau_z}{\tau_y|h|},\ \lambda_x = \frac{f_x h}{\tau_x\tau_z},\ \lambda_y = \frac{f_y h}{\tau_y\tau_z},\ ,\ \lambda_z = \frac{f_z h}{\tau_z^2}$$

Démonstration : On applique une nouvelle fois la propriété (2.4).

Définition 2.6 *Cette loi est appelée NIG (Normal Inverse Gaussian) et une écriture simplifiée, obtenue en utilisant (2.8) est*

$$f_{NIG}(\lambda_z, \lambda_x, \lambda_y, \delta_x, \delta_y; x, y) = \frac{\delta_x\delta_y e^{\lambda_z}}{2\pi}\frac{(q(x,y)^{-1/2} + p^{1/2})}{q(x,y)}e^{-\sqrt{pq(x,y)}}e^{\delta_x\lambda_x x + \delta_y\lambda_y y}$$

avec $p = \lambda_z^2 + \lambda_x^2 + \lambda_y^2$ et $q(x,y) = 1 + \delta_x^2 x^2 + \delta_y^2 y^2$.

Le vecteur paramètre $\theta = (\delta_x, \delta_y, \lambda_x, \lambda_y, \lambda_z)$ appartient à $\Theta = (\mathbb{R}^+)^2 \times (\mathbb{R})^2 \times \mathbb{R}^+$.

c) Modèle 3, GHD : généralisation

Cette fois-ci le temps de fécondation est supposé suivre une GIG sans valeur particulière pour le premier paramètre α. Alors, γ est une GHD de paramètres

$$\alpha,\ \delta_x = \frac{\tau_z}{\tau_x|h|},\ \delta_y = \frac{\tau_z}{\tau_y|h|},\ \lambda_x = \frac{f_x h}{\tau_x\tau_z},\ \lambda_y = \frac{f_y h}{\tau_y\tau_z},\ \lambda_z = \frac{f_z h}{\tau_z^2}$$

Le vecteur paramètre $\theta = (\delta_x, \delta_y, \lambda_x, \lambda_y, \lambda_z, \alpha)$ appartient à
$\Theta = (\mathbb{R}^+)^2 \times (\mathbb{R})^2 \times \mathbb{R}^+ \times \mathbb{R}$.

Ce modèle englobe les deux modèles précédents. En effet, les trois modèles diffèrent seulement dans la valeur du paramètre α : pour le modèle 1, GTM, on a $\alpha = 1/2$; pour le modèle 2, NIG, on a $\alpha = 3/2$; et pour le modèle 3, GHD, le paramètre α est libre. De plus, ce paramètre est lié à l'influence de la végétation sur la trajectoire des grains de pollen (plus il sera proche de 1.5, moins la végétation aura d'influence sur la trajectoire par exemple.)

2.4.3 Passage en coordonnées polaires

On va effectuer le calcul pour le modèle 2, NIG.

La fonction de dispersion individuelle associée au modèle 2 est, d'après 2.4.2 b) :

$$f_{NIG}(x, y; \theta) = \frac{\delta^2 e^{\lambda_z}}{2\pi}\frac{(q(x,y)^{-1/2} + p^{1/2})}{q(x,y)}e^{-\sqrt{pq(x,y)}}e^{\delta(\lambda_x x + \lambda_y y)}$$

avec $p = \lambda_z^2 + \lambda_x^2 + \lambda_y^2$, $q(x,y) = 1 + \delta^2(x^2 + y^2)$ et θ le vecteur des paramètres.

Ici, on a supposé aussi $\delta_x = \delta_y = \delta$.

On pose $x = r\cos\omega$ et $y = r\sin\omega$ avec $r \geq 0$ et $\omega \in [0, 2\pi[$.

Alors, on a

$$f_{NIG}(x, y; \theta)dxdy = C(\theta)\frac{q(r)^{-1/2} + p^{1/2}}{q(r)} \exp(-\sqrt{pq(r)} + \delta r(\lambda_x \cos\omega + \lambda_y \sin\omega))rdrd\theta$$

avec $q(r) = 1 + \delta^2 r^2$.

On peut écrire $\lambda_x \cos\omega + \lambda_y \sin\omega = (\lambda_x^2 + \lambda_y^2)\cos(\omega - \omega_1)$ avec $\omega_1 \in [0, 2\pi[$ tel que $\cos\omega_1 = \dfrac{\lambda_x}{\sqrt{\lambda_x^2 + \lambda_y^2}}$ et $\sin\omega_1 = \dfrac{\lambda_y}{\sqrt{\lambda_x^2 + \lambda_y^2}}$.

Alors f_{NIG} s'écrit sous la forme :

$$f_{NIG}(r\cos\omega, r\sin\omega; \theta) = C(\theta)g_\theta(r)\exp(\kappa r\cos(\omega - \omega_1))$$

avec $\kappa = \delta(\lambda_x^2 + \lambda_y^2)$, C une fonction ne dépendant que des paramètres du modèle, g une fonction dépendant de r et du vecteur paramètre θ.

On constate que la loi de Ω conditionnellement à $\{R = r\}$ suit une distribution de Von Mises (Batschelet 1981) et donc la direction du vent rentre en compte dans l'expression de la fonction de dispersion individuelle. Pour ces trois modèles, la pollinisation n'est donc pas isotrope (contrairement au colza par exemple).

D'autre part, sous cette forme, on constate que les variables aléatoires R et Ω ne sont pas indépendantes, contrairement à l'hypothèse faite par Tufto *et al* (1997) qui supposent R et Ω indépendantes pour leur modèle.

Remarque : Le même raisonnement peut s'appliquer pour le modèle 1, GTM, et le modèle 3, GHD. Les fonctions s'écrivent également comme le produit d'une fonction ne dépendant que de r et de la densité d'une loi de Von Mises.

2.5 Nouveaux modèles proposés

2.5.1 Étude d'un processus d'Ornstein-Uhlenbeck intégré

On définit (V_t) par $dV_t = -fV_t dt + \tau dB_t$, $V_0 = \eta$ avec η indépendante de $(B_t)_t$, $f > 0$ et $\tau > 0$. (V_t) est donc un processus d'Ornstein-Uhlenbeck.

Il est connu que la solution de cette équation différentielle stochastique (obtenue en appliquant la formule d'Itô) est de la forme

$$V_t = e^{-ft}\eta + \tau e^{-ft}\int_0^t e^{fs}dB_s$$

On définit alors le processus $(X_t)_{t\geq 0}$ par $X_t = \displaystyle\int_0^t V_s ds$.

Définition 2.7 *$(X_t)_{t\geq 0}$ est appelé un processus d'Ornstein-Uhlenbeck intégré, noté par la suite O-U intégré.*

On s'intéresse au cas d'un processus non stationnaire : on suppose donc que $V_0 = v_0$.

Proposition 2.3 *:*
Le processus X_t est égal à $X_t = v_0\dfrac{1-e^{-ft}}{f} + \dfrac{\tau}{f}\displaystyle\int_0^t \left(1-e^{-f(t-u)}\right) dB_u$.
De plus, on a $E(X_t) = m_t(f, v_0) = v_0\dfrac{1-e^{-ft}}{f}$ *et*
$Var(X_t) = \sigma_t^2(f,\tau) = \dfrac{\tau^2}{f^2}\left(t + \dfrac{4e^{-ft} - 3 - e^{-2ft}}{2f}\right)$

Démonstration :
En appliquant le théorème de Fubini pour les intégrales stochastiques (Protter 1992 page 159), on a

$$\begin{aligned}
X_t &= \int_0^t \left\{ e^{-fs}v_0 + \tau e^{-fs}\int_0^s e^{fu} dB_u \right\} ds \\
X_t &= v_0\int_0^t e^{-fs}\, ds + \tau\int_0^t e^{-fs}\left\{\int_0^s e^{fu} dB_u\right\} ds \\
X_t &= v_0\frac{1-e^{-ft}}{f} + \tau\int_0^t dB_u\, e^{fu}\left\{\int_u^t e^{-fs}\, ds\right\} \\
X_t &= v_0\frac{1-e^{-ft}}{f} + \frac{\tau}{f}\int_0^t \left(1-e^{-f(t-u)}\right) dB_u \qquad (2.10)
\end{aligned}$$

De cette écriture, on en déduit directement l'espérance de X_t donnée ci-dessus.
Pour la variance, on a

$$\begin{aligned}
Var(X_t) &= \frac{\tau^2}{f^2}\int_0^t \left(1-e^{-f(t-s)}\right)^2 ds \\
&= \frac{\tau^2}{f^2}\int_0^t \left(1 + e^{-2f(t-s)} - 2e^{-f(t-s)}\right) ds
\end{aligned}$$

d'où le résultat.

Pour les deux modèles proposés ci-dessous, un grain de pollen est toujours considéré comme une particule soumise à un champ de forces. Certaines composantes sont modélisées de façon plus précise : en effet, suivant le cas, on s'intéresse à la modélisation des composantes du vecteur vitesse et non plus aux composantes de la trajectoire.

2.5.2 Modélisation de la vitesse dans le plan vertical

Dans ce premier modèle, les composantes horizontales de la trajectoire (X_t, Y_t) restent deux mouvements browniens avec drift indépendants. La loi du couple (X_t, Y_t) est donc inchangée.
Par contre, pour la composante verticale Z_t, en se basant sur les équations de Langevin (provenant de l'équation des forces en physique), on suppose que la vitesse verticale V_t est modélisée par :

$$\begin{cases} dZ_t = V_t dt \\ dV_t = f_z dt - \beta V_t dt + F_t dt \end{cases}$$

f_z est une force résultant de la gravité et F_t représente les contributions de la force exercée par le fluide (ici le vent) sur la particule durant son trajet, qui n'ont pas déjà été mises dans le terme linéaire de friction $-\beta V_t$. F_t est modélisé en général par un mouvement brownien standard.

La trajectoire $(P_t)_{t\geq 0}$ de ce modèle, appelé modèle 4, est donc modélisée par

$$\begin{cases} dX_t = f_x dt + \tau_x dB_t^1 \\ dY_t = f_y dt + \tau_y dB_t^2 \\ dZ_t = V_t dt \end{cases} \tag{2.11}$$

avec

$$\begin{cases} dV_t = (c_z - \beta V_t)dt + \tau_z dB_t^3 \\ V_0 = v_0 = 0 \end{cases} \tag{2.12}$$

où les $(B_t^i)_{i=1,2,3}$ sont 3 mouvements browniens indépendants (et $P_0 = (0,0,0)$) ;
τ_x, τ_y, τ_z sont positifs, $\beta > 0$ et $c_z < 0$.
Il est à noter que τ_z n'a pas tout à fait la même signification physique que pour les modèles 1 à 3.

Lemme 2.1 *Pour $\beta > 0$, l'équation (2.12) admet pour solution*

$$V_t = \frac{c_z}{\beta}(1 - e^{-\beta t}) + \tau_z e^{-\beta t} \int_0^t e^{\beta s} dB_s$$

Démonstration :
On pose $S_t = \exp(\beta t) V_t$. La formule d'Itô donne alors :

$$dS_t = c_z \exp(\beta t) dt + \tau_z \exp(\beta t) dB_t$$

d'où $S_t = v_0 + \frac{c_z}{\beta}(\exp(\beta t) - 1) + \tau_z \int_0^t \exp(\beta s) dB_s$. On en déduit le résultat. ◇

D'après le calcul effectué en à la proposition 2.3, Z_t peut s'écrire sous la forme :

$$Z_t = \frac{c_z}{\beta}t - \frac{c_z}{\beta}\frac{1-e^{-\beta t}}{\beta} + \frac{\tau_z}{\beta}\int_0^t \left(1-e^{-\beta(t-s)}\right) dB_s \tag{2.13}$$

On suppose toujours que le temps de fécondation est le premier temps de passage de la trajectoire à la hauteur h. On pose donc $\nu_h = inf\{t > 0, Z_t \leq h\}$.

Propriété 2.5 *Sous l'hypothèse que $c_z < 0$ et $h < 0$, le temps d'arrêt $\nu_h = inf\{t > 0, Z_t \leq h\}$, où Z_t est donné en 2.13, est fini presque sûrement.*

Démonstration : On a

$$\begin{aligned}\mathbb{P}(\nu_h = +\infty) &= \mathbb{P}(\forall t, Z_t > h) = \mathbb{P}(\lim_{t\to+\infty} \searrow \bigcap_{u\leq t}\{Z_u > h\}) \\ &= \lim_{t\to+\infty} \mathbb{P}(\bigcap_{u\leq t}\{Z_u > h\}) \leq \lim_{t\to+\infty}\mathbb{P}(Z_t > h)\end{aligned}$$

En utilisant l'égalité (2.13) et la proprosition 2.3, on a $\mathrm{E}(Z_t) = \frac{c_z t}{\beta}(1+O(1))$ et quand t tend vers l'infini $h - \mathrm{E}(Z_t) > 0$. En appliquant l'inégalité de Bienaymé-Tchebytchev, on obtient alors

$$\mathbb{P}(Z_t > h) = \mathbb{P}(Z_t - \mathrm{E}(Z_t) > h - \mathrm{E}(Z_t)) \leq \frac{Var(Z_t)}{(h-\mathrm{E}(Z_t))^2}$$

Toujours, d'après la proposition 2.3, on obtient $\mathrm{E}(Z_t) \sim \frac{c_z}{\beta}t$ et $\mathrm{Var}(Z_t) \sim \frac{\tau_z^2}{c_z^2}t$ d'où $\lim_{t\to+\infty}\mathbb{P}(Z_t > h) = 0$ et le résultat. ◇

Ici, ce n'est plus la densité du temps de premier passage à un niveau h pour un mouvement brownien avec drift qui nous intéresse. On souhaite, en fait, la densité du temps de premier passage à un niveau h pour un processus d'Ornstein-Uhlenbeck intégré. La détermination de la fonction de densité de ce processus a été étudiée depuis plusieurs années, mais il n'existe pas de solutions exactes analytiques. Souvent les résultats obtenus donnent des informations sur les moments ou sur la transformée de Laplace de la densité du premier temps de passage, mais pas sur l'expression explicite de la densité. C'est pour cela que l'on va approcher la densité de ν_h.

Calcul de la fonction de dispersion individuelle γ :

Théorème 2.1 *On suppose que le processus Z_t est défini par 2.13 avec $c_z < 0$ et $h < 0$. Alors la densité du temps d'arrêt ν_h, définie sur $\mathbb{R}^+$, peut être approchée par*

$$p_\theta(t) = \frac{\beta g'(\beta t)}{\sqrt{2\pi g(\beta t)}}\left[\frac{b_z H(\beta t)+c_z}{g(\beta t)} - \frac{b_z H'(\beta t)}{g'(\beta t)}\right]\exp\left(-\frac{(b_z H(\beta t)+c_z)^2}{2g(\beta t)}\right) \tag{2.14}$$

où $b_z = \frac{-c_z}{\tau_z\sqrt{\beta}}, c_z = \frac{-\beta^{3/2}h}{\tau_z}$, $H(t) = 1-e^{-t}-t$ et $g(t) = t - 1.5 + 2e^{-t} - 0.5e^{-2t}$.

Démonstration :
Pour commencer, à l'aide d'un changement de temps, on se ramène à l'approximation de la densité du premier temps de passage d'un mouvement brownien traversant une courbe dépendant du temps.

On voit facilement que l'on peut écrire ν_h sous la forme :
$\nu_h(Y) = inf\{t > 0, Y_t \leq f_\theta(t)\}$ où $Y_t = \int_0^t \left(1 - e^{-\beta(t-s)}\right) dB_s$ et
$f_\theta(t) = \frac{\beta h}{\tau_z} - \frac{c_z}{\tau_z} t + \frac{c_z}{\tau_z \beta}(1 - e^{-\beta t})$.

On définit le processus $\tilde{Y}_t = \int_0^t \left(1 - e^{-(t-u)}\right) dB_u$. Alors, on a $Y_t \stackrel{\mathcal{L}}{=} \frac{1}{\sqrt{\beta}} \tilde{Y}_{\beta t}$. D'où

$$\nu_h(Y) \stackrel{\mathcal{L}}{=} inf\{t > 0, \tilde{Y}_{\beta t} \leq \sqrt{\beta} f_\theta(t)\} = \frac{1}{\beta} inf\{u > 0, \tilde{Y}_u \leq \tilde{f}_\theta(u)\} = \frac{1}{\beta} \tilde{\nu}(\tilde{Y})$$

avec $\tilde{f}_\theta(u) = \bar{b}_z H(u) + \bar{c}_z$ où $\bar{b}_z = \frac{c_z}{\tau_z \sqrt{\beta}}$ et $\bar{c}_z = \frac{\beta^{3/2} h}{\tau_z}$ et $H(u) = (1 - e^{-u} - u)$.

$(\tilde{Y}_t)$ est une martingale locale continue avec $\tilde{Y}_0 = 0$ car c'est une intégrale d'Itô. De plus,

$$g(t) = < \tilde{Y} >_t = \int_0^t \left(1 - e^{-(t-s)}\right)^2 ds = t + \frac{4e^{-t} - 3 - e^{-2t}}{2}$$

(et donc on a $\lim_{t \to +\infty} < \tilde{Y} >_t = +\infty$.)
On définit $T(s) = inf\{t > 0, < \tilde{Y} >_t > s\}$. g réalisant une bijection croissante de $[0, +\infty[$ sur $[0, +\infty[$, on a

$$T(s) = inf\{t > 0, g(t) > s\} = g^{-1}(s).$$

Par un théorème de changement de temps (par exemple Rogers et Williams (1994), page 64), on a alors $\tilde{Y}_{T(s)} \stackrel{\mathcal{L}}{=} \tilde{B}_s$ où $\tilde{B}_s$ est un mouvement brownien standard.

Par suite,

$$inf\{s > 0, \tilde{Y}_{T(s)} \leq \tilde{f}_\theta(g^{-1}(s))\} \stackrel{\mathcal{L}}{=} inf\{s > 0, \tilde{B}_s \leq \tilde{f}_\theta(g^{-1}(s))\} = g(\tilde{\nu}(\tilde{Y}))$$

donc $\tilde{\nu}(\tilde{Y}) \stackrel{\mathcal{L}}{=} g^{-1}\left(inf\{s > 0, \tilde{B}_s \leq \tilde{f}_\theta(g^{-1}(s))\}\right)$.

En utilisant la propriété du mouvement brownien $B_t \stackrel{\mathcal{L}}{=} -B_t$ et en posant
$a_\theta(s) = -\tilde{f}_\theta(g^{-1}(s)) = -\bar{b}_z H(g^{-1}(s)) + \bar{c}_z = b_z H(g^{-1}(s)) + c_z$, on en déduit que

$$\nu_h(Y) \stackrel{\mathcal{L}}{=} \frac{1}{\beta} g^{-1}\left(inf\{s > 0, \tilde{B}_s \geq a_\theta(s)\}\right) \tag{2.15}$$

On va maintenant donner une approximation de la densité du temps d'atteinte $inf\{s > 0, \tilde{B}_s \geq a_\theta(s)\}$, l'approximation étant notée $r_\theta(t)$.

La fonction a_θ définie est C^1 sur l'intervalle $[0, +\infty)$ avec $a_\theta(0) = c_z > 0$ (car $h < 0, c_z < 0$ et $\beta > 0$). De plus, a est concave (g étant convexe).

On peut alors approcher la densité du temps de premier passage d'un mouvement brownien par une courbe dépendant du temps à l'aide d'un théorème dû à Durbin (1992). On obtient alors

$$r_\theta(t) = \frac{1}{\sqrt{2\pi t}} \left(\frac{a_\theta(t)}{t} - a'_\theta(t) \right) \exp\left(-\frac{a_\theta(t)^2}{2t} \right)$$

(L'erreur commise lors de l'approximation est inférieure à $\sqrt{\pi}/\Gamma(0.5) \times e(t)$ où $e(t)$ représente le maximum de $|\frac{a(r)-a(s)}{r-s} - a'(r)|$ pour $0 < s < r \leq t$. Ainsi plus la pente de la fonction a varie lentement, plus la convergence vers la vraie fonction de densité est rapide.)

On en déduit alors la densité $p(t)$ de ν_h à l'aide de l'égalité (2.15) :

$$p_\theta(t) = \frac{\beta g'(\beta t)}{\sqrt{2\pi g(\beta t)}} \left[\frac{b_z H(\beta t) + c_z}{g(\beta t)} - \frac{b_z H'(\beta t)}{g'(\beta t)} \right] \exp\left(-\frac{(b_z H(\beta t) + c_z)^2}{2g(\beta t)} \right)$$

◇

On peut alors obtenir une fonction de dispersion individuelle :

Proposition 2.4 *On suppose :*
- la trajectoire $(P_t)_{t\geq 0}$ est définie par (3.5) avec $\tau_x = \tau_y = \tau$.
- le temps de fécondation vérifie $T_F = \nu_h$ et on approche sa fonction de densité par la fonction p_{θ_z} définie ci-dessus.
Alors, la loi du point de fécondation (X_{T_F}, Y_{T_F}) admet pour densité
$\gamma(\theta; x, y) = \dfrac{\lambda^2}{\pi} \exp\left(2\lambda(\delta_x x + \delta_y y)\right)$

$$\int_0^{+\infty} \frac{1}{t} \exp\left(-(\delta_x^2 + \delta_t^2)t - \frac{\lambda^2(x^2+y^2)}{t} \right) p_1(t, \theta_z) dt$$

où $p_1(t, \theta_z) = p_{\theta_z}(\dfrac{t}{\beta})$ *et* $\theta_z = (b_z, c_z)$, $\theta = (\delta_x, \delta_y, \lambda, b_z, c_z)$, *avec* $\delta_x^2 = \dfrac{f_x^2}{2\tau^2\beta}$,
$\delta_y^2 = \frac{f_y^2}{2\tau^2\beta}$, $\lambda^2 = \dfrac{\beta}{2\tau^2}$ *et* $\theta \in \mathbb{R}^2 \times (\mathbb{R}^+)^3$.

Démonstration :
Il suffit de constater que X_t suit une loi normale de moyenne $f_x t$ et de variance $\tau^2 t$; Y_t une loi normale de moyenne $f_y t$ et de variance $\tau^2 t$ et ν_h admet pour densité $p_{\theta_z}(t)$. Alors, en appliquant l'égalité (2.3) et en faisant le changement de variable $t = \beta u$, on obtient le résultat.

2.5.3 Modélisation du vecteur vitesse dans le plan horizontal

Le deuxième modèle proposé est basé sur la modélisation des composantes du vecteur vitesse dans le plan horizontal par deux processus d'Ornstein-Uhlenbeck indépendants. Cela permet de ne plus considérer constants les paramètres décrivant la vitesse et la direction du vent tout au long de la trajectoire du grain de pollen (à la différence des modèles basés sur trois mouvements browniens avec drift).
D'autre part, contrairement à Tufto *et al* (1997), ces processus ne sont pas supposés stationnaires. Cela permet d'introduire une vitesse de vent minimale à laquelle les grains de pollen sont émis.

La composante verticale Z_t est à nouveau modélisée par un mouvement brownien avec drift, indépendant de (X_t, Y_t).

La trajectoire $(P_t)_{t\geq 0}$ s'écrit alors (dans un premier temps) sous la forme :

$$\begin{cases} dX_t = V_t^x dt \\ dY_t = V_t^y dt \\ dZ_t = f_z dt + \tau_z dB_t^3 \end{cases} \tag{2.16}$$

avec $\begin{cases} dV_t^x = -c_x V_t^x dt + \tau_x dB_t^1 \\ V_0^x = v_0^x \end{cases}$ et $\begin{cases} dV_t^y = -c_y V_t^y dt + \tau_y dB_t^2 \\ V_0^y = v_0^y \end{cases}$

où $P_0 = (0,0,0)$ et les $(B_t^i)_{i=1,2,3}$ sont 3 mouvements browniens indépendants ; τ_x, τ_y, τ_z sont positifs, c_x, c_y sont supposés positifs et f_z est supposé négatif.
On peut remarquer que les paramètres τ_x et τ_y n'ont pas tout à fait la même signification physique par rapport aux modèles précédents.

Le temps de fécondation est à nouveau pris égal au temps de premier passage au niveau h. Comme $f_z < 0$, $T_h = \inf\{t > 0, Z_t = h\}$ est fini presque sûrement $(h < 0)$ et suit donc une GIG.

Calcul de la fonction de dispersion individuelle γ :

La première idée était de supposer que les trajectoires dans le plan horizontal sont en régime stationnaire strict, c'est-à-dire que les variables V_0^x et V_0^y admettent pour lois respectives une $\mathcal{N}\left(0, \frac{\tau_x^2}{2c_x}\right)$ et une $\mathcal{N}\left(0, \frac{\tau_y^2}{2c_y}\right)$. Alors, dans ce cas, les processus (X_t) et (Y_t) sont stationnaires strictes. En particulier ils sont centrés donc leur moyenne ne dépend pas du temps. Cela n'est pas envisageable. C'est pourquoi, on va plutôt considérer des processus non stationnaires.

Si on suppose que la trajectoire $(P_t)_{t\geq 0}$ est définie par (3.4), la loi du couple (X_t, Y_t) s'obtient facilement grâce à la proposition 2.3 et la fonction de dispersion individuelle se calcule alors à l'aide de l'équation 2.3. Elle s'exprime sous forme d'une intégrale. Cependant, l'algorithme utilisé pour estimer les paramètres ne converge pas. En particulier, on se heurte à des problèmes numériques pour le calcul de certaines intégrales.

C'est pourquoi, cela a conduit à définir une nouvelle fonction de dispersion individuelle basée sur l'approximation des fonctions $m_t(c, v_0)$ et $\sigma_t^2(c, \tau)$ définies à la propriété 2.3, par des fonctions $\bar{m}_t(c, v_0)$ et $\bar{\sigma}_t{}^2(\tau)$ respectivement.
Par analogie avec les modèles GTM, NIG et GHD définis en 2.4.2, on définit $\bar{m}_t(v_0) = v_0 t$.
Pour la fonction $\sigma_t^2(c, \tau)$, en effectuant un développement limité à l'ordre 3, on prend $\bar{\sigma}_t{}^2(\tau) = \tau^2 t^3$.

Remarques : La variance, ici, est en t^3 alors que pour les modèles précédents elle était en t. Et le paramètre c n'apparaît plus dans ce nouveau modèle.

Ainsi, on définit le modèle 5 par :

(i) X_t suit une loi normale de moyenne $\bar{m}_t(v_0^x)$ et de variance $\bar{\sigma}_t{}^2(\tau_x)$, avec $v_0^x \in \mathbb{R}$ et $\tau_x > 0$.

(ii) Y_t suit une loi normale de moyenne $\bar{m}_t(v_0^y)$ et de variance $\bar{\sigma}_t{}^2(\tau_y)$, avec $v_0^y \in \mathbb{R}$ et $\tau_y > 0$.

(iii) $Z_t = f_z t + \tau_z B_t$ avec $f_z < 0$, $\tau_z > 0$.

(iv) X, Y et Z sont indépendants.

Proposition 2.5 *On suppose :*
- la trajectoire $(P_t)_{t\geq 0}$ est définie par le modèle 5 décrit ci-dessus.
- le temps de fécondation vérifie $T_F = T_h = inf\{t > 0, Z_t = h\}$.

Alors, la loi du point de fécondation (X_{T_F}, Y_{T_F}) admet pour densité

$$\tilde{\gamma}(\theta; x, y) = \tilde{C}(\theta) \int_0^{+\infty} \frac{\exp\left(-\frac{\lambda_z}{u} - u\right) \exp\left(-\frac{(x-w_x u)^2}{2a_x^2 u^3} - \frac{(y-w_y u)^2}{2a_y^2 u^3}\right)}{2\pi a_x a_y u^{9/2}} du \qquad (2.17)$$

avec $\theta = (\lambda_z, a_x, a_y, w_x, w_y)$, et θ appartenant à $\Theta = (\mathbb{R}^+)^3 \times \mathbb{R}^2$ et
$\lambda_z = \rho\eta$, $a_x^2 = \dfrac{\tau_x^2}{\rho^3}$, $a_y^2 = \dfrac{\tau_y^2}{\rho^3}$, $w_x = \dfrac{v_0^x}{\rho}$, $w_y = \dfrac{v_0^y}{\rho}$.
De plus, $\tilde{C}(\theta) = \dfrac{\sqrt{\lambda_z} e^{2\sqrt{\lambda_z}}}{\sqrt{\pi}}$.

Démonstration :
T_h suit toujours une GIG de paramètres $\dfrac{3}{2}, \dfrac{f_z^2}{2\tau_z^2}$ et $\frac{h^2}{2\tau_z^2}$.
De plus, ici, en utilisant l'égalité (2.3), on commence par écrire,
si $\omega = (\rho, \eta, \tau_x, \tau_y, c_x, c_y, v_0^x, v_0^y)$ avec $\rho = \dfrac{f_z^2}{2\tau_z^2}$, $\eta = \dfrac{h^2}{2\tau_z^2}$

$$\tilde{\gamma}(\omega; x, y) = C(\omega) \int_0^{+\infty} \frac{\exp\left(-\frac{\eta}{t} - \rho t\right)}{t^{3/2}} \frac{\exp\left(-\frac{(x-v_0^x t)^2}{2\tau_x^2 t^3} - \frac{(y-v_0^y t)^2}{2\tau_y^2 t^3}\right)}{2\pi\tau_x\tau_y t^3} dt$$

avec $C(\omega) = \dfrac{\sqrt{\eta} e^{2\sqrt{\eta\rho}}}{\sqrt{\pi}}$.

Puis, on effectue le changement de variable $u = \rho t$ dans l'intégrale.
On obtient alors le résultat souhaité avec les notations introduites dans la proposition. ◇

Calcul de la fonction de dispersion globale dans le cas continu :

Pour les deux modèles proposés dans ce paragraphe, on utilisera la formule définie en (2.2) dans un cadre continu pour calculer la fonction de dispersion globale $\mu(x, y)$. Ci-dessous se trouvent les calculs nécessaires.

On a : $\mu(x, y) = \dfrac{SB(x, y)}{SB(x, y) + SJ(x, y)}$ avec

$$SB(x, y) = \iint_{(x', y') \in B} \gamma(x - x', y - y')\, dx' dy'$$

$$SJ(x, y) = \iint_{(x', y') \in J} \gamma(x - x', y - y')\, dx' dy$$

On suppose que le domaine B est un rectangle défini par

$$B = \{(x, y) \in \mathbb{R}^2,\ a_1 \leq x \leq a_2 \ et\ b_1 \leq y \leq b_2\}$$

Le domaine J est un rectangle

$$J_1 = \{(x, y) \in \mathbb{R}^2,\ a_3 \leq x \leq a_4 \ et\ b_3 \leq y \leq b_4\}$$

avec $a_3 < a_1$, $a_2 < a_4$, $b_3 < b_1$ et $b_2 < b_4$, auquel on retire le domaine B.
On peut donc écrire J sous la forme

$J = \{(x, y) \in \mathbb{R}^2,\ a_1 \leq x \leq a_2\, et\, (b_2 \leq y \leq b_4 \ \ ou\, b_3 \leq y \leq b_1)\} \bigcup$

$$\{(x, y) \in \mathbb{R}^2,\ b_3 \leq y \leq b_4\, et\, (a_2 \leq x \leq a_4\, ou\, a_3 \leq x \leq a_1)\}$$

On rappelle que la fonction erf est définie par

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \qquad \forall x \in \mathbb{R}$$

Calculons $SB(x, y)$.
On a $SB(x, y) = \displaystyle\int_{x'=a_1}^{a_2} \int_{y'=b_1}^{b_2} \gamma(x - x', y - y')\, dx' dy'$.

On pose

$$I(x, m_t, \sigma_t, c, d) = \frac{1}{\sqrt{2\pi}\sigma_t} \int_c^d \exp\left(-\frac{(x - x' - m_t)^2}{2\sigma_t^2}\right) dx' \qquad (2.18)$$

où m_t représente la moyenne du processus (X_t) et σ_t^2 sa variance.
On suppose que la densité du temps d'atteinte est $d(t,\theta)$. En appliquant le théorème de Fubini, on a

$$SB(x,y) = \tilde{C}(\theta) \int_{t=0}^{+\infty} d(t,\theta) I(x, m_t^x, \sigma_t^x, a_1, a_2) I(y, m_t^y, \sigma_t^y, b_1, b_2)\, dt$$

En effectuant le changement de variable $u = \dfrac{x - x' - m_t}{\sqrt{2}\sigma_t}$, on obtient

$$\begin{aligned} I(x, m_t, \sigma_t, c, d) &= -\frac{1}{\sqrt{\pi}} \int_{\frac{x-c-m_t}{\sqrt{2}\sigma_t}}^{\frac{x-d-m_t}{\sqrt{2}\sigma_t}} \exp(-u^2)\, du \\ &= \frac{1}{2}\left(\operatorname{erf}\left(\frac{x-c-m_t}{\sqrt{2}\sigma_t} \right) - \operatorname{erf}\left(\frac{x-d-m_t}{\sqrt{2}\sigma_t} \right) \right) \end{aligned}$$

D'où

Lemme 2.2 *On a*

$$SB(x,y) = \tilde{C} \int_0^{+\infty} d(t,\theta) I(x, m_t^x, \sigma_t^x, a_1, a_2) I(y, m_t^y, \sigma_t^y, b_1, b_2)\, dt.$$

De la même façon, on obtient en décomposant le domaine jaune (noté J)

Lemme 2.3 $SJ(x,y) = \tilde{C} \int_0^{+\infty} d(t,\theta) \times$
$\{ I(x, m_t^x, \sigma_t^x, a_1, a_2) \left(I(y, m_t^y, \sigma_t^y, b_2, b_4) + I(y, m_t^y, \sigma_t^y, b_3, b_1) \right) +$
$I(y, m_t^y, \sigma_t^y, b_3, b_4) \left(I(x, m_t^x, \sigma_t^x, a_2, a_4) + I(x, m_t^x, \sigma_t^x, a_3, a_1) \right) \}\, dt.$

2.5.4 Synthèse sur les modélisations de la trajectoire

Aux paragraphes 2.4 et 2.5, ont été décrites différentes modélisations pour la trajectoire d'un grain de pollen de maïs, vu comme une particule soumise à un champ de force, basées sur l'utilisation de mouvements browniens avec drift et de processus d'Ornstein-Uhlenbeck intégrés. On a alors obtenu plusieurs fonctions, paramétriques, de dispersion individuelles efficaces.
Par la suite, à partir de l'expérience décrite en (2.2), on va estimer les paramètres des différents modèles à l'aide d'une méthode de quasi-vraisemblance. Cela va permettre dans un premier temps de les comparer à l'aide de méthodes graphiques (comparaison des graphiques des résidus réduits). Ensuite, on peut relier les paramètres estimés avec les données météorologiques et les données physiques du maïs. On peut donc comparer les valeurs estimées et les valeurs observées. Cela nous permet d'en déduire le(s) modèle(s) le(s) plus proche(s) de la réalité. (Dans Klein *et al* (2003) la comparaison avec les données météorologiques n'était pas concluante)

Ce travail a pour but, dans un premier temps, de simuler, dans des conditions physiques proches de la réalité, des périodes de pollinisation du maïs et par suite d'étudier la dispersion du pollen, par exemple en faisant varier la taille et la forme des champs. Pour cela, on a besoin de tenir compte des caractéristiques physiques des grains de pollen (poids, hauteur des fleurs femelles) et de paramètres liés au vent (direction, intensité). Le travail de Klein (2000) sur la fonction de dispersion pour le maïs a été intégré dans le modèle MAPOD développé par l'unité Eco-Innov de l'INRA Grignon et le laboratoire ESE d'Orsay (Angevin *et al*, 2001). Ce modèle permet de décrire en milieu homogène des dispersions de gènes sur des parcelles agricoles dans des conditions réalistes.
Un second objectif est de modéliser, à partir des résultats obtenus, le flux de pollen quand le paysage n'est plus homogène, c'est-à-dire quand deux champs sont séparés par un sol nu ou une autre culture. C'est l'objectif du chapitre 3.

2.6 Estimation des paramètres

A la section 2.3.2, on a défini le modèle statistique de la façon suivante :

$$N_k = n_t\mu(\theta; x_k, y_k) + \varepsilon_k \text{ avec } E(\varepsilon_k) = 0 \text{ et } Var(\varepsilon_k) = \sigma^2 n_t v(\theta, b; \mu(x_k, y_k))$$

où les $(\varepsilon_k)_k$ sont supposées indépendantes, n_t représente le nombre total de grains sur l'épi localisé en (x_k, y_k) et N_k le nombre de grains bleus sur cet épi. Le paramètre introduit σ^2 est un paramètre de dispersion.

Dans un premier temps, on présente les trois fonctions de variance envisagées. Puis, on décrit la méthode statistique utilisée : méthode de quasi-vraisemblance. Enfin, on donne les résultats et analyse des résultats pour les 5 modèles décrits précédemment et les 3 fonctions de variance.

2.6.1 Fonctions de variance proposées

Trois fonctions de variance v sont envisagées dans ce travail.

1. Fonction de type binomiale :

   Supposons que $y_i = \sum_{i=1}^{n_i} R_{i,j}$ où les $R_{i,j}$ sont des variables aléatoires ne pouvant prendre que deux valeurs, c'est-à-dire que ce sont des variables aléatoires de Bernoulli de paramètre p_i. Si les variables $R_{i,j}$ sont indépendantes, on retrouve l'hypothèse faite par Klein *et al* (2003) : N_k suit une loi binomiale de paramètres $(n, \mu(x_k, y_k))$.
   Maintenant, si les variables $R_{i,j}$ sont corrélées entre elles et si on note d le coefficient de corrélation entre $R_{i,j}$ et $R_{i,k}$ pour $j \neq k$ alors on a toujours $E(y_i) = n_i p_i$ (par linéarité de l'espérance).
   D'autre part, on a $\mathrm{Var}(y_i) = \sum_{j=1}^{n_i} \mathrm{Var}(R_{i,j}) + \sum_{j=1}^{n_i} \sum_{j \neq k} \mathrm{Cov}(R_{i,j}, R_{i,k})$.
   Le coefficient de corrélation d, entre $R_{i,j}$ et $R_{i,k}$, est $d = \dfrac{\mathrm{Cov}(R_{i,j}, R_{i,k})}{\sqrt{\mathrm{Var}(R_{i,j})\,\mathrm{Var}(R_{i,k})}}$.
   Comme $\mathrm{Var}(R_{i,j}) = \mathrm{Var}(R_{i,k}) = p_i(1-p_i)$, il vient $\mathrm{Cov}(R_{i,j}, R_{i,k}) = dp_i(1-p_i)$.
   D'où

$$\begin{aligned} \mathrm{Var}(y_i) &= \sum_{j=1}^{n_i} p_i(1-p_i) + \sum_{j=1}^{n_i} \sum_{j \neq k} dp_i(1-p_i) \\ &= (1 + d(n_i - 1))n_i p_i(1-p_i) \end{aligned} \tag{2.19}$$

   D'où le choix d'une fonction de variance dite de type binomiale :
   $v_1(\mu) = \mu(1-\mu)$.

   Remarques :
   Si $d > 0$, cela signifie un plus grand écart entre les nombres de "succès" que ce qui est attendu dans le cas où les variables observées sont indépendantes.

Une autre possibilité d'inadéquation du modèle peut être la variabilité dans les probabilités de réponse. Cela peut se produire lorsque l'ensemble des observations ne sont pas prises dans les mêmes conditions expérimentales. Des calculs (par exemple Collett 1991) conduisent à l'équation (2.19) avec $d > 0$. Ainsi, dans le cas de la sur-dispersion, on ne peut pas distinguer les effets dus à une corrélation entre les données et les effets dus à une variabilité entre les probabilités de réponse.

2. Fonction de type linéaire :

   En étudiant les données expérimentales, on remarque que les valeurs observées pour les proportions de nombres de grains bleus sur un épi, sont essentiellement comprises entre 0 et 0.2 et ne dépassent pas 0.5. Ainsi, il est important de modéliser la fonction de variance sur cet intervalle (et non sur tout l'intervalle $[0, 1]$).

   De plus, on souhaite prendre une fonction de variance non nulle en 0. En effet, il y a beaucoup de valeurs observées proches de 0 et pour l'estimation paramétrique des fonctions de dispersion individuelles, la méthode statistique utilisée (voir paragraphe suivant pour la description détaillée de la méthode de quasi-vraisemblance) fait intervenir la fonction de variance au dénominateur des équations de quasi-vraisemblance, d'où l'apparition de grandes valeurs.
   Afin de pondérer de façon correcte les observations significatives, on considère donc une fonction de variance de type linéaire, définie pour $\mu \in [0, 0.5]$, par $v_2(\mu, a) = (a + \mu)$ où $a \in \mathbb{R}^+$.

3. Fonction de type exponentiel :

   Enfin, en traçant le graphique des variances empiriques par rapport à l'espérance, en ayant regroupé les données par paquets de taille suffisante, on constate que la courbe croit légèrement pour les valeurs de l'espérance comprises entre 0 et 0.2 , puis croît fortement pour celles supérieures à 0.2 .
   Cela nous amène à considérer une fonction de variance de type "exponentielle" de la forme $v_3(\mu, b) = \exp(b\mu)$ avec $b > 0$ (toujours pour $\mu \in [0, 0.5]$).

2.6.2 Méthode

a) Z-estimateur :

On suppose que l'on observe $(Y_1, ..., Y_n)$ un n-échantillon d'une loi P_θ, où $\theta \in \Theta$ est le vecteur paramètre, de dimension k. On souhaite estimer ce paramètre θ.
On suppose que l'estimateur $\hat{\theta}$ de θ_0 vérifie un système d'équations de la forme

$$\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi_\theta(Y_i) = 0.$$

La fonction $\psi = (\psi_{\theta,1}, ...\psi_{\theta,k})$ est appelée la fonction score et l'estimateur $\hat{\theta}$ Z-estimateur.

On fait les hypothèses suivantes :
- $\Psi_n(\theta)$ converge en probabilité vers $\Psi(\theta) = E(\psi_\theta(X_1))$.
- La suite $(\hat{\theta}_n)_n$ converge en probabilité vers θ_0 quand n tend vers $+\infty$ avec $\Psi(\theta_0) = 0$.
- Pour tout x, la fonction $\theta \mapsto \psi_\theta(x)$ admet des dérivées partielles secondes et continues par rapport à θ.
- La matrice, $(k \times k)$, des dérivées premières partielles de Ψ_n, notée $\dot{\Psi}_n(\theta_0)$, converge en probabilité vers la matrice $(k \times k)$ $\Gamma_{\theta_0} = E(\dot{\psi}_{\theta_0})$ inversible.
- On note enfin $I(\theta) = E(\psi_\theta \psi_\theta^T)$.

Alors on a le théorème suivant (Van der Vaart 1998 par exemple)

Théorème 2.2 $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_k\left(0, (\Gamma_{\theta_0})^{-1} I(\theta_0) (\Gamma_{\theta_0}^T)^{-1}\right)$

b) Méthode de maximum de vraisemblance :

On suppose toujours que l'on observe $(Y_1, ..., Y_n)$ un n-échantillon ayant pour fonction de densité p_θ. On souhaite estimer le vecteur paramètre θ de dimension k. L'estimateur du maximum de vraisemblance, noté $\hat{\theta}_{V,n}$, maximise la fonction de log-vraisemblance définie par $L_n(X, \theta) = \sum_{i=1}^{n} \ln(p_\theta(X_i))$. Cela équivaut à résoudre le système d'équations, pour $j = 1, ..., k$:

$\Psi_n(\theta) = \sum_{i=1}^{n} \psi_{\theta,j}(X_i) = 0$ avec ici $\psi_{\theta,j}(x) = \dfrac{\partial}{\partial \theta_j} \ln(p_\theta(x))$.

Sous les conditions de régularité ci-dessus, on a le théorème de convergence en loi asymptotique suivant :

Théorème 2.3 $\sqrt{n}(\hat{\theta}_{V,n} - \theta_0) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}_k(0, I^{-1}(\theta_0))$

La démonstration du théorème est basée sur l'utilisation d'une formule de Taylor pour la fonction Ψ_n et sur les égalités suivantes :

$$E(\psi_\theta) = 0 \text{ et } E(\dot{\psi}_\theta) = -I(\theta) \tag{2.20}$$

obtenues en différenciant l'égalité $\displaystyle\int p_\theta \, dP = 1$.

La matrice $I(\theta)$ est la matrice d'information de Fisher.

Remarque : Si on considère un modèle admettant une log-vraisemblance de la forme $\sigma^{-2}\{y^T\theta - b(\theta) - c(y, \sigma)\}$, alors en différenciant l'expression et en supposant que le support de la loi ne dépende pas de θ, on trouve :

$$E(Y) = \mu(\theta) = b'(\theta) \text{ et } \mathrm{Cov}(Y) = \sigma^2 b"(\theta) = \sigma^2 V(\mu).$$

c) Méthode de maximum de quasi-vraisemblance :

Au paragraphe 2.3.2, nous avons fait des hypothèses uniquement sur l'espérance et la variance des variables aléatoires N_k. Ainsi, leurs lois ne nous sont pas connues et on ne peut pas estimer les paramètres des modèles à l'aide de la méthode de maximum de vraisemblance décrite ci-dessus. Nous allons donc utiliser une méthode de quasi-vraisemblance introduite par Wedderburn (1974) et généralisée par McCullagh (1983) par exemple.

On se donne un vecteur aléatoire $Y = (Y_1, ..., Y_n)$ de coordonnées indépendantes avec $E(Y_i) = \mu_i(\beta)$ et $\mathrm{Var}(Y_i) = \sigma^2 v(\mu_i)$ où v est une fonction donnée et où l'on note $\mu_i = \mu_i(\beta)$.

La fonction de quasi-vraisemblance est construite à partir du système d'équations différentielles suivant :

Soit $R_i(\mu_i; y_i) = \dfrac{\partial K}{\partial \mu_i}(\mu_i; y_i) = \dfrac{y_i - \mu_i}{v(\mu_i)}$.

Le logarithme de la fonction de quasi-vraisemblance, noté K, est alors défini par

$$K(\mu; y) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} R_i(t, y_i)\, dt = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{v(\mu_i)}\, dt$$

L'estimateur de quasi-vraisemblance, noté $\hat{\beta}_{QV}$, maximise la fonction K. Cela revient donc à résoudre le système d'équations, pour $j = 1, ..., k$:

$$U_j(\beta) = \sum_{i=1}^{n} \frac{\partial \mu}{\partial \beta_j}(\beta; y_i) \frac{y_i - \mu_i}{v(\mu_i)} = 0$$

ou sous forme matricielle $U(\beta) = D^T V^{-1}(\mu)\,(y - \mu) = 0$ avec $D = \dfrac{\partial \mu}{\partial \beta}$ et

$V(\mu) = Diag(v(\mu_1), ..., v(\mu_n))$, matrice $n \times n$ diagonale.

U est donc la fonction "quasi-score". Ici, on a donc $\psi_{\theta_j}(x) = \dfrac{\partial \mu}{\partial \beta_j}(\beta; x) \dfrac{x - \mu_i}{v(\mu_i)}$.

En fait, on a (par exemple Wedderburn 1974) $E(R_i) = 0$, $\mathrm{Var}(R_i) = \dfrac{1}{v(\mu_i)}$,

$E\left(\dfrac{\partial R_i}{\partial \mu_i}\right) = \dfrac{1}{v(\mu_i)}$ et $E\left(\dfrac{\partial K}{\partial \beta_i}\dfrac{\partial K}{\partial \beta_j}\right) = -E\left(\dfrac{\partial^2 K}{\partial \beta_i \partial \beta_j}\right) = \dfrac{1}{V(\mu)}\dfrac{\partial \mu}{\partial \beta_i}\dfrac{\partial \mu}{\partial \beta_j}$.

Ainsi la fonction $U(\beta)$ est centrée et de matrice de covariance :

$$\sigma^2 i_\beta = \sigma^2 D^T V^{-1} D = E(\psi_\theta \psi_\theta^T)$$

La log-quasi-vraisemblance va donc se comporter "de la même façon" que la log-vraisemblance (la preuve du théorème (2.2) étant basée sur l'utilisation d'une formule de Taylor et les égalités (2.20).

Plus précisément, on a (McCullagh 1983)

$$\sqrt{n}(\hat{\beta}_{QV} - \beta_0) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}_k\left(0, \sigma^2 i^{-1}(\beta_0)\right)$$

où $-i(\beta_0)$ représente l'espérance de la matrice des dérivées secondes de K.

Le paramètre σ^2 est estimé par la variance résiduelle :

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \mu_i(\hat{\beta}))^2}{v(\mu_i(\hat{\beta}))} \tag{2.21}$$

d) Cas étudié :

Pour la fonction de type binomiale, c'est exactement le cas traité en c).
Par contre, pour les deux autres fonctions, il y a un paramètre en plus dans la fonction de variance.

On note p la dimension du vecteur des paramètres θ et on suppose que b est de dimension 1. Alors les équations de quasi-vraisemblance sont données pour $i = 1, ..., p$ par

$$U_i(\theta, b) = \sum_{k=1}^{n}\frac{\partial\mu}{\partial\theta_i}(\theta; x_k, y_k)\frac{N_k - n_t\mu(\theta; x_k, y_k)}{g(\theta, b, n_t; x_k, y_k)}$$

et l'on rajoute l'équation suivante :

$$U_{p+1}(\theta, b) = \sum_{k=1}^{n}\frac{\partial g}{\partial b}(\theta, b, n_t; (x_k, y_k))\,\frac{(N_k - n_t\mu(\theta; x_k, y_k))^2 - g(\theta, b, n_t; (x_k, y_k))}{g^2(\theta, b, n_t; (x_k, y_k))} \tag{2.22}$$

L' estimateur de quasi-vraisemblance de (θ, b) , noté $(\hat{\theta}, \hat{b})$, est alors défini par $U_i(\hat{\theta}, \hat{b}) = 0$ pour tout $i = 1, ..., p$ et $U_{p+1}(\hat{\theta}, \hat{b}) = 0$.
Le paramètre σ^2 est toujours estimé par la variance résiduelle définie en (2.21) avec ici $\mu_i(\theta) = n\mu(\theta; (x_i, y_i))$ et le terme $v(\mu_i(\hat{\beta}))$ est remplacé par $g(\theta, b, n; (x_k, y_k))$.

L'estimateur $(\hat{\theta}, \hat{b})$ vérifie les mêmes propriétés asymptotiques décrites ci-dessus en c). En particulier, il est consistant et converge en loi vers une loi normale centrée quand le nombre d'observations n tend vers l'infini.

En effet, l'équation (2.22) est construite à partir d'un argument asymptotique (Huet *et al* 1996). Si Z est la somme de n variables aléatoires identiquement distribuées d'espérance μ et de variance σ^2, le théorème central limite dit que, pour n suffisamment grand, la loi de $\frac{(Z-n\mu)}{\sqrt{n}\sigma}$ peut être approchée par une loi normale centrée et de variance 1. Ainsi, l'équation (2.22) est construite comme si les moments jusqu'à l'ordre 4 des variables $\varepsilon_i = Y_i - n\mu_i$ étaient égaux aux moments d'une loi normale centrée et de variance $\sigma^2 v(\theta, b, n_i; (x_i, y_i))$. Alors, quand n est suffisamment grand, $\mathrm{Var}(\varepsilon_i) \approx 2\sigma^4 v^2(\theta, b, n_i; (x_i, y_i))$.

Intervalles de confiance :
On peut donc construire des intervalles de confiance. Un intervalle de confiance pour le paramètre θ_i de niveau asymptotique 95% est, si $\sqrt{n}(\hat{\theta}_i - \theta_{0,i})$ converge en loi vers une $\mathcal{N}(0, \hat{e}^2)$ quand n tend vers l'infini , $IC = [\hat{\theta}_i - 1.96\hat{e}, \hat{\theta}_i + 1.96\hat{e}]$.

Pour le paramètre σ^2, $\dfrac{n\hat{\sigma^2}}{\sigma_0^2}$ converge en loi, quand n tend vers l'infini, vers un χ^2 à

$(n-p)$ degrés de liberté, où n représente le nombre de données et p la dimension du vecteur paramètre θ (Collett (1991) ou McCullagh et Nelder (1989)).
On en déduit un intervalle de confiance pour σ^2 de niveau asymptotique 95 % de la forme

$$IC(\hat{\sigma^2}) = \left[\frac{n\hat{\sigma^2}}{q_{0.975}}, \frac{n\hat{\sigma^2}}{q_{0.025}} \right]$$

avec $q_{0.025}$ et $q_{0.975}$ quantiles d'un χ^2 à $(n-p)$ degrés de liberté.

2.6.3 Mise en oeuvre

Pour les calculs, on a utilisé le langage C et le logiciel nls2 [1], qui est un ensemble de fonctions Splus ou R et de programmes C (avec la librairie NAG). Il utilise l'algorithme de Gauss-Marquardt pour estimer les paramètres d'un modèle de régression non-linéaire sur un ensemble donné d'observations.
L'algorithme de Gaus-Marquardt est une modification de l'algorithme de Gauss-Newton. Il utilise les expressions explicites des fonctions des dérivées de la fonction de dispersion μ et de la fonction de variance v. Il améliore la stabilité numérique.

Pour calculer la fonction μ, fonction de dispersion globale définie en (3.1) et (2.2), il faut effectuer le calcul d'une double somme discrète ou d'une double intégrale, suivant les cas.
Pour les modèles 1 à 3, on a utilisé un maillage discret basé sur l'emplacement des plantes, c'est-à-dire un pas de discrétisation de 0.8 mètres pour l'axe des x et de 0.2 mètres pour l'axe des y. (cf la figure 2.1)
Pour les modèles 4 et 5, un maillage discret conduisait au calcul d'une double somme discrète et d'une intégrale généralisée. Cela étant trop coûteux en temps, on a donc utilisé la définition de la fonction μ dans le cadre continu. Ainsi, d'après les calculs effectués en 2.5, on a pu se ramener au calcul d'une seule intégrale, ce qui a considérablement diminué les temps de calcul par ordinateur.

2.6.4 Résultats

Ci-dessous, se trouvent les tableaux des estimations des paramètres pour les modèles 1 à 5 définis en détails aux paragraphes 2.4 et 2.5. Les écart-types sont calculés à partir des données réelles.

On rappelle que $\sigma^2 = 1 + d(n_t - 1)$ où n_t est le nombre total de grains par épi.

1. Modèle 1 : GTM

Ce modèle est décrit au paragraphe 2.4.2 a). Il consiste à supposer que la végétation est dense et limite donc les trajectoires. On a vu que la fonction de dispersion individuelle était alors une GTM, dont l'expression est :

[1] Pour plus de renseignements : http ://www.inra.fr/bia/produits/logiciels/.

$$f_{GTM}(\lambda_z, \lambda_x, \lambda_y, \delta; x, y) = \frac{\delta^2 \lambda_z e^{\lambda_z}}{2\pi} \frac{e^{-\sqrt{pq(x,y)}}}{\sqrt{q(x,y)}} e^{\delta(\lambda_x x + \lambda_y y)}$$

où on a fait l'hypothèse que $\tau_x = \tau_y$ donc $\delta_x = \delta_y = \delta$.

	Variance binomiale		Variance linéaire		Variance exponentielle	
Paramètres	Estimation	Ecart-type	Estimation	Ecart-type	Estimation	Ecart-type
δ	11.04	64.85	10.18	2.55	4.776	1.5119
λ_x	-0.0001	0.0602	-0.0004	0.0004	-0.007	0.0024
λ_y	0.0096	0.0094	0.0113	0.003	0.0315	0.0104
λ_z	0.0156	0.0965	0.0175	0.004	0.0433	0.0142
a	-	-	$1.093\ 10^{-5}$	$3.567\ 10^{-7}$	-	-
b	-	-	-	-	74.32	0.5821
σ^2	9588 [9115 ; 10098]		14.64 [13.90 ; 15.37]		3.403 [3.235 ; 3.584]	
d	24.39		0.0347		0.0061	

TAB. 2.1 – Résultats des estimations des paramètres pour le modèle 1 : GTM

2. Modèle 2 : NIG

Ce modèle est décrit au paragraphe 2.4.2 b). On ne suppose plus que la végétation arrête les trajectoires. Celles-ci s'arrêtent la première fois où elles atteignent le niveau des fleurs femelles. On a vu que la fonction de dispersion individuelle était alors une NIG, dont l'expression est :

$$f_{NIG}(\lambda_z, \lambda_x, \lambda_y, \delta; x, y) = \frac{\delta^2 e^{\lambda_z}}{2\pi} \frac{(q(x,y)^{-1/2} + p^{1/2})}{q(x,y)} e^{-\sqrt{pq(x,y)}} e^{\delta(\lambda_x x + \lambda_y y)}$$

où on a fait l'hypothèse que $\tau_x = \tau_y$ donc $\delta_x = \delta_y = \delta$.

A ce niveau-là, on constate que, pour la valeur estimée du paramètre b dans la fonction de variance, v_3, de type exponentielle, il y a "explosion", c'est-à-dire que la fonction v_3 atteint rapidement de grandes valeurs quand μ croît. ($\mu \in [0, 0.5]$)
De plus, cette fonction de variance apporte une structure aux résidus réduits lors de leur représentation graphique en fonction des ajustés (cf figures 2.15 et 2.16 pages 52-53). De même lors de la représentation des résidus réduits sur le champ : pour les résidus négatifs, il apparaît une sorte de cercle tout autour du champ de maïs à grains bleus.
On peut également remarquer que, pour le modèle 1, cela modifie significativement la valeur estimée du paramètre δ.
–> On exclut cette modélisation par la suite.

	Variance binomiale		Variance linéaire		Variance exponentielle	
Paramètres	Estimation	Ecart-type	Estimation	Ecart-type	Estimation	Ecart-type
δ	0.5176	0.0837	0.5177	0.0225	0.5306	0.0269
λ_x	- 0.0056	0.0509	-0.0096	0.0065	-0.0308	0.0047
λ_y	0.1808	0.0244	0.1914	0.0143	0.1854	0.0156
λ_z	0.0561	0.0835	0.0669	0.0204	0.0761	0.0189
a	-	-	$1.175\ 10^{-5}$	$3.589\ 10^{-7}$	-	-
b	-	-	-	-	95.33	0.6891
σ^2	145.9		9.261		2.930	
	[138.7 ; 156.7]		[8.801 ; 9.755]		[2.786 ; 3.086]	
d	0.3687		0.0210		0.0049	

TAB. 2.2 – Résultats des estimations des paramètres pour le modèle 2 : NIG

3. Modèle 3 : GHD

Ce modèle est décrit au paragraphe 2.4.2 c). Ici, la fonction de dispersion individuelle considérée est une GHD avec le paramètre α quelconque. Elle est de la forme :

$$f_{GHD}(\alpha, \lambda_z, \lambda_x, \lambda_y, \delta; x, y) = \frac{\lambda_z^{1-\alpha}\delta^2(p/q(x,y))^{\frac{\alpha}{2}}}{2\pi} \frac{\mathcal{K}_\alpha(\sqrt{pq(x,y)})}{\mathcal{K}_{1-\alpha}(\lambda_z)} e^{\delta(\lambda_x x + \lambda_y y)}$$

où on a fait l'hypothèse que $\tau_x = \tau_y$ donc $\delta_x = \delta_y = \delta$.

	Variance binomiale		Variance linéaire	
Paramètres	Estimation	Ecart-type	Estimation	Ecart-type
δ	0.5985	0.2334	0.5850	0.0571
λ_x	-0.0046	0.0217	-0.0082	0.0057
λ_y	0.1555	0.0746	0.1683	0.0205
λ_z	0.0799	0.0573	0.0853	0.0151
α	1.3986	0.1804	1.4133	0.0461
a	-	-	$1.166\ 10^{-5}$	$3.534\ 10^{-7}$
σ^2	155.6		9.263	
	[147.9 ; 163.8]		[8.808 ; 9.752]	
d	0.3934		0.0210	

TAB. 2.3 – Résultats des estimations des paramètres pour le modèle 3 : GHD

Pour les modèles avec une variance type binomiale, on obtient des résultats assez proches de ceux de Klein *et al*, qui avaient utilisé une méthode de maximum de vraisemblance pour l'estimation des paramètres.
Pour plus de détails, on se reportera au paragraphe (2.17) sur la comparaison des

estimations avec les paramètres physiques.

4. Modèle 4 :

Ce modèle est décrit au paragraphe 2.5.2. Il est basé sur la modélisation de la trajectoire à l'aide de deux mouvements browniens avec drift indépendants pour les coordonnées (X_t, Y_t) ; et d'un processus d'Ornstein-Uhlenbeck intégré pour la composante verticale (Z_t).

	Variance binomiale		Variance linéaire	
Paramètres	Estimation	Ecart-type	Estimation	Ecart-type
δ_x	-0.1822	0.3221	-0.1903	0.1010
δ_y	1.4359	0.4796	1.3340	0.1620
λ	0.1571	0.0179	0.1820	0.0078
b_z	0.4683	0.0532	0.4417	0.0163
c_z	0.1208	0.0647	0.1299	0.0216
a	-	-	$8.948\ 10^{-6}$	$4.803\ 10^{-8}$
σ^2	153.6		18.81	
	[146.03 ; 161.76]		[17.88 ;19.82]	
d	0.3883		0.0453	

TAB. 2.4 – Résultats des estimations des paramètres pour le modèle 4

Remarque : A cause de problèmes numériques, on a approché les fonctions $\frac{H(u)}{g(u)}$ et $\frac{H'(u)}{g'(u)}$ par -1 dans la fonction de densité p_{θ_z}.

5. Modèle 5 :

Ce modèle est décrit au paragraphe 2.5.3. Il est basé sur la modélisation de la trajectoire à l'aide de deux processus d'Ornstein-Uhlenbeck indépendants et non stationnaires pour les coordonnées (X_t, Y_t) ; et d'un mouvement brownien avec drift pour la composante verticale (Z_t).

2.6.5 Analyse des résultats

Rappelons qu'au vu des résultats, on a exclu la variance de type exponentielle.

En comparant les écarts-type pour les différents modèles avec la variance de type binomiale et la variance de type linéaire, on constate que les écarts-type semblent meilleurs pour la variance type linéaire car ils sont plus petits et toujours significatifs. En effet, pour le modèle 1, GTM, les écarts-type pour la variance type binomiale ne sont pas significatifs pour δ (on a $\hat{\delta} = 11.04$ et $\hat{e} = 64.85$), λ_x et λ_z, paramètres devant être positifs ($\hat{\lambda_z} = 0.0156$ et $\hat{e} = 0.0965$). De même, pour le modèle 2, NIG, on a pour le paramètre λ_z, $\hat{\lambda_z} = 0.0561$ et $\hat{e} = 0.0835$.

	Variance binomiale		Variance linéaire	
Paramètres	Estimation	Ecart-type	Estimation	Ecart-type
λ_z	0.3034	0.0470	0.3058	0.0480
a_x	7.514	0.9423	7.498	0.09423
a_y	10.13	1.233	10.07	1.234
w_x	-0.7779	0.1826	-0.8749	0.1910
w_y	4.873	0.5155	4.969	0.5173
a	-	-	$1.006\ 10^{-5}$	$2.630\ 10^{-7}$
σ^2	7.674 [7.296 ; 8.082]		7.089 [6.7406 ; 7.466]	
d	0.0170		0.0155	

TAB. 2.5 – Résultats des estimations des paramètres pour le modèle 5

- - > Cela incite à choisir la fonction de variance de type linéaire.

De plus, le premier modèle, GTM, est une GHD avec le paramètre $\alpha = 1/2$ et le second modèle, NIG, est une GHD avec le paramètre $\alpha = 3/2$.
Au paragraphe suivant, on testera donc si on peut accepter $\alpha = 3/2$ (ou $\alpha = 1/2$) en ayant observé que pour le modèle 3, GHD, le paramètre α estimé valait 1.40 et 1.41.

Ci-dessous, le graphe pour le modèle 2, NIG, des fonctions de variance de type binomiale et linéaire pour les valeurs des paramètres estimés.
On peut voir que pour μ autour de 0.2, la fonction de variance de type binomiale est de l'ordre de 10 fois plus grande que la fonction de variance de type linéaire. Pour comparaison, la fonction de variance de type exponentielle est alors 10^8 fois plus grande que celle de type linéaire.

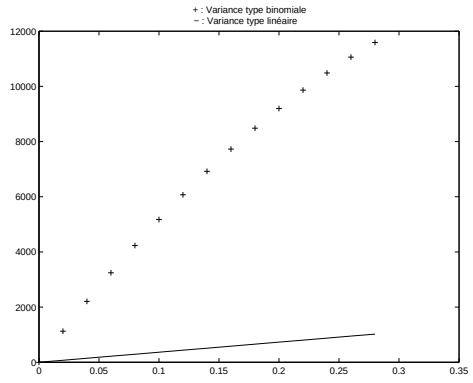


FIG. 2.4 – Fonctions de variance de types binomiale et linéaire pour le modèle NIG

2.7 Validation des résultats

Après l'estimation des paramètres pour les différentes modélisations, il nous faut valider les résultats obtenus et essayer de déterminer si certaines modélisations sont meilleures que d'autres.
Pour commencer, on utilisera un critère statistique de sélection de modèle. Cependant, un critère de sélection classique ne peut pas être utilisé. En effet, on ne dispose pas d'une log-vraisemblance pour utiliser le critère d'Akaïke (Akaïke 1973), les lois des variables aléatoires N_k n'étant pas connues. On ne peut pas utiliser non plus le coefficient C_p de Mallows (Mallows 1973), qui suppose que les variances des erreurs sont constantes car dans notre cas, elles ne le sont pas. Néanmoins, il existe des extensions de ces critères de sélection dans le cas où on utilise une méthode de quasi-vraisemblance. Hurvich et Tsai (1995) ont introduit un critère type Akaïke, explicité au paragraphe (2.7.1).

D'autre part, des méthodes graphiques sont à notre disposition pour comparer les différents modèles reposant sur l'étude des résidus réduits et des ajustés, l'étude des courbes de dispersion obtenues.

Enfin, on peut également comparer les valeurs des paramètres estimés avec les paramètres physiques mesurés, ce qui est intéressant pour choisir le modèle le mieux adapté aux données.

2.7.1 Critère de sélection

Des hypothèses sur l'espérance et la variance des variables aléatoires N_k ont été faites et une méthode de quasi-vraisemblance a été utlisée pour l'estimation des paramètres.
Hurvich et Tsai (1995) ont étendu le critère de sélection de modèle d'Akaïke au modèle de quasi-vraisemblance de la façon suivante :

$$AICc = \log(\sigma^2) + \frac{2(p+1)}{n-p-2}$$

avec n le nombre d'observations et p la dimension de θ.
Le critère d'information d'Akaïke correspondant est

$$AIC = \log(\sigma^2) + 1 + \frac{2(p+1)}{n}$$

Ce critère présente l'avantage de pouvoir comparer des modèles qui ne sont pas emboîtés. Ainsi, en théorie, on pourrait comparer les modèles 1 à 5. Cependant, on ne le fera pas ici car les calculs pour l'estimation des paramètres n'ont pas été faits de la même façon. En effet, pour les modèles 1 à 3, pour calculer la fonction de dispersion globale $\mu(x, y)$, on a utilisé une somme discrète, contrairement aux modèles 4 et 5, pour lesquels on a calculé la fonction $\mu(x, y)$ à l'aide d'une intégrale (donc de façon continue).

Voici les valeurs obtenues pour les modèles 1 à 3 :

	GTM	NIG	GHD
Variance de type binomiale	9.172	4.986	5.051
Variance de type linéaire	2.6876	2.2299	2.2308

TAB. 2.6 – Tableau donnant les valeurs de AICc dans les différents cas

- Pour la variance de type binomiale :
  On préfère le modèle 2, NIG. Mais pour le modèle 3, GHD, la valeur de AICc est presque égale. On note également que, pour le modèle 1, GTM, on obtient une valeur de AICc nettement plus grande.
- Pour la variance de type linéaire :
  Les valeurs de AICc sont presque égales pour les modèles NIG et GHD, avec un très léger avantage pour le modèle NIG. La valeur de AICc, pour le modèle 1, GTM, est encore plus grande mais l'écart avec les deux autres modèles est réduit par rapport aux modèles avec la variance de type binomiale.

On préfère donc le modèle 2, NIG, avec une variance de type linéaire.

Pour les modèles 4 et 5, le même critère conduit à préférer le modèle 5 pour les deux types de fonctions de variance.

2.7.2 Tests sur le paramètre α de la loi GHD

Les deux premiers modèles étant des sous-modèles du modèle 3, cela donne la possibilité d'effectuer des tests de rapport de vraisemblance (dans le cas d'un modèle non linéaire hétéroscédastique ; cf Huet *et al* 1996) sur le paramètre α.

On teste l'hypothèse H_0 : "$\alpha = 3/2$" contre l'alternative H_1 : "$\alpha \neq 3/2$"
(c'est-à-dire le modèle est une NIG).
La statistique de test, Sv, converge en loi asymptotiquement, quand n tend vers l'infini, vers un χ^2 à un degré de liberté. On souhaite effectuer un test au niveau asymptotique 5%. Si Z est une variable aléatoire suivant un χ^2 à un degré de liberté, alors le quantile $q_{0.05}$ tel que $\mathbb{P}(Z > q_{0.05}) = 0.05$ est égal à 3.84.

Pour la modélisation avec la variance de type binomiale, on obtient $Sv = 5.4$ donc on rejette l'hypothèse H_0 au niveau asymptotique 5%.

Pour la modélisation avec la variance de type linéaire, on obtient $Sv = 2.26$ donc on accepte l'hypothèse H_0 au niveau asymptotique 5%.

Pour le modèle 1, cela revient à tester H_0 : "$\alpha = 1/2$" contre l'alternative H_1 : "$\alpha \neq 1/2$" . Dans les deux cas (variances de type binomiale et linéaire) , on rejette l'hypothèse H_0 au niveau asymptotique 5%, ce qui paraît logique.

	GTM	NIG
Variance de type binomiale	9544	5.40
Variance de type linéaire	72.4	2.26

TAB. 2.7 – Tableau donnant les valeurs de la statistique de test dans les différents cas

2.7.3 Étude des résidus réduits

Définition 2.8 *Soit $\hat{\theta}$ l'estimateur du vecteur des paramètres θ_0. Les ajustés sont les valeurs $(\hat{\mu}(x_k, y_k))_k$ définies par $\hat{\mu}(x_k, y_k) = \mu(\hat{\theta}; x_k, y_k)$.*

Définition 2.9 *Les résidus réduits sont définis par*

$$R_k = \frac{N_k - n\hat{\mu}(x_k, y_k)}{\sqrt{\sigma^2 nv(\hat{\mu}(x_k, y_k))}}$$

Si le modèle étudié est convenable, la plupart des résidus réduits doit être compris entre -2 et 2 et le graphique ne doit pas présenter de structure particulière.

Cependant, pour le cas étudié, on a une certaine structure de base due à la nature des données.
- En effet, ce sont les données du comptage de nombres de grains bleus par épi. Et pour un certain nombre d'épis, il y a un faible nombre de grains bleus, par exemple, $N_k = 0, 1, 2..$, donc de très faibles valeurs pour μ. Cela se traduit par l'apparition de certaines "lignes" sur les graphiques des résidus réduits en fonction des ajustés, en particulier la ligne horizontale située au niveau 0.
- Pour la représentation sur le champ, la direction du vent est de la droite vers la gauche du champ à peu près horizontalement. De ce fait, on observe un grand nombre de N_k nuls pour les capteurs se situant à gauche du champ central. D'où un nombre plus important de ce côté de résidus réduits négatifs car on a tendance à surestimer les N_k. (En comparant ces graphiques avec le graphique des observations, pour le côté gauche du champ, il y a concordance entre les valeurs de N_k nulles ou très petites et les résidus négatifs.)

Graphiques des résidus réduits en fonction des ajustés :

Une première façon, pour comparer les différentes modélisations, est de tracer les graphiques des résidus réduits en fonction des ajustés. On a utilisé une échelle logarithmique pour l'axe des abscisses, c'est-à-dire pour les ajustés.
Leurs représentations ne nous apportant pas beaucoup de renseignements, elles se trouvent en annexe (2.10).

On note tout de même :

- Pour la variance de type linéaire, les résidus réduits, pour les valeurs de μ supérieures à 10^{-4}, sont mieux répartis que pour la variance de type binomiale où les résidus sont petits. (surtout compris entre -1 et 1).

- La structure des résidus réduits quand la variance est de type exponentielle pour les modèles 1, GTM, et 2, NIG : apparition de “lignes de ‘niveau”.

- Pour les variances de type binomiale et linéaire du modèle 5, les résidus réduits sont mieux répartis. Cependant, il y a un nombre beaucoup plus élevé de forts résidus réduits supérieurs à 2 en valeurs absolue, d’où sans doute une moins bonne modélisation.

Graphique des résidus réduits représentés sur le champ :

Une deuxième façon de comparer les résidus réduits est de les représenter sur le champ de l’expérience. Les graphiques 2.7 à 2.16 ci-après représentent, pour chaque modèle, à gauche les résidus réduits positifs et à droite les résidus réduits négatifs.

Le coin en haut à gauche et le champ central de maïs à grains bleu ne comportent pas de capteurs. Sur les graphiques, ils apparaissent en blanc.
Par ailleurs, sur les graphiques des résidus réduits positifs, les résidus réduits négatifs apparaissent en blanc et vice versa pour les graphiques des résidus réduits négatifs.

1. Fonction de Variance de type binomiale :
   Les figures (2.5), (2.6), (2.7), (2.8) et (2.9)représentent les graphiques des résidus réduits pour les modèles 1 à 5.

   Pour les modèles 1 à 4, les résidus réduits sont surtout compris entre -1 et 1 .

   Pour le modèle 5, ils sont plus grands mais il y a de forts résidus positifs supérieurs à 2 un peu partout du côté droit du champ (sens du vent) et quelques résidus forts négatifs dans le coin haut droit à côté du champ de maïs à grains bleus.

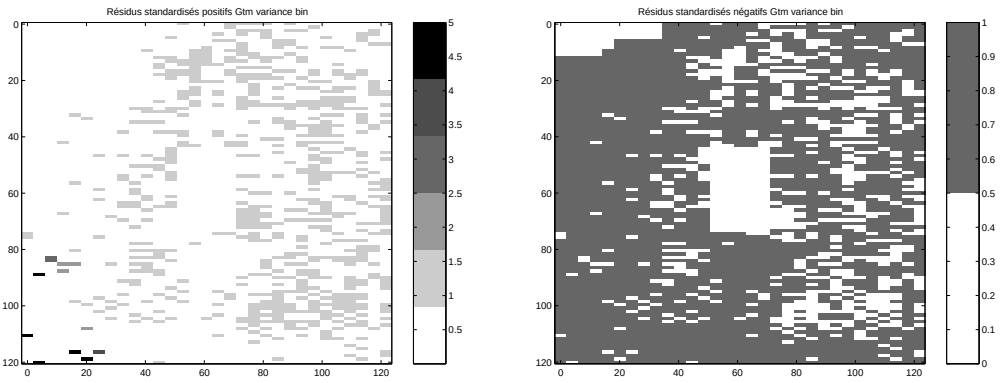


FIG. 2.5 – Résidus réduits sur le champ pour le modèle 1, GTM, avec la variance de type binomiale

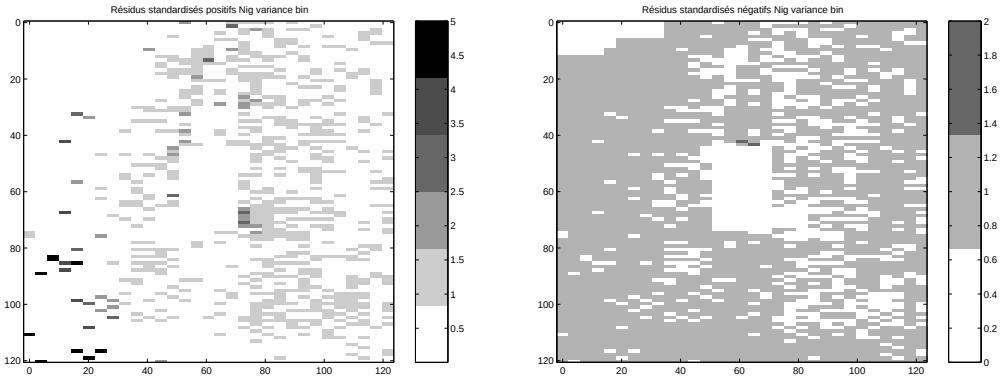


FIG. 2.6 – Résidus réduits sur le champ pour le modèle 2, NIG, avec la variance de type binomiale

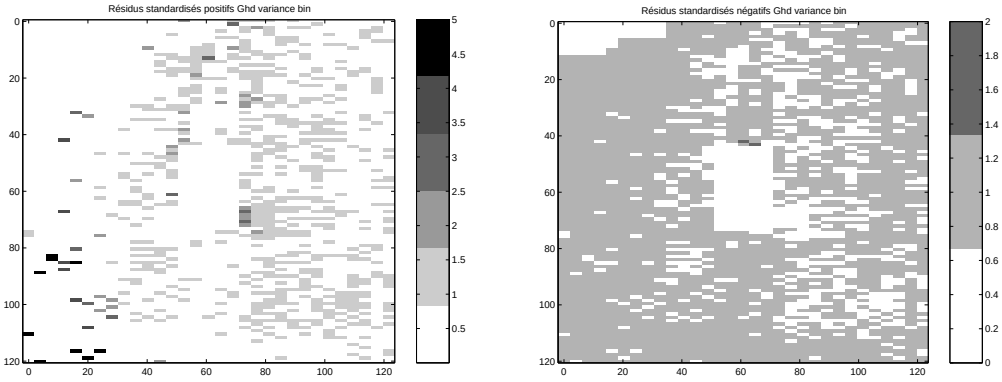


FIG. 2.7 – Résidus réduits sur le champ pour le modèle 3, GHD, avec la variance de type binomiale

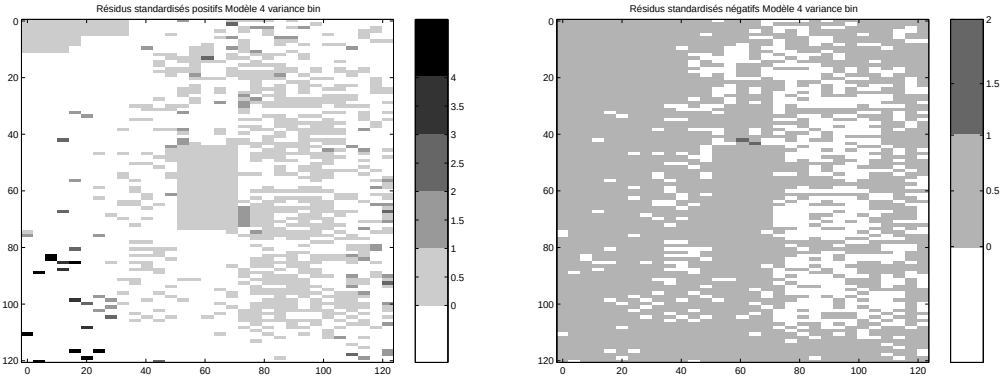


FIG. 2.8 – Résidus réduits sur le champ pour le modèle 4 avec la variance de type binomiale

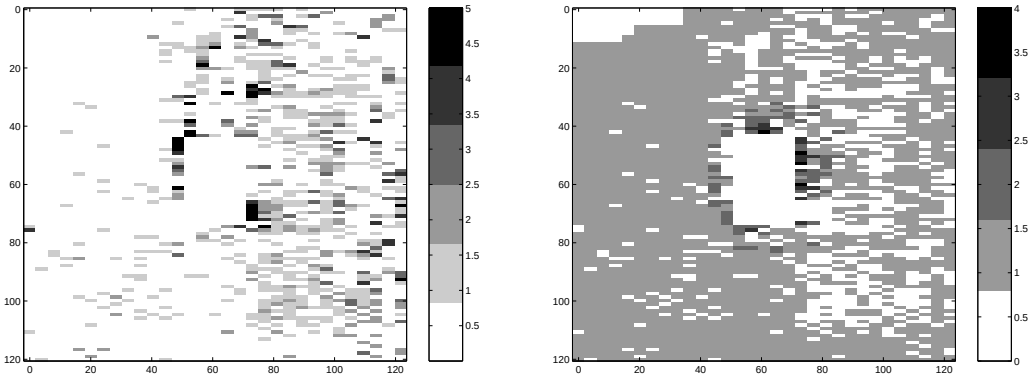


FIG. 2.9 – Résidus réduits sur le champ pour le modèle 5 avec la variance de type binomiale

2. Fonction de Variance de type linéaire :

   Les figures (2.10), (2.11), (2.12), (2.13) et (2.14) représentent les graphiques des résidus réduits pour les modèles 1 à 5.

   Tout d'abord, dans l'ensemble, ces graphiques sont meilleurs que ceux de la variance de type binomiale car les résidus réduits sont compris entre -2 et 2.

   On remarque aussi que, pour les modèles 2, NIG, et 3, GHD, les graphiques sont semblables.

   En comparant les graphiques des modèles 1, GTM, et 2, NIG, on note de forts résidus pour les deux modèles au milieu en haut du champ ainsi que sur le côté gauche en bas. Par contre, on note de forts résidus pour le modèle 1, GTM, sur le bord droit du champ, que l'on ne retrouve pas sur le graphique des résidus réduits du modèle 2, NIG. Cela confirme le fait que le modèle 2, NIG, est meilleur que le modèle 1, GTM, comme on l'avait vu précédemment.

   Pour le modèle 4, on constate l'apparition de résidus positifs élevés sur la droite du champ comparé au modèle 2. Ce modèle est donc moins bon.

   Pour le modèle 5, les graphiques pour les deux modèles de variance sont semblables. Ainsi, les résidus élevés du côté gauche du champ, pour les trois premiers modèles, ont disparu. Par contre, il y a de forts résidus positifs supérieurs à 2 un peu partout du côté droit du champ et des résidus forts négatifs dans le coin haut droit à côté du champ de maïs à grains bleus. Ce modèle est donc moins bon.

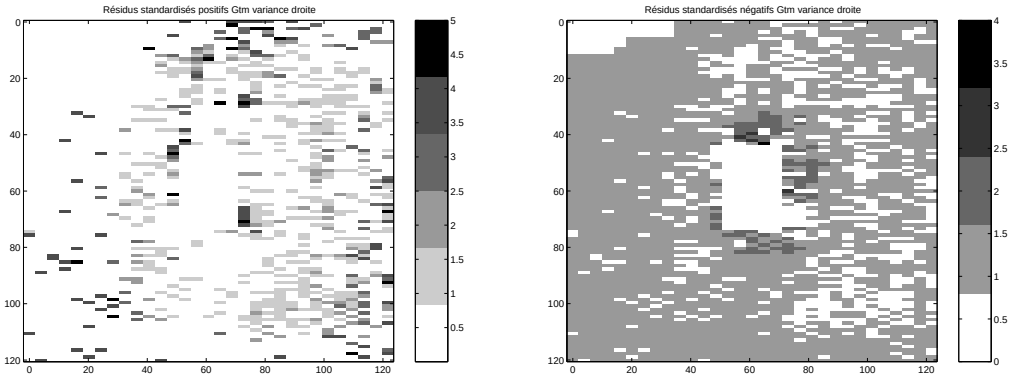


FIG. 2.10 – Résidus réduits sur le champ pour le modèle 1, GTM, avec la variance de type linéaire

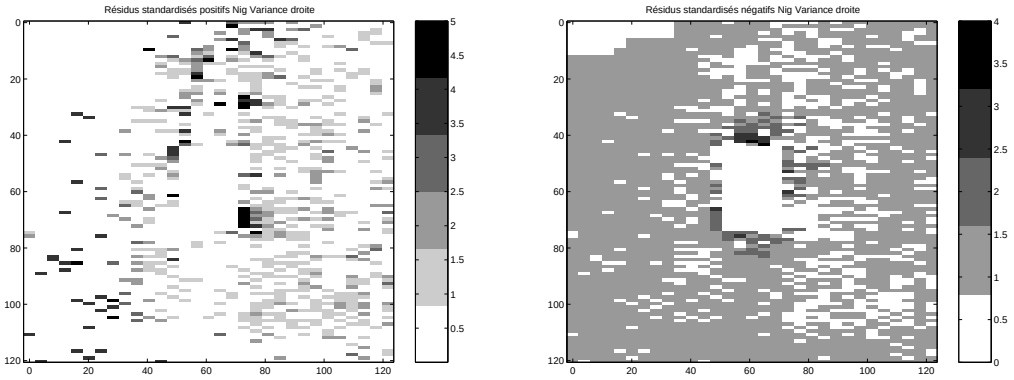


FIG. 2.11 – Résidus réduits sur le champ pour le modèle 2, NIG, avec la variance de type linéaire

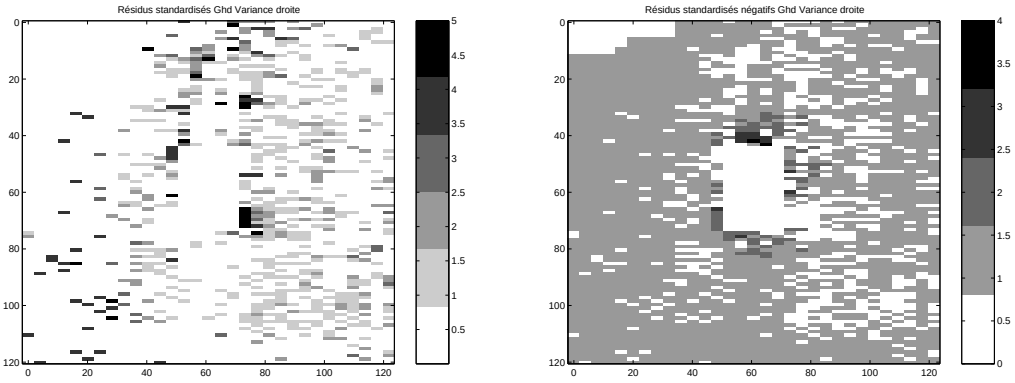


FIG. 2.12 – Résidus réduits sur le champ pour le modèle 3, GHD, avec la variance de type linéaire

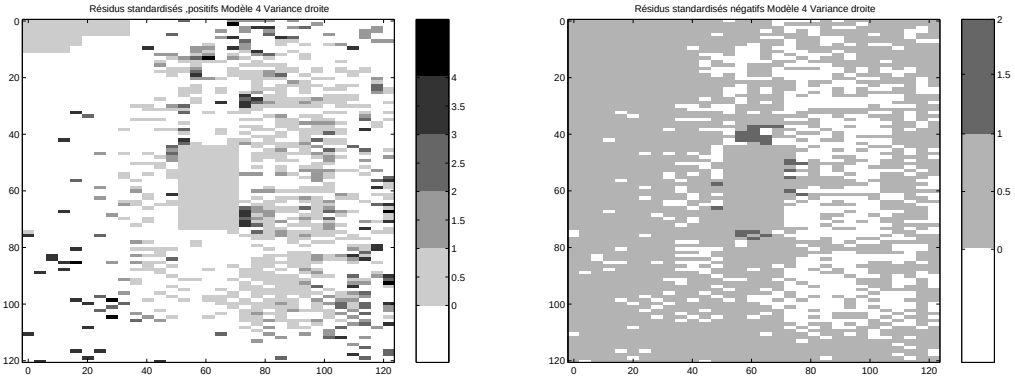


FIG. 2.13 – Résidus réduits sur le champ pour le modèle 4 avec la variance de type linéaire

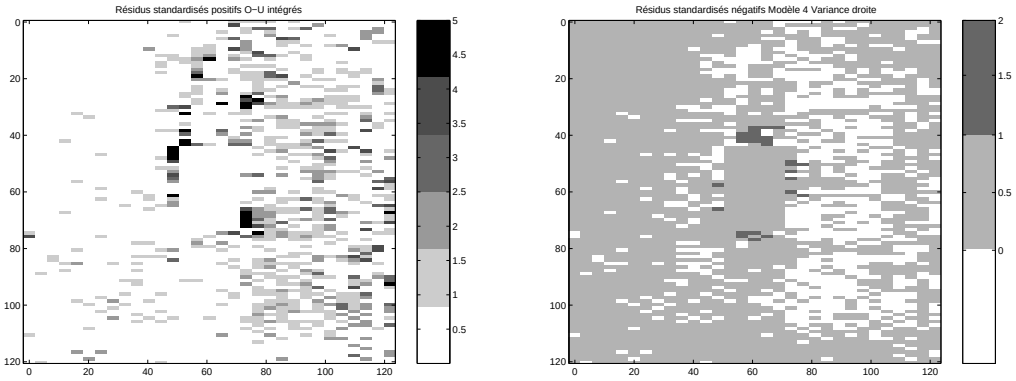


FIG. 2.14 – Résidus réduits sur le champ pour le modèle 5 avec la variance de type linéaire

3. Fonction de Variance de type exponentielle :
   Les figures (2.15) et (2.16), représentent les graphiques des résidus réduits pour les modèles 1 et 2.

   On constate l'apparition d'une structure pour les résidus négatifs tout autour du champ de maïs à grains bleus pour les deux modèles. Cela nous conduit à exclure cette fonction de variance.

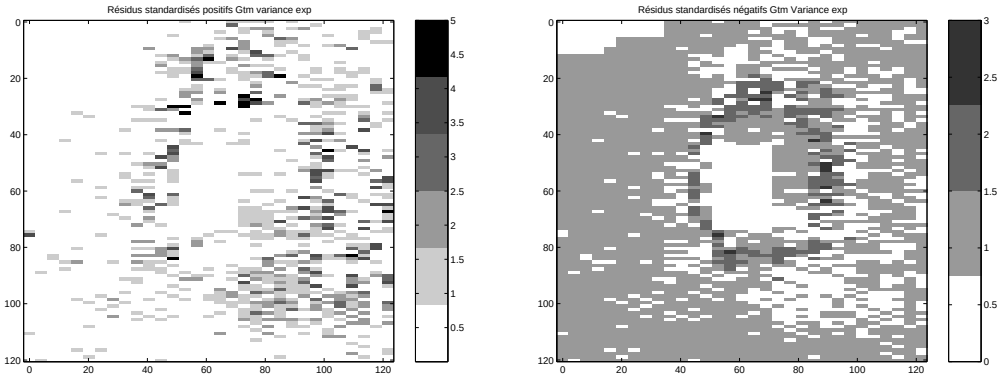


FIG. 2.15 – Résidus réduits sur le champ pour le modèle 1, GTM, avec la variance de type exponentielle

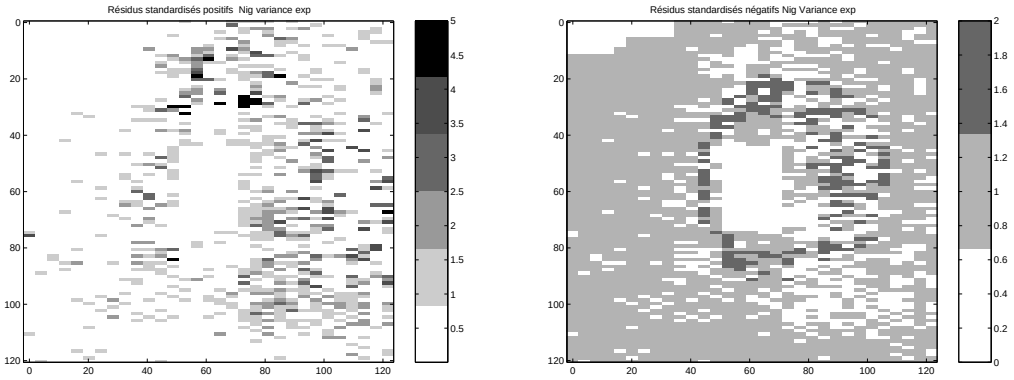


FIG. 2.16 – Résidus réduits sur le champ pour le modèle 2, NIG, avec la variance de type exponentielle

L'analyse des résultats et toutes ces études graphiques nous conduisent à choisir

le modèle 2 : NIG avec une modélisation de la fonction de variance de type linéaire

2.7.4 Courbes des fonctions de dispersion individuelles

On peut également tracer les courbes des fonctions de dispersion individuelles avec les différentes valeurs des paramètres obtenues lors des estimations, pour la variance de type linéaire.
On présente ici les courbes des fonctions de dispersion individuelles suivant l'axe du vent, celles-ci permettant de mieux visualiser les renseignements apportés que les courbes en trois dimensions.

On remarque que la fonction de dispersion individuelle de la GTM décroît beaucoup plus rapidement que celles des modèles NIG et GHD, et vers des valeurs quasiment nulles.
Pour les modèles NIG et GHD, on constate que la courbe de la NIG est au-dessus de celle du modèle GHD.
Pour le modèle 5, la courbe de dispersion commence par décroître très rapidement puis elle se stabilise au niveau de la courbe du modèle 3, GHD.

De façon plus précise, les fonction de dispersion individuelles des modèles 1 à 5 admettent des maxima en $(0,0)$ valant respectivement 0.29, 0.045, 0.049, 0.30 et 0.09.
Le modèle 4 admet donc un maximum plus élevé. Sa courbe décroît au début plus rapidement que les modèles 2 et 3 et ensuite prend des valeurs semblables au modèle 2.

De plus, on remarque que dans tous les cas, les fonctions de dispersion individuelles ne sont pas isotropes. En effet, dans le sens opposé à celui du vent, les courbes décroissent plus rapidement que dans le sens du vent où les courbes sont plus étalées. Les variables aléatoires R et Θ ne sont donc pas indépendantes comme l'avaient supposé Tufto *et al* (1997) et comme on l'avait déjà vu au paragraphe 2.5.3.

2.7.5 Étude des ajustés

On peut aussi, pour compléter notre étude, représenter pour les différentes modélisations, lorsque la fonction de variance est de type linéaire, les ajustés sur le champ et les valeurs observées sur ce champ.
Les graphiques pour les différents modèles se trouvent en annexe (2.10).
Comme on l'a remarqué au paragraphe précédent, les modèles 1, GTM, et 4 sont plus concentrés autour du champ bleu que les modèles 2, NIG, et 3, GHD.
Le modèle 5 prend plus en compte le côté gauche du champ par rapport aux modèles 1 à 3. De plus, les ajustés prennent des valeurs entre 10^{-5} et 0.5 Pour le modèle 1, GTM, les ajustés prennent des valeurs comprises entre 10^{-12} et 0.5 et pour les modèles 2, NIG, 3, GHD, et modèle 4, les ajustés prennent des valeurs entre 10^{-9} et 0.5 .

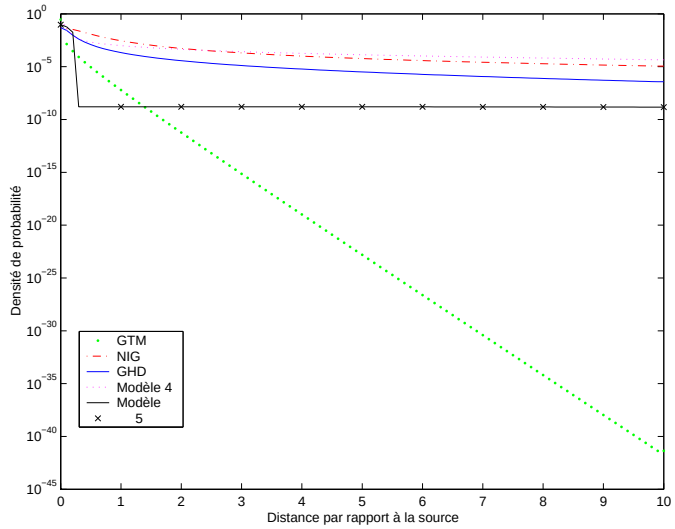


FIG. 2.17 – Fonctions de dispersion individuelles pour les cinq modèles et la variance de type linéaire, dans la direction du vent

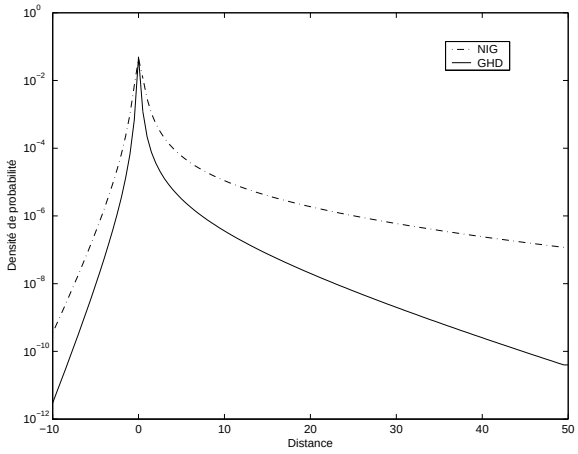


FIG. 2.18 – Fonctions de dispersion individuelles pour les modèles 2 et 3 et la variance de type linéaire, dans l'axe du vent

2.7.6 Comparaison avec les paramètres physiques

a) Description des paramètres physiques

Deux types de paramètres physiques sont utilisés dans les modèles proposés : des paramètres biologiques et des paramètres basés sur les données météorologiques.

Les paramètres biologiques :

1. La différence de hauteur h entre les plantes mâles et femelles, mesurée sur un échantillon de 25 plantes.

2. La vitesse de sédimentation représentant le paramètre f_z, calculée à partir de mesures également.

Les paramètres physiques basés sur les données météorologiques :

Pour les calculer, Klein *et al* (2003) ont utilisé les données de Météo-France pour la direction et l'intensité du vent.

1. Les paramètres f_x et f_y représentent l'intensité moyenne du vent dans chaque direction. Elles sont calculées directement à partir des données météo.
2. Les paramètres de variance τ_x, τ_y et τ_z sont des paramètres liés à la turbulence. Une étude des turbulences dans une canopée de maïs a montré que :

   * Il n'y a pas de différence notable entre τ_x et τ_y.

   * Les valeurs des paramètres de turbulence varient suivant la hauteur où le grain de pollen se trouve. Dans le tableau se trouvent donc les valeurs minimales, moyennes et maximales, calculées à l'aide des données météo également.

Tous ces paramètres sont reportés dans le tableau ci-dessous.

	Vitesse moyenne du vent		
Paramètres	Minimum	Moyenne	Maximum
Drift vertical, f_z $(m.s^{-1})$		0.183	
Différence de hauteur, h (m)		0.831	
Vitesse du vent $(m.s^{-1})$		1	
Drift horizontal : f_x $(m.s^{-1})$		-0.056	
f_y $(m.s^{-1})$		0.998	
Variance verticale, τ_z $(m.s^{-1})$	0.35	1.175	2
Variances horizontales, $\tau_x = \tau_y$ $(m.s^{-1})$	0.65	1.325	2

TAB. 2.8 – Paramètres physiques basés sur les mesures

b) Estimation des paramètres physiques

Dans le tableau suivant, se trouvent les estimations des paramètres physiques calculées à partir des estimations du paragraphe (2.6.4) pour le modèle 2, NIG, pour différentes hypothèses :

– Les variables aléatoires N_k sont supposées être des binomiales (modèle proposé par Klein *et al* (2003)).

- La variance des variables aléatoires N_k est de type binomiale (avec un paramètre de dispersion).
- La variance des variables aléatoires N_k est de type linéaire (avec un paramètre de dispersion).

Enfin, dans la dernière colonne, se trouve le modèle 3, GHD, avec la variance des variables aléatoires N_k de type linéaire.

Paramètres	NIG Klein *et al.*	NIG Var de type bin	NIG Var de type lin	GHD Var de type lin
Drift horizontal : f_x $(m.s^{-1})$	-0.074	-0.042	-0.061	-0.036
f_y $(m.s^{-1})$	1.74	1.37	1.22	0.742
Variance verticale, τ_z $(m.s^{-1})$	2.37	1.65	1.51	1.33
Variances horizontales, $\tau_x = \tau_y$ $(m.s^{-1})$	5.70	3.83	3.51	2.75

TAB. 2.9 – Estimations des paramètres physiques pour les différents modèles

Pour le modèle 4, les estimations faites avec une variance de type linéaire donnent :

Paramètres $(m.s^{-1})$	f_x $(m.s^{-1})$	f_y $(m.s^{-1})$	τ_z $(m.s^{-1})$	$\tau_x = \tau_y$ $(m.s^{-1})$	β
Estimations	-0.154	1.081	0.105	0.974	0.346

TAB. 2.10 – Estimations des paramètres physiques pour le modèle 4 avec une variance de type linéaire

Pour le modèle 5, les estimations faites avec une variance de type linéaire donnent :

Paramètres $(m.s^{-1})$	τ_z	τ_x	τ_y	v_0^x	v_0^y	Norme de v_0
Estimations	0.371	0.318	0.427	0.106	0.604	0.613

TAB. 2.11 – Estimations des paramètres physiques pour le modèle 5 avec une variance de type linéaire

c) Analyse

- Klein *et al.* (2003) avaient multiplié par deux les paramètres relatifs à la vitesse du vent car, en comparant les paramètres physiques calculés et estimés, il avait

remarqué que les valeurs étaient différentes avec la vitesse de vent observée alors qu'avec une vitesse de vent doublée, l'ensemble des estimations se rapprochaient. Une raison évoquée était le fait d'avoir négligé dans le modèle l'existence d'une vitesse minimale pour l'émission des grains de pollen.

- Pour le modèle 2, NIG, avec les modélisations de fonctions de variance que nous avons introduites, on constate :
* Pour la fonction de variance de type binomiale, les résultats sont déjà améliorés, par rapport à ceux de Klein *et al* (2003).
* Pour la fonction de variance de type linéaire, les résultats sont encore meilleurs. En effet, on n'a pas besoin de doubler la vitesse moyenne du vent et donc d'envisager des hypothèses supplémentaires . On obtient des résultats convenables pour les paramètres f_x, f_y et τ_z. Seul le paramètre de variance horizontal τ est nettement supérieur à la valeur calculée.

- Si l'on compare enfin les modèles 2 et 3 avec une fonction de variance de type linéaire, on remarque que pour les paramètres f_x, f_y, le modèle 2 est meilleur ; et pour le paramètre de variance horizontal τ, l'écart entre l'estimation du modèle 3 et la valeur calculée se réduit par rapport aux autres modèles.

Pour le modèle 4, l'estimation des paramètres f_y et τ est correcte. Par contre pour le paramètre f_x, elle est sur-estimée. Pour τ_z, on obtient une valeur sous-estimée mais il est à noter que certains phénomènes de turbulence ont été pris en compte dans le terme de friction.

Pour le modèle 5, on voit qu'on obtient une estimation du paramètre τ_z (0.371) proche de la valeur minimale calculée (0.35). Pour les paramètres τ_x et τ_y, on constate qu'ils ont des valeurs assez proches (ce qui confirme l'hypothèse de les supposer égaux).
Enfin, cette modélisation permet d' obtenir des renseignements sur la valeur de la vitesse minimale d'émission des grains de pollen : environ 0.6 $m.s^{-1}$.

Cela nous conforte dans le choix d'une fonction de dispersion individuelle de la forme NIG, avec une fonction de variance de type linéaire.

2.7.7 Discussion sur l'estimation des paramètres

L'analyse des résultats obtenus pour l'estimation des paramètres des différents modèles proposés et les différentes méthodes de validation employées (critère de sélection de modèle ; études graphiques des résidus réduits et des courbes des fonctions de dispersion individuelles et comparaison avec les paramètres physiques) conduisent à choisir une fonction de dispersion individuelle de la forme NIG, avec la fonction de variance de type linéaire.

Pour les modèles 4 et 5, on a modélisé plus finement les composantes du vent dans le plan vertical et le plan horizontal (x, y). On aurait donc pu s'attendre à obtenir de meilleurs résultats. Or, il y a plus de résidus élevés que dans les autres modèles,

en particulier autour du champ de maïs bleu, donc les modèles sont moins adaptés aux données.
On remarque tout de même, grâce aux études graphiques, que le modèle 4 se rapproche du modèle NIG. Cependant les relations entre paramètres du modèles et paramètres physique sont moins bonnes.
Cela peut s'expliquer par le fait qu'au lieu d'utiliser la définition de la fonction de dispersion globale dans le cadre discret, on l'a utilisée dans le cadre continu. En effet, la puissance des ordinateurs utilisés n'a pas permis d'estimer les paramètres du modèle dans le cadre discret. La fonction de dispersion individuelle est définie à l'aide d'une intégrale (généralisée) et pour calculer la fonction de dispersion globale, il faut alors effectuer une double sommation. On a donc utilisé la définition dans le cadre continu ce qui a permis de nous ramener à une seule intégrale à calculer et on a dû effectuer certaines approximations en raison des problèmes numériques rencontrés.

2.8 Conclusion et perspectives

Tout ce travail est basé sur l'étude du grain de pollen de maïs, vu comme une particule soumise à un champ de force. A partir d'hypothèses faites sur la trajectoire du grain de pollen, les fonctions de dispersion individuelles sous-jacentes ont été calculées. En effet, on a vu l'intérêt de travailler avec cette fonction : elle permet d'estimer une quantification robuste de la dispersion efficace du pollen et ainsi d'effectuer des prédictions. On a vu également que c'était un avantage du point de vue des temps de calculs avec un ordinateur par rapport aux modèles physiques basés sur des simulations de trajectoires (Loubet *et al*, 2004).

Partant du travail de Klein *et al* (2003), on a tout d'abord remis en cause l'hypothèse faite sur les variables aléatoires N_k, représentant le nombre de grains bleus sur les épis de maïs. Klein *et al* (2003) supposaient que les N_k suivaient des lois binomiales de paramètres $(n_k, \mu(x_k, y_k))$. On a fait seulement des hypothèses sur les fonctions d'espérance et de variance des N_k.
D'une part, pour des raisons biologiques (hétérogénéité temporelle du vent et variabilité de la période de fécondation), l'hypothèse d'indépendance des génotypes ne semble pas vérifiée, et on a introduit un paramètre de dispersion dans la fonction de variance (Collett 1991).
D'autre part, après étude des données expérimentales, en plus d'une fonction de variance de type binomiale, on a introduit deux autres fonctions de variance : l'une de type linéaire et l'autre de type exponentielle.

On a donc ensuite essayé ces différentes modélisations pour la fonction de variance sur des modèles mécanistes :
Dans un premier temps, on a repris les modèles proposés par Klein et al (2003), basés sur un mouvement brownien avec drift dans $\mathbb{R}^3$ et différentes hypothèses pour la définition du temps de fécondation (prédominance de la végétation limitant les trajectoires ou prédominance du sol).

Dans un deuxième temps, on a introduit deux nouveaux modèles :
- Les composantes horizontales de la trajectoire sont modélisées par deux mouvements browniens avec drift et la composante verticale, par un processus d'Ornstein-Uhlenbeck intégré.
- Les composantes de la vitesse dans le plan horizontal sont modélisées par des processus d'Ornstein-Uhlenbeck non stationnaires, permettant ainsi d'introduire une vitesse minimale d'émission des grains de pollen, souvent démontrée. Klein *et al* (2003) avançaient cette raison pour leurs estimations moyennes par rapport aux valeurs des paramètres physiques liés à la vitesse du vent.

Pour les deux derniers modèles, des approximations ont été nécessaires. En particulier, pour le modèle 54, on a utilisé un théorème d'approximation du temps de premier passage, d'une courbe dépendant du temps, pour un mouvement brownien.

La seconde partie du travail a consisté à estimer les paramètres des différentes familles paramétriques de fonctions individuelles obtenues précédemment, à partir des données expérimentales disponibles. N'ayant fait des hypothèses que sur l'espérance et la variance des variables aléatoires N_k, une méthode de quasi-vraisemblance a été utilisée pour l'estimation.

Une fois l'estimation des paramètres effectuée, les résultats ont été analysés afin de déterminer quel modèle mécaniste et quelle fonction de variance ajuste le mieux les observations. Pour cela, on a tout d'abord utilisé un critère de sélection de modèle type Akaïke et effectué des tests sur les trois premiers modèles, les deux premiers modèles étant des sous-modèles du troisième.
Ensuite, on a employé des méthodes graphiques : étude des résidus réduits en fonction des ajustés, sur le champ ; étude des courbes des fonctions de dispersion obtenues, en particulier dans l'axe du vent.
Enfin, on a comparé les valeurs estimées des paramètres physiques avec celles mesurées par les biologistes au cours de l'expérience.
Les résultats ont amené à choisir le modèle 2, NIG ; les composantes de la trajectoire du grain de pollen pouvant être modélisées par trois mouvements browniens avec drift indépendants et le temps de fécondation défini comme le temps de premier passage au niveau des fleurs femelles. De plus, la modélisation de la fonction de variance avec un paramètre de dispersion et une fonction de type linéaire permet d'obtenir de meilleurs résultats que Klein *et al* (2003).
Ainsi, les modélisations plus précises des modèles 4 et 5 ne se sont pas révélées être meilleures contrairement à ce qu'on aurait pu penser.

L'objectif, maintenant, est d'utiliser ces résultats pour la modélisation du flux de pollen du maïs en milieu hétérogène, c'est-à-dire lors d'une discontinuité du couvert végétal.
Les capacités prédictives des modèles développés dans ce chapitre sont limitées aux conditions climatiques et topographiques dans lesquelles ils ont été développées. Ils ne permettent donc pas de décrire de façon satisfaisante la dispersion du pollen au niveau d'une discontinuité dans un paysage, qui aurait pour effet d'accélérer ou de ralentir la trajectoire.

En effet, le but de tout ce travail, à la base, est de pouvoir prédire quelle distance doit séparer deux champs pour qu'il n'y ait pas de pollution génétique d'une culture par l'autre.
La difficulté essentielle est que l'hypothèse (H1) :
"Toutes les plantes dispersent leur pollen suivant la même fonction de dispersion individuelle γ"
ne tient plus.
Effectivement, la fonction de dispersion individuelle dépend de la distance entre la plante émettrice, liée à cette fonction, et la discontinuité du couvert.

C'est le sujet du chapitre suivant.

Annexes

2.9 Annexe A : Lois hyperboliques généralisées sur $\mathbb{R}$

Loi du premier temps de passage d'un niveau b pour un mouvement brownien avec drift

Dans le cas d'un mouvement brownien standard, la loi du temps de premier passage d'un niveau h est donnée par la propriété suivante :

Propriété 2.6 *:*
Soit $(W_t)_{t\geq 0}$ un mouvement brownien standard. On définit le temps d'atteinte du niveau b, pour $b \in \mathbb{R}^*$, par $T_b = \inf\{t > 0, W_t = b\}$.

(i) *T_b est presque sûrement fini.*

(ii) *La transformée de Laplace est pour $\alpha > 0$:*

$$E\left(e^{-\alpha T_b}\right) = e^{-|b|\sqrt{2\alpha}} \tag{2.23}$$

(iii) *T_b admet pour fonction de densité :*

$$f(t,b) = \frac{|b|}{\sqrt{2\pi}\, t^{3/2}} \exp\left(-\frac{b^2}{2t}\right) \mathbf{1}_{t>0} \tag{2.24}$$

Dans le cas d'un mouvement brownien avec drift, on a :

Propriété 2.7 *:*
Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité.
On considère le processus $(Z_t)_{t\geq 0}$ défini par $Z_t = at + W_t$ où $(W_t)_t$ est un $\mathcal{F}_t$-mouvement brownien standard et $a \neq 0$. On définit le temps d'atteinte du niveau b, pour $b \neq 0$, par $R_b = \inf\{t > 0, Z_t = b\}$. Alors, R_b admet pour fonction de densité :

$$f(t,b,a) = \frac{|b|\, e^{ab}}{\sqrt{2\pi}} \frac{1}{t^{3/2}} \exp\left(-\frac{b^2}{2t} - \frac{a^2}{2}t\right) \mathbf{1}_{t>0}$$

Si a et b sont de même signe alors R_b est presque sûrement fini.
On dit que R_b admet pour loi une IG (Inverse Gaussian) de paramètres (a, b).

Définition 2.10 *Plus généralement, une variable aléatoire suit une Generalized Inverse Gaussian (GIG) de paramètres (α, ρ, η) si elle admet pour fonction de densité*

$$f_{GIG}(\alpha, \rho, \eta; t) = \frac{1}{I(\alpha, \rho, \eta)} t^{-\alpha} \exp\left(-\rho t - \frac{\eta}{t}\right) \mathbf{1}_{t\geq 0} \tag{2.25}$$

où la constante de normalisation $I(\alpha, \rho, \eta)$ est égale à

$$I(\alpha, \rho, \eta) = 2\left(\frac{\eta}{\rho}\right)^{(1-\alpha)/2} \mathcal{K}_{1-\alpha}(2\sqrt{\rho\eta})$$

α est un réel et η et ρ sont positifs.

$K_\nu(x)$ représente la fonction modifiée de Bessel de troisième ordre (pour une définition voir Abramowitz et Stegun (1972) par exemple).

Ainsi, une variable aléatoire admettant pour loi une Inverse Gaussienne de paramètres (a, b) est en fait une GIG de paramètres $\alpha = \frac{3}{2}$, $\rho = \frac{a^2}{2}$ et $\eta = \frac{b^2}{2}$.

Démonstration de la propriété 2.7 :
On écrit $Z_t = at + W_t = m(t) + W_t$ où $m(t) = at = \int_0^t a\, ds$.
On définit $L_t = \exp\left(\int_0^t a\, dW_s - \frac{1}{2}\int_0^t a^2\, ds\right) = \exp\left(aW_t - \frac{a^2}{2}\, t\right)$.
On sait que $(L_t)_{t\geq 0}$ est une martingale (en appliquant le critère de Novikov par exemple).

Le théorème de Girsanov nous dit alors qu'il existe une unique mesure de probabilité $\mathbb{Q}$ définie sur $(\Omega, \mathcal{F}_\infty)$, où $\mathcal{F}_\infty = \sigma(\mathcal{F}_s, s \geq 0)$, tel que pour tout $t \geq 0$ et pour tout $A \in \mathcal{F}_t$, $\mathbb{Q}(A) = \int_A L_t\, d\mathbb{P}$.
C'est-à-dire que le processus $(Z_t)_t$ sous $\mathbb{P}$ a la même loi que le processus $(W_t)_t$ sous $\mathbb{Q}$. Ainsi, la loi sous $\mathbb{P}$ du temps d'atteinte R_b du processus $(Z_t)_t$ est aussi la loi sous $\mathbb{Q}$ du temps d'atteinte du niveau b, noté T_b, du mouvement brownien $(W_t)_t$.
On a donc pour $t \geq 0$,

$$\mathbb{P}(R_b \leq t) = \mathbb{Q}(T_b \leq t) = E_{\mathbb{P}}(\mathbf{1}_{\{T_b \leq t\}} L_t)$$

L'ensemble $\{T_b \leq t\}$ appartient à $\mathcal{F}_t \cap \mathcal{F}_{T_b} = \mathcal{F}_{t\wedge T_b}$.
Sur cet ensemble, on a $L_{t\wedge T_b} = L_{T_b} = \exp\left(ab - \frac{1}{2}a^2 T_b\right)$ car $t \wedge T_b = T_b$ et $W_{T_b} = b$.
On a alors, en utilisant un théorème d'arrêt pour les martingales

$$\begin{aligned}
\mathbb{P}(R_b \leq t) &= E_{\mathbb{P}}\left[\mathbf{1}_{\{T_b\leq t\}} L_t\right] = E_{\mathbb{P}}\left[\mathbf{1}_{\{T_b\leq t\}} E(L_t|\mathcal{F}_{t\wedge T_b})\right] \\
&= E_{\mathbb{P}}\left[\mathbf{1}_{\{T_b\leq t\}} L_{t\wedge T_b}\right] \qquad \text{car } (L_t)_t \text{ est une martingale} \\
&= E_{\mathbb{P}}[\mathbf{1}_{\{T_b\leq t\}} L_{T_b}] = E_{\mathbb{P}}\left[\mathbf{1}_{\{T_b\leq t\}} \exp\left(ab - \frac{1}{2}a^2 T_b\right)\right] \qquad (2.26) \\
&= \int_0^t \exp\left(ab - \frac{1}{2}a^2 s\right) f(s,b)\, ds
\end{aligned}$$

où $f(s, b)$ a été calculée en (2.24).
On en déduit que la densité de R_b est

$$f(t, b, a) = \frac{|b|\, e^{ab}}{\sqrt{2\pi}} \frac{1}{t^{3/2}} \exp\left(-\frac{b^2}{2t} - \frac{a^2}{2}t\right) \mathbf{1}_{t>0}$$

Remarque : En faisant tendre t vers $+\infty$ dans (2.26), on obtient

$$\mathbb{P}(R_b < +\infty) = \exp(ab) E\left[\exp\left(-\frac{1}{2}a^2 T_b\right)\right]$$

Alors, en utilisant (2.23), on obtient $\mathbb{P}(R_b < +\infty) = \exp(ab - |ab|)$.
Ainsi, un mouvement brownien avec drift $a \neq 0$ atteint le niveau $b \neq 0$ avec probabilité égale à 1 si et seulement si a et b ont le même signe.

Loi Normale Inverse Gaussienne : NIG

La fonction de Bessel modifiée de troisième ordre, notée $K_\nu(x)$, est par définition solution d'une équation différentielle de $\mathbb{R}$ (voir Abramovitz et Stegun 1972 par exemple pour plus de détails).
Elle vérifie la propriété suivante (par exemple Prudnikov *et al* 1986) :

Propriété 2.8 *: Pour tout réel ν et pour tous réels positifs p, q, on a*

$$\int_0^{+\infty} t^{\nu-1} \exp\left(-pt - \frac{q}{t}\right) dt = 2\left(\frac{q}{p}\right)^{\nu/2} \mathcal{K}_\nu(2\sqrt{pq}) \tag{2.27}$$

Définition 2.11 *Une variable aléatoire admet pour loi une Normale Inverse Gaussienne, notée NIG, si elle admet pour fonction de densité*

$$g(x;\alpha,\beta,\mu,\delta) = a(\alpha,\beta,\mu,\delta)\, q\left(\frac{x-\mu}{\delta}\right)^{-1} K_1\left(\delta\alpha\, q\left(\frac{x-\mu}{\delta}\right)\right) \exp(\beta x) \tag{2.28}$$

où la constante de normalisation $a(\alpha,\beta,\mu,\delta)$ est égale à
$a(\alpha,\beta,\mu,\delta) = \pi^{-1}\alpha \exp\left(\delta\sqrt{\alpha^2-\beta^2} - \beta\mu\right)$
et $q(x) = \sqrt{1+x^2}$
De plus, on a $\delta > 0$, $\mu \in \mathbb{R}$ et $0 \leq |\beta| < \alpha$. On notera cette loi $NIG(\alpha,\beta,\mu,\delta)$.

Remarque : La loi normale $\mathcal{N}(\mu,\sigma^2)$ apparaît comme un cas limite pour $\beta = 0$, α tendant vers $+\infty$ et $\delta/\alpha = \sigma^2$.
Et la loi de Cauchy est le cas spécial $NIG(0,0,1,0)$.

En fait, la loi NIG peut être vue comme la loi d'un processus subordonné. Dans notre travail, nous nous sommes intéressés au cas particulier suivant

Propriété 2.9 *On considère le processus bidimensionnel $(X_t, Z_t)_{t\geq 0}$ défini par*

$$\begin{cases} X_t = ct + W_t^1 \\ Z_t = at + W_t^2 \end{cases} \tag{2.29}$$

où $(W_t^1)_t$ et $(W_t^2)_t$ sont deux mouvements browniens standards indépendants avec $a > 0$ et $c \in \mathbb{R}$.
On définit pour $b > 0$, $R_b = \inf\{t > 0, Z_t = b\}$.
(On pourrait également prendre a et b négatifs tous les deux.)
Alors la loi de la variable aléatoire X_{R_b} est une NIG de paramètres $\alpha = \sqrt{a^2+c^2}$, $\beta = c$, $\mu = 0$ et $\delta = |b|$.

Démonstration :
Soit φ une fonction mesurable positive. Les variables aléatoires X_t et R_b étant indépendantes, d'après une formule de Bayes, on a

$$E(\varphi(X_{R_b})) = \int_x \int_t \varphi(x)\, f_{X_t}(x)\, f_{R_b}(t)\, dt\, dx = \int_x dx\, \varphi(x) \left\{ \int_{t>0} f_{X_t}(x)\, f_{R_b}(t)\, dt \right\}$$

où f_{X_t} représente la densité de X_t et f_{R_b} représente la densité du temps d'atteinte R_b. Ainsi la densité de X_{R_b} est $g(x) = \int_{t>0} f_{X_t}(x)\, f_{R_b}(t)\, dt$.
D'après la propriété (2.7), on connaît f_{R_b}. D'autre part, X_t suit une loi normale de moyenne ct et de variance t d'où

$$\begin{aligned} g(x) &= \int_{t>0} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(x-ct)^2}{2t}\right) \times \frac{|b|\, e^{ab}}{\sqrt{2\pi}} \frac{1}{t^{3/2}} \exp\left(-\frac{b^2}{2t} - \frac{a^2}{2}t\right) dt \\ &= \frac{|b|\, e^{ab}}{2\pi}\, e^{2cx} \int_{t>0} \frac{1}{t^2} \exp\left(-\frac{(b^2+x^2)}{2t} - \frac{(a^2+c^2)}{2}\, t\right) dt \end{aligned} \tag{2.30}$$

Alors, en utilisant la formule (2.27), on obtient

$$g(x) = \frac{e^{ab}}{2\pi}\, e^{cx} \frac{\sqrt{a^2+c^2}}{\sqrt{1+x^2/b^2}} K_1\left(|b|\sqrt{a^2+c^2}\sqrt{1+x^2/b^2}\right)$$

Ainsi, X_{R_b} suit une NIG de paramètres $\alpha = \sqrt{a^2+c^2}, \beta = c, \mu = 0$ et $\delta = |b|$.

Lois hyperboliques généralisées

La fonction de densité définie en (2.28), représentant la loi NIG, est, en fait, un cas particulier de la loi hyperbolique généralisée, introduite par Barndorff-Nielsen (1997).

Définition 2.12 *Une variable aléatoire suit une loi hyperbolique généralisée si elle admet pour fonction de densité*

$$\begin{aligned} h(x;\lambda,\alpha,\beta,\delta,\mu) = a(\lambda,\alpha,\beta,\delta) \left(\delta^2 + (x-\mu)^2\right)^{(\lambda-0.5)/2} \\ \times K_{\lambda-0.5}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right) \exp\left(\beta(x-\mu)\right) \end{aligned} \tag{2.31}$$

où $a(\lambda,\alpha,\beta,\delta) = \dfrac{(\alpha^2-\beta^2)^{\lambda/2}}{\sqrt{2\pi}\,\alpha^{\lambda-0.5}\,\delta^{\lambda}\,K_\lambda\left(\delta\sqrt{\alpha^2-\beta^2}\right)}$

On a $\alpha > 0$ *paramètre de forme,* β *avec* $0 \leq |\beta| < \alpha$ *paramètre de biais,* $\mu \in \mathbb{R}$ *paramètre de localisation,* $\delta > 0$ *paramètre d'échelle et* $\lambda \in \mathbb{R}$.
On notera cette loi $GHD(\lambda,\alpha,\beta,\delta,\mu)$.

Remarque : Pour $\lambda = -1/2$, on trouve la loi Normale Inverse Gaussienne.
Une autre valeur intéressante est le cas où $\lambda = 0.5$.

Comme au paragraphe précédent, si Z suit une $GIG(1-\lambda, \delta, \sqrt{\alpha^2 - \beta^2})$ et si la loi de X sachant $\{Z = z\}$ est une loi normale de moyenne $\mu + \beta z$ et de variance z, alors la loi de X est une $GHD(\lambda, \alpha, \beta, \delta, \mu)$.
Cette propriété est démontrée dans le cas de la dimension 2 au paragraphe 2.5.1

Il existe une extension de la définition d'une GHD au cas multidimensionnel. En particulier, au paragraphe 2.5.1 de ce chapitre se trouve la définition dans le cas de la dimension 2.

Ci-dessous, les représentations graphiques pour $\alpha = 1, \beta = 0.9, \delta = 1, \mu = 0$ d'une $GHD(\lambda, \alpha, \beta, \delta, \mu)$ pour $\lambda = -1/2, 1/2$ et 0 .

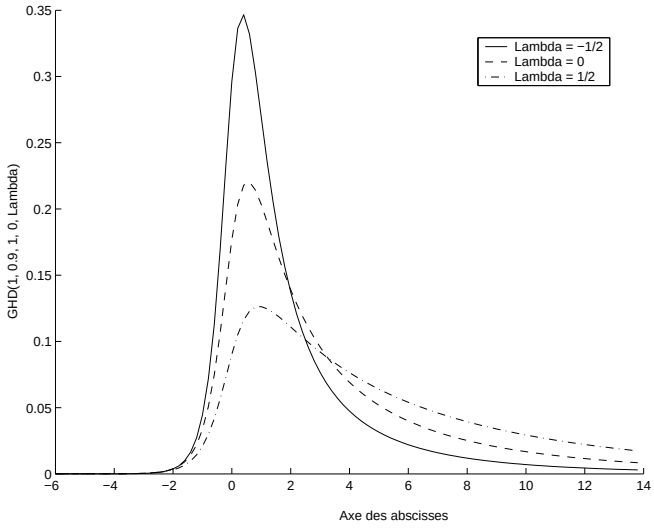
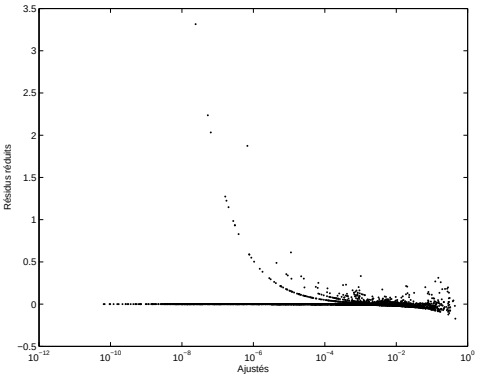


FIG. 2.19 – Représentations graphiques pour $\alpha = 1, \beta = 0.9, \delta = 1, \mu = 0$ d'une $GHD(\lambda, \alpha, \beta, \delta, \mu)$ pour $\lambda = -1/2, 1/2$ et 0

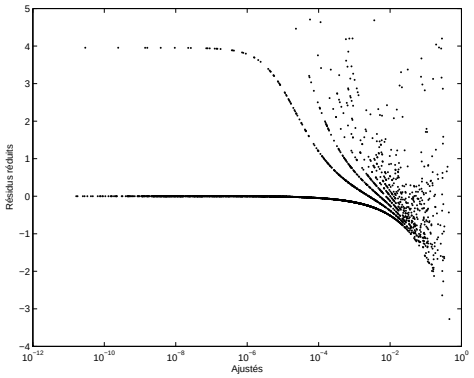
2.10 Annexe B : Graphiques des résidus

Graphiques des résidus réduits en fonction des ajustés

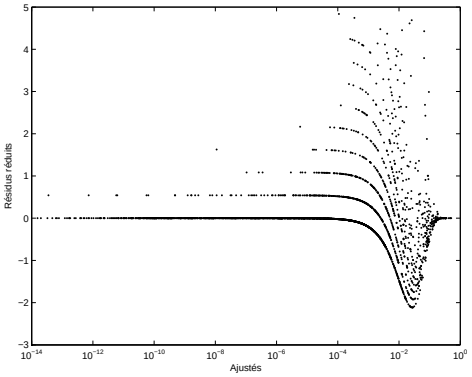
Modèle 1 : GTM



Variance de type binomiale



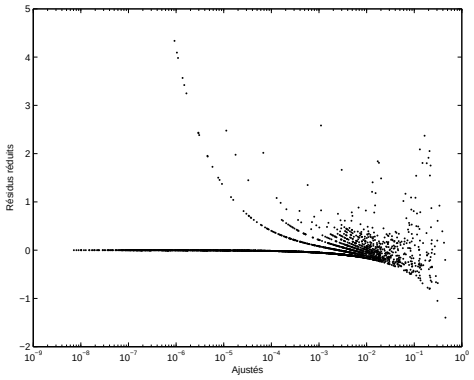
Variance de type linéaire



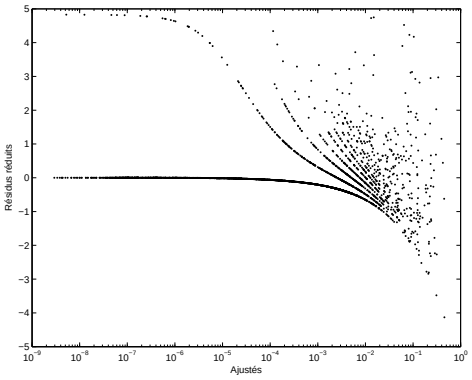
Variance de type exponentielle

FIG. 2.20 – Résidus réduits par rapport aux ajustés pour le modèle 1, GTM

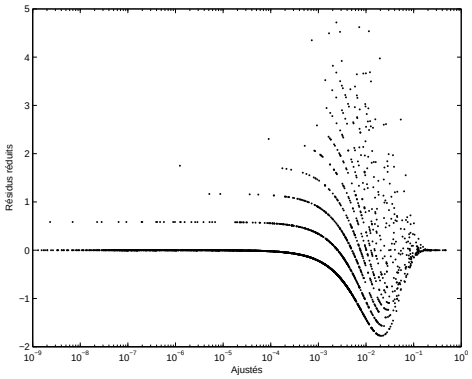
Modèle 2 : NIG



Variance de type binomiale



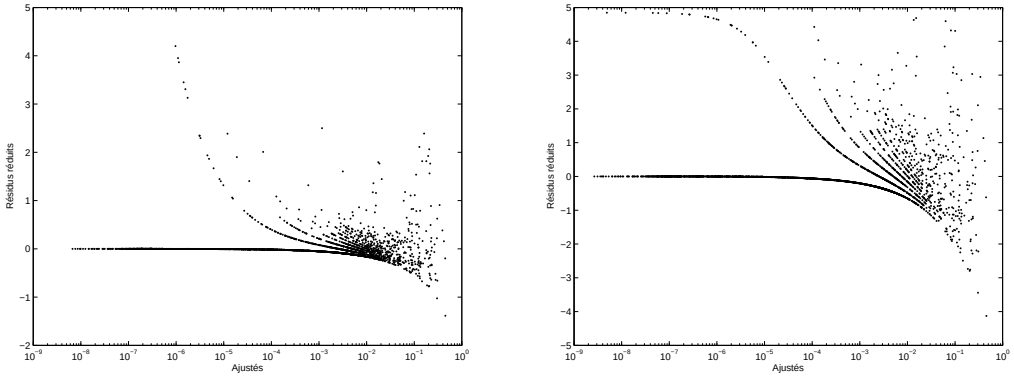
Variance de type linéaire



Variance de type exponentielle

FIG. 2.21 – Résidus réduits par rapport aux ajustés pour le modèle 2, NIG

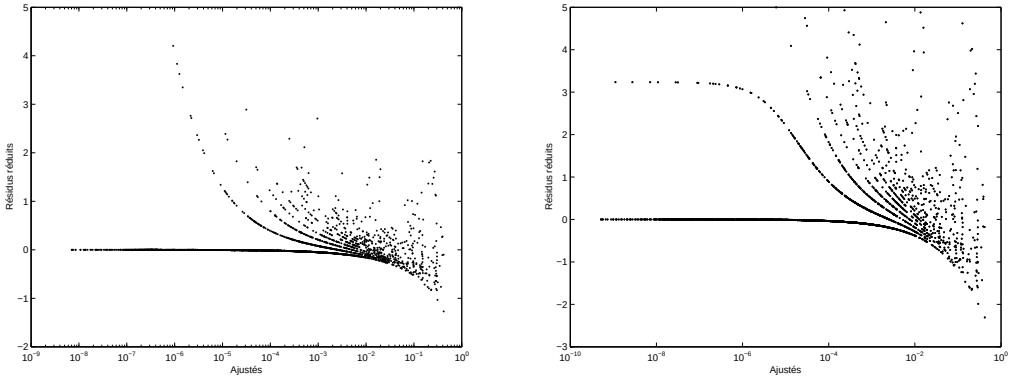
Modèle 3 : GHD



Variance de type binomiale Variance de type linéaire

FIG. 2.22 – Résidus réduits par rapport aux ajustés pour le modèle 3, GHD

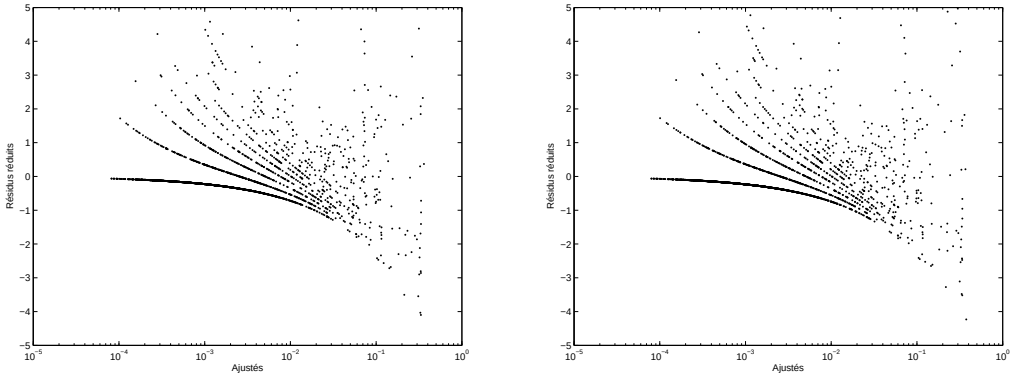
Modèle 4



Variance de type binomiale Variance de type linéaire

FIG. 2.23 – Résidus réduits par rapport aux ajustés pour le modèle 4

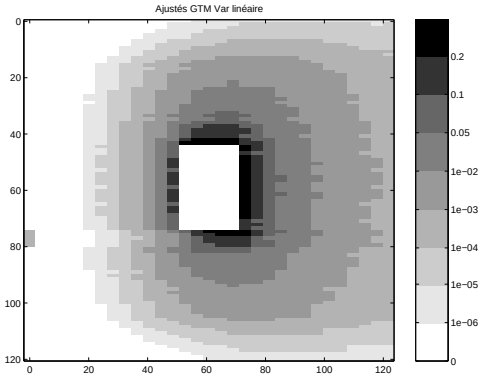
Modèle 5



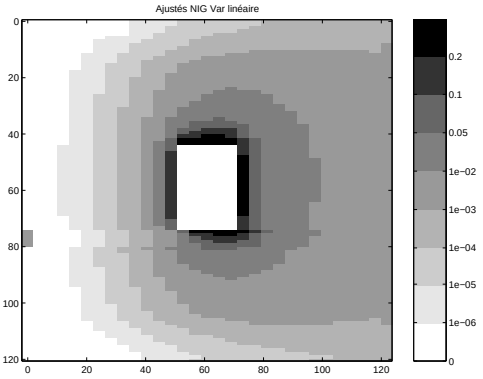
Variance de type binomiale Variance de type linéaire

FIG. 2.24 – Résidus réduits par rapport aux ajustés pour le modèle 5

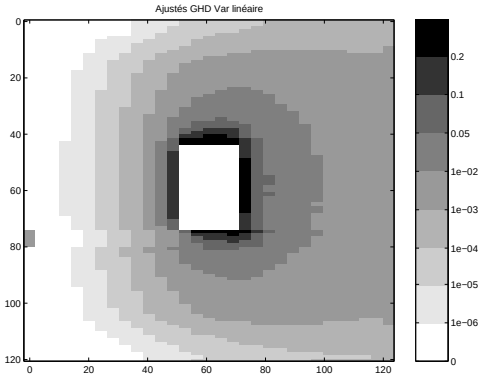
Graphiques des ajustés sur le champ



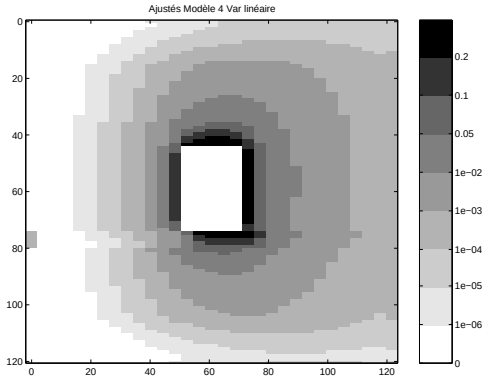
Ajustés pour le modèle 1, GTM



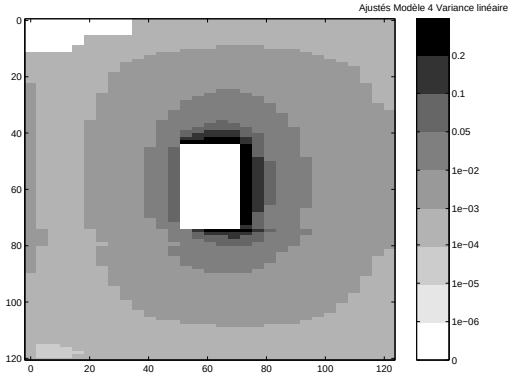
Ajustés pour le modèle 2, NIG



Ajustés pour le modèle 3, GHD



Ajustés pour le modèle 4



Ajustés pour le modèle 5

FIG. 2.25 – Ajustés sur le champ pour les différents modèles avec la fonction de variance de type linéaire

2.11 Annexe C : Parametric models for corn pollen dispersal using diffusion processes and statistical estimation

This is a joined work with Catherine Larédo.

2.11.1 Introduction

During the last thirty years, the use of genetically modified plants increased in many domains such as medicine, agriculture (plants resistance to insects and diseases for example) or environment. In opposition to improvements and possible economic advantages due to GMO culture, arises the issue of possible risks on health and environment. Hence it is important to study pollen dispersion in order to handle these risks. Indeed the pollen dispersion can involve a diffusion of modified genes to a non modified variety of the same specie. That is why it is necessary to define a minimal distance between two fields in order to ensure that the cross pollination will be minimal and inferior to a tolerance level (1 % in Europe now).

In the following the study focus on corn pollen dispersion which is based only on wind. One can remark that this study type can be applied to seed or spore dispersal (Portnoy and Wilson 1993, McCartney and Fitt 1998). On the other hand, for example, the oilseed rape is a more complex case because pollen is dispersed by wind and insects (Lavigne *et al* 1998).
The dispersion can be studied on two levels : on the level of an agricultural landscape (pollen dispersion study over long distances) or on the level of two fields contiguous or separated by another culture. The work presented here consists in studying the pollen dispersion in homogeneous landscape, i.e. when two fields are contiguous.
The used experiment type permits to observe directly the proportion of grains having the gene marker in sampled plants at different distances from a marked source (first proposed by Bateman 1947). In particular for corn, most of experiments are performed with marker non trangenic coloring grains in blue.

Then two approaches exist to study pollen flows. The first one is to use the backward dispersal function which represents the observed dispersion and is directly deduced from the observations (Morris *et al* 1994). However this function is tied to the experiment design. But a major objective is to be able to predict pollen cloud and gene movements using obtained results (in particular to develop models valid in various agricultural landscapes and different climatic conditions). Hence it is necessary to have a more robust measure of dispersion independent of the field design. This is the second approach using the individual dispersal function (Tufto *et al* 1997, Nurminiemi *et al* 1998 or Klein *et al* 2003) and called "forward" method in opposition

to the "backward" one. But this function is not directly observable.

Two main methods are used for the individual dispersal function. The first method, called empirical method, is to consider exponential or decreasing power type functions, or a compromise (Nurminiemi *et al* 1998 , Klein 2000). But those functions are isotropic and therefore not adapted to the corn case for which the dispersion is based only on wind. So it is necessary to take into account a wind dominant direction. That leads to use a second method using "mechanist" models, studied in the following, as Klein *et al* 2003, Tufto *et al* 1997. In this method, the pollen grain is considered as a particle. This permits to model the corn pollen path while taking into account the major phenomenas during the dispersion : physical and biological corn pollen characteristics, parameters tied to wind (intensity and direction) and field of forces (gravity in particular).

This paper is organized as follow. First the experiment used to estimate and to validate a model is described. Second the backward and forward dispersion functions are defined and the relation between them are detailed. The next section aims to present the mathematical model and the statistical method used to estimate parameters of different proposed parametric individual dispersal functions. A quasi-likelihood method is used with the introduction of an overdispersion parameter. The next section describes studied models for corn pollen grain path using diffusion processes, in particular Brownian motion with drift and Ornstein-Uhlenbeck integrated processes. It also describes the relation existing between this path and the associated individual dispersal function. Finally in the last sections the results are given and compared to the observed data to choose the most fitted model. Moreover the estimated parameters values are compared with corn biological and physical parameters and available meteorological data. Those data were obtained independently of observed data.

2.11.2 Data description and statistical problem

a) Data description

The experimental data result from an experiment on a corn field performed during summer 1998 near Montargis (France) by AGPM.

The main characteristic of corn pollen flow dispersion is that it is based only on wind. Pollen is emitted at the male flower height taken as origin of the vertical axis and the pollination happens at the female flower height h with h negative.

Two pollen sources are used. One contains homozygous plants having a dominant genetic marker coloring the corn grains in blue. The other source contains common homozygous corn plants not genetically marked. The plants from the second source are used as receptors, and each of those offspring plant being genetically marked (so being blue colored) results from a fecundation by a corn seed from the first source. Therefore, the number of blue-marked grains in an ear reflects exactly the pollination intensity by the first source upon the second source.

The corn field from the experiment is a square of length approximately 120 meters (the upper left corner is cropped by a road, therefore there is no culture in that

corner). It has been cultivated according to the following model : 155 lines spaced of 0.8 m, and on each line 800 plants are sown, spaced of 0.15 m. In the middle of this field a square of length 20 meters of plants producing blue grains has been sown. Elsewhere yellow corn has been sown.

On the whole, 2937 ears have been sampled. In fact to be more accurate 101 lines were sampled (1 line every 3 lines, and every line at proximity of the blue corn square). On each line 31 ears were taken, so about one every 4 meters. Then, on each ear, the number of blue grains has been counted. (For computations, the total number of grains on a ear is taken constant equal to 394 (estimated mean grains number on a corn ear).

Below is given the representation of blue grains observed proportions on each sampled ear in Figure 2.26.

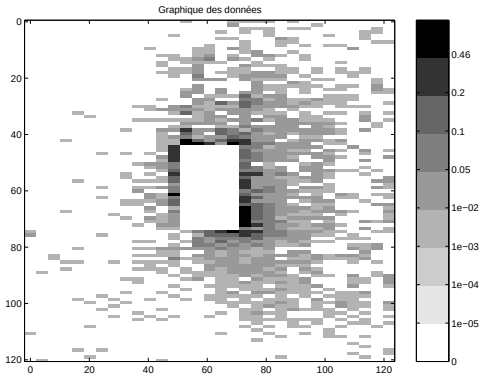


FIG. 2.26 – Blue grains observed proportions on sampled ears.

Moreover some meteorological data are available. It permits to compare them with some estimated physical parameters that can be extracted from the proposed models, also for validation purpose.

b) Modeling the pollen dispersal and pollination

Two functions can be defined to describe pollen dispersal.
The first one is linked to phenomena occurring at the pollen source and is called the forward approach.
The effective individual dispersal function $\gamma(x, y)dxdy$ describes the probability that a pollen grain emitted at point $(0, 0)$ falls and fertilizes a plant located in $((x, y), (x + dx, y + dy))$. It is a two-dimensional probability density function.

The second function is the backward dispersal function, $\mu(x, y)$. It represents the probability that a corn grain located at point (x, y) is pollinated by the marked source.
This function is the most intuitive and the most used by biological people. As a fact pollen is overabundant. This implies that pollination occurs according to the pollen

cloud composition above a plant (This takes into account the competition between pollen grains in the pollen cloud).

The framework in this document is :
(A1) *Pollen is dispersed following the same individual dispersal function γ for each plant.*
(A2) *The same amount of pollen is produced by all plants, whatever their genotype.*
(A3) *There is no intrinsic genetic differences between both plants (viability, germination rate, fertilization effectiveness).*

Assumption (A1) doesn't take into account pollination at the field extremities where there is a border effect. However it is justified for large fields because only few plants are impacted by this effect.

Under these assumptions, there exists a relation between the backward and the individual dispersal function. This approach was used by Tufto *et al* (1997) and Klein *et al* (2003).

Let set two different sources of pollen being in competition : a source S_A of blue marked plants at points $(x_k, y_k)_{k=1,\dots,S_A}$ and a source S_B of yellow plants at points $(x_k, y_k)_{k=1,\dots,S_B}$. Then the pollen cloud composition above a plant located at (x, y), defining $\mu(x, y)$ can be written :

$$\mu(x,y) = \frac{\sum_{k=1}^{S_A} \gamma(x - x_k, y - y_k)}{\sum_{k=1}^{S_A} \gamma(x - x_k, y - y_k) + \sum_{k=1}^{S_B} \gamma(x - x_k, y - y_k)} \tag{2.32}$$

A continuous approximation can be used when the plants density is high enough. Then the discrete sum is replaced by an integral and the μ function can be written as a non linear convolution product : $\mu(x,y) = \dfrac{\gamma * \mathbb{I}_A(x,y)}{\gamma * \mathbb{I}_{A \bigcup B}(x,y)}$

The number of blue grains on an ear located at point (x, y), deriving of the experiment, represents the noisy observation of the backward function $\mu(x, y)$. However μ is linked to the experiment design (shape, size for example) contrary to the individual dispersal function γ. Hence the aim is to estimate γ.

c) The statistical problem

Let n_k denote the total number of grains on an ear located at point (x_k, y_k) and N_k the number of observed blue grains on this ear.
Since the observations are counting data, the first approach is to assume that the random variables N_k are binomials with parameters $(n_k, \mu(\theta; x_k, y_k))$. But this leads to a form of noise modeling which is too specific.

Therefore a more general statistical model is considered :

$$N_k = n_k \mu(\theta; x_k, y_k) + \varepsilon_k \text{ with } E(\varepsilon_k) = 0 \text{ and } Var(\varepsilon_k) = \sigma_k^2 n_k v(\theta, b; (x_k, y_k))$$

where the $(\varepsilon_k)_k$ are assumed independent and $\sigma_k^2 > 0$.
Hence this is a non-linear deconvolution problem which is not classical in statistic. That is why, in the following, only parametric individual dispersal functions are

proposed, under the shape $\{\gamma(\theta; x, y), \theta \in \Theta \; and \; (x, y) \in \mathbb{R}^2\}$ where Θ is a subset of $\mathbb{R}^p$ $(p \geq 1)$.

2.11.3 Parametric models for pollen dispersal and pollination

There exists mainly two ways to model the individual dispersal function. The first way, called empirical method, is to consider isotropic, exponential, or decreasing power type functions, or a compromise (Nurminiemi *et al* (1998), for rapeseed Klein, 2000). For corn the dispersion is only based on wind and therefore it is necessary to take into account a wind dominant direction. This leads to another approach using "mechanist" models where the pollen grain is considered as a particle. Let $P_t = (X_t, Y_t, Z_t)$ be the position of a pollen grain at time t and $P_0 = (0, 0, 0)$. and T_F the pollination time which represents the time where the pollen grain path stops on a female flower.

Property 2.1 *By taking into account the three following assumptions :*
(i) T is a positive random variable a.s. finite with density f on $\mathbb{R}^{+*}$.
(ii) the process (X_t, Y_t) admits a density $g_t(x, y)$ for all $t > 0$.
(iii) T and (X_t, Y_t) are independent.
Then the process (X_T, Y_T) has a density on $\mathbb{R}^2$

$$\gamma(x, y) = \int_0^{+\infty} g_t(x, y) f(t) dt \tag{2.33}$$

Proof : Let ϕ a positive measurable function. According to a Bayes formula and the fact that T_F and (X_t, Y_t) are independent random variables, we have :

$$\begin{aligned} \mathrm{E}[\phi(X_{T_F}, Y_{T_F})] &= \int\int_{(x,y)\in\mathbb{R}^2} \int_{t=0}^{+\infty} \phi(x, y) g_T(x, y | T = t) f(t) \, dt \, dx \, dy \\ &= \int\int_{(x,y)\in\mathbb{R}^2} \phi(x, y) \left\{ \int_{t=0}^{+\infty} g_t(x, y) f(t) \, dt \right\} dx \, dy \end{aligned}$$

Hence the result is obtained.
Summary of previous results

a) Modeling the path

In this section, the pollen grain path is modeled by a Brownian motion with drift of $\mathbb{R}^3$. This permits to take into account in particular mean wind intensity and atmospheric turbulence due to wind. So the path (P_t) can be written :

$$\begin{cases} dX_t = f_x dt + \tau_x dB_t^1 \\ dY_t = f_y dt + \tau_y dB_t^2 \\ dZ_t = f_z dt + \tau_z dB_t^3 \end{cases} \tag{2.34}$$

with τ_x, τ_y, τ_z being positive and f_z being assumed negative. Also $(B_t^i)_{i=1,2,3}$ are three independent Brownian motions.

The parameters f_x and f_y represent the wind mean velocity and the parameter f_z represents the velocity of the pollen grain resulting from gravity. The variance parameters of these stochastic processes represent atmospheric turbulences.
In this case, the joined distribution of (X_t, Y_t) is a normal distribution with mean $(f_x t, f_y t)$ and covariance matrix $\begin{pmatrix} \tau_x^2 t & 0 \\ 0 & \tau_y^2 t \end{pmatrix}$
Klein *et al* (2003) proposed three models for pollination time which will be used in the following for estimation.

Model 1 : Exponential hitting time The first assumption is that vegetation is the main factor that stops pollen grains. Tufto *et al* (1997) modeled the stopping of the path after a random time T_e being an exponential distribution of parameter λ (positive) and independent of the path. Moreover there is fecundation only when a pollen grain hits a female flower. Hence the T_F pollination time distribution is defined as the distribution of the stopping time T_e conditionally to the event $\{Z_{T_e} = h\}$. Then T_F is a GIG (Generalized Inverse Gaussian) with parameters $\frac{1}{2}$, $\lambda + \frac{f_z^2}{2\tau_z^2}$ and $\frac{h^2}{2\tau_z^2}$. (Klein *et al* 2003)

Model 2 : First hitting time of a level Here the pollen grain path stops when it hits the female flower height. Hence the pollination time is the first-passage time at the height h and $T_F = T_h = inf\{t > 0, Z_t = h\}$. As f_z is negative, T_h is a.s finite and is a GIG with parameters $\frac{3}{2}$, $\frac{f_z^2}{2\tau_z^2}$ and $\frac{h^2}{2\tau_z^2}$. (see for example Karatzas and Shreve, 1991)

Model 3 : Generalization The pollination time is a GIG with density function on $\mathbb{R}$ defined by :

$$f_{GIG}(\alpha, \rho, \eta; t) = \frac{1}{I(\alpha, \rho, \eta)} t^{-\alpha} e^{-\rho t - \frac{\eta}{t}} \mathbb{I}_{t \geq 0}$$

Here the α parameter is unspecified. It includes both previous models and permits to take into account the vegetation influence on the pollination time.

b) Individual dispersal functions

For the three cases, using property (2.1), the individual dispersal functions expression are explicit. In fact, if T_F is a GIG then the dispersal function γ is a GHD (Generalized Hyperbolic Distribution) (Barndorff-Nielsen 1997).
For model 1, γ has a simplified expression and is called GTM (for Generalized Tufto Model).
For model 2, γ is a NIG (Normal Inverse Gaussian) having the shape

$$f_{NIG}(\lambda_z, \lambda_x, \lambda_y, \delta_x, \delta_y; x, y) = \frac{\delta_x \delta_y e^{\lambda_z}}{2\pi} \frac{(q(x,y)^{-1/2} + p^{1/2})}{q(x,y)} e^{-\sqrt{pq(x,y)}} e^{\delta_x \lambda_x x + \delta_y \lambda_y y}$$

And for the more generalized model 3, γ has the shape

$$f_{GHD}(\alpha, \lambda_z, \lambda_x, \lambda_y, \delta_x, \delta_y; x, y) = \frac{\lambda_z^{1-\alpha}\delta_x\delta_y(p/q(x,y))^{\frac{\alpha}{2}}}{2\pi}\frac{\mathcal{K}_\alpha(\sqrt{pq(x,y)})}{\mathcal{K}_{1-\alpha}(\lambda_z)}e^{\delta_x\lambda_x x+\delta_y\lambda_y y}$$

where $p = \lambda_z^2 + \lambda_x^2 + \lambda_y^2$, $q(x,y) = 1 + \delta_x^2x^2 + \delta_y^2y^2$ and

$$\alpha,\ \delta_x = \frac{\tau_z}{\tau_x|h|},\ \delta_y = \frac{\tau_z}{\tau_y|h|},\ \lambda_x = \frac{f_x h}{\tau_x\tau_z},\ \lambda_y = \frac{f_y h}{\tau_y\tau_z},\ \lambda_z = \frac{f_z h}{\tau_z^2}$$

The parameters vector $\theta = (\delta_x, \delta_y, \lambda_x, \lambda_y, \lambda_z, \alpha)$ lies in $\Theta = (\mathbb{R}^+)^2 \times (\mathbb{R})^2 \times \mathbb{R}^+ \times \mathbb{R}$. ($K_\nu(x)$ is the modified Bessel function of third kind (for a detailed definition see for example Abramovitz and Stegun 1972)).

2.11.4 Proposed models

From a physical point of view, previous models are only based on path components of the pollen grain. In this section, a pollen grain is still seen as a particle. But enhanced models are introduced by taking into account the components of the pollen grain speed.

a) Model 4

In the first new model, the process (X_t, Y_t) representing components in the horizontal plan is kept as a two-dimensional Brownian motion with drift. But the vertical velocity component is modeled using the Langevin equation. It permits to introduce the fact that the pollen grain is subject to a force resulting from gravity but also from a wind force. Hence, the path $(P_t)_{t\geq 0}$ is modeled by

$$\begin{cases} dX_t = f_x dt + \tau_x dB_t^1 \\ dY_t = f_y dt + \tau_y dB_t^2 \\ dZ_t = V_t dt \end{cases} \tag{2.35}$$

with

$$dV_t = (c_z - \beta V_t)dt + \tau_z dB_t^3,\ \ V_0 = v_0 \tag{2.36}$$

It is assumed that τ_x, τ_y, τ_z are positive, $\beta > 0$ and $c_z < 0$. Moreover $(B_t^i)_{i=1,2,3}$ are assumed to be independent.
So V_t represents the vertical velocity of the pollen grain. c_z is a force resulting from gravity and $\tau_z dB_t^3$ represents the contributions of the force applied by the wind on the particle during its trajectory whose are not already in the friction term $-\beta V_t$.

Lemma 2.1 *The process Z_t is called an integrated Ornstein-Uhlenbeck process and can be written in the form :*

$$Z_t = \frac{c_z}{\beta}t + \left(v_0 - \frac{c_z}{\beta}\right)\frac{1 - e^{-\beta t}}{\beta} + \frac{\tau_z}{\beta}\int_0^t \left(1 - e^{-\beta(t-s)}\right) dB_s \tag{2.37}$$

Moreover we have $E(Z_t) = \frac{c_z}{\beta}t + (v_0 - \frac{c_z}{\beta})\frac{1-e^{-\beta t}}{\beta}$ *and*
$Var(Z_t) = \frac{\tau_z^2}{\beta^2}\left(t + \frac{4e^{-\beta t} - 3 - e^{-2\beta t}}{2\beta}\right)$

Proof : First, for $\beta > 0$, the equation (2.36) admits for solution

$$V_t = \frac{c_z}{\beta} + \left(v_0 - \frac{c_z}{\beta}\right)e^{-\beta t} + \tau_z e^{-\beta t}\int_0^t e^{\beta s} dB_s$$

Indeed it is sufficient to apply the Itô formula to the process $S_t = \exp(\beta t)V_t$.

The use of the Fubini theorem for stochastic integral (Protter, 1992) leads to :

$$\begin{aligned} Z_t &= \frac{c_z}{\beta}t + \int_0^t \left\{\left(v_0 - \frac{c_z}{\beta}\right)e^{-\beta s} + \tau_z e^{-\beta s}\int_0^s e^{\beta u} dB_u\right\} ds \\ &= \frac{c_z}{\beta}t + \left(v_0 - \frac{c_z}{\beta}\right)\frac{1-e^{-\beta t}}{\beta} + \frac{\tau_z}{\beta}\int_0^t \left(1 - e^{-\beta(t-u)}\right) dB_u \end{aligned} \qquad (2.38)$$

From above, we deduce directly the expectation of Z_t.
For the variance, we have

$$\begin{aligned} Var(Z_t) &= \frac{\tau_z^2}{c_z^2}\int_0^t \left(1 - e^{-\beta(t-s)}\right)^2 ds \\ &= \frac{\tau_z^2}{c_z^2}\int_0^t \left(1 + e^{-2\beta(t-s)} - 2e^{-f\beta(t-s)}\right) ds \end{aligned}$$

Hence we obtain the result.

Here, the joined distribution of (X_t, Y_t) is the same as in §2.11.3 and the pollination time is again the first-hitting time at level $h < 0$. So $\nu_h = inf\{t > 0, Z_t \leq h\}$.

Lemma 2.2 *Assuming that* $c_z < 0$ *and* $h < 0$*, the stopping time* $\nu_h = inf\{t > 0, Z_t \leq h\}$ *where* Z_t *is defined in 2.37, is almost surely finite.*

Proof : We have

$$\begin{aligned} \mathbb{P}(\nu_h = +\infty) &= \mathbb{P}(\forall t, Z_t > h) = \mathbb{P}(\lim_{t\to+\infty} \searrow \bigcap_{u\leq t}\{Z_u > h\}) \\ &= \lim_{t\to+\infty} \mathbb{P}(\bigcap_{u\leq t}\{Z_u > h\}) \leq \lim_{t\to+\infty} \mathbb{P}(Z_t > h) \end{aligned}$$

Using the equality (2.37) and Lemma 2.1, we have $\mathrm{E}(Z_t) = \frac{c_z t}{\beta}(1 + O(1))$ and when t tends to infinity $h - \mathrm{E}(Z_t) > 0$. Applying the Bienayme-Tchebytchev inequality, then we obtain

$$\mathbb{P}(Z_t > h) = \mathbb{P}(Z_t - \mathrm{E}(Z_t) > h - \mathrm{E}(Z_t)) \leq \frac{Var(Z_t)}{(h - \mathrm{E}(Z_t))^2}$$

Still according to Lemma 2.1, we obtain $\mathrm{E}(Z_t) \sim \frac{c_z}{\beta}t$ and $\mathrm{Var}(Z_t) \sim \frac{\tau_z^2}{c_z^2}t$ hence $\lim_{t\to+\infty}\mathbb{P}(Z_t > h) = 0$ and the result. ◇

The distribution of ν_h is the density of the first passage time at level h for an integrated Ornstein-Uhlenbeck process, which is different from previous models. Although the calculus of the density of this kind of process has been studied for several years there is no known explicit analytic expression. So to be able to go further an approximation of the distribution of ν_h is required.

Individual dispersion function computation :

Theorem 2.1 *We assume that the process Z_t is defined by 2.37 with $c_z < 0$, $h < 0$ and $v_0 = 0$. Then the density of ν_h, defined on $\mathbb{R}^+$, can be approximated by*

$$p(t) = \frac{\beta g'(\beta t)}{\sqrt{2\pi g(\beta t)}}\left[\frac{b_z H(\beta t) + c_z}{g(\beta t)} - \frac{b_z H'(\beta t)}{g'(\beta t)}\right]\exp\left(-\frac{(b_z H(\beta t) + c_z)^2}{2g(\beta t)}\right) \quad (2.39)$$

where $b_z = \frac{-c_z}{\tau_z\sqrt{\beta}}, c_z = \frac{-\beta^{3/2}h}{\tau_z}$, $H(t) = 1-e^{-t}-t$ *and* $g(t) = t-1.5+2e^{-t}-0.5e^{-2t}$.

Proof : First, a time change leads to approximate the first passage density of a Brownian motion crossing a curved boundary, depending of time.

It is easy to see that ν_h can be written under the shape :
$\nu_h(Y) = inf\{t > 0, Y_t \le f_\theta(t)\}$ where $Y_t = \int_0^t \left(1 - e^{-\beta(t-s)}\right) dB_s$ and
$f_\theta(t) = \frac{\beta h}{\tau_z} - \frac{c_z}{\tau_z}t + \frac{c_z}{\tau_z\beta}(1 - e^{-\beta t})$.

We define the process $\tilde{Y}_t = \int_0^t \left(1 - e^{-(t-u)}\right) dB_u$. Then we have $Y_t \overset{\mathcal{D}}{=} \frac{1}{\sqrt{\beta}}\tilde{Y}_{\beta t}$. Therefore

$$\nu_h(Y) \overset{\mathcal{D}}{=} inf\{t > 0, \tilde{Y}_{\beta t} \le \sqrt{\beta} f_\theta(t)\} = \frac{1}{\beta} inf\{u > 0, \tilde{Y}_u \le \tilde{f}_\theta(u)\} = \frac{1}{\beta}\tilde{\nu}(\tilde{Y})$$

with $\tilde{f}_\theta(u) = \bar{b}_z H(u) + \bar{c}_z$, $\bar{b}_z = \frac{c_z}{\tau_z\sqrt{\beta}}$, $\bar{c}_z = \frac{\beta^{3/2}h}{\tau_z}$ and $H(u) = (1 - e^{-u} - u)$.

Note that $(\tilde{Y}_t)$ is a continuous local martingale as the result of an Itô integral, with $\tilde{Y}_0 = 0$.
Let us define $g(t) = <\tilde{Y}>_t$. Then

$$g(t) = \int_0^t \left(1 - e^{-(t-s)}\right)^2 ds = t + \frac{4e^{-t} - 3 - e^{-2t}}{2}$$

(so $\lim_{t\to+\infty} <\tilde{Y}>_t = +\infty$.)
And the function g is an increasing bijection from $[0, +\infty[$ to $[0, +\infty[$.

Define now $T(s) = inf\{t > 0, < \tilde{Y} >_t > s\}$. Therefore

$$T(s) = inf\{t > 0, g(t) > s\} = g^{-1}(s).$$

A time change theorem (for example Rogers and Williams (1994), p 64) gives $\tilde{Y}_{T(s)} \stackrel{\mathcal{D}}{=} \tilde{B}_s$ where $\tilde{B}_s$ is a Standard Brownian motion.

Then

$$inf\{s > 0, \tilde{Y}_{T(s)} \leq \tilde{f}_\theta(g^{-1}(s))\} \stackrel{\mathcal{D}}{=} inf\{s > 0, \tilde{B}_s \leq \tilde{f}_\theta(g^{-1}(s))\} = g(\tilde{\nu}(\tilde{Y}))$$

hence $\tilde{\nu}(\tilde{Y}) \stackrel{\mathcal{D}}{=} g^{-1}\left(inf\{s > 0, \tilde{B}_s \leq \tilde{f}_\theta(g^{-1}(s))\}\right)$.

Using the following property of Brownian motion $B_t \stackrel{\mathcal{L}}{=} -B_t$ and noting $a_\theta(s) = -\tilde{f}_\theta(g^{-1}(s)) = -\bar{b}_z H(g^{-1}(s) + \bar{c}_z = b_z H(g^{-1}(s)) + c_z$, we deduce that

$$\nu_h(Y) \stackrel{\mathcal{D}}{=} \frac{1}{\beta} g^{-1}\left(inf\{s > 0, \tilde{B}_s \geq a_\theta(s)\}\right) \tag{2.40}$$

Now we approximate the density of the stopping time $inf\{s > 0, \tilde{B}_s \geq a_\theta(s)\}$ by a density function noted $r_\theta(t)$.
The defined function a_θ is C^1 on the interval $(0, +\infty)$ with $a_\theta(0) = c_z > 0$ (because $h < 0, c_z < 0$ and $\beta > 0$). Moreover, a is concave (because g is convex).

Applying an approximation theorem for the first passage density of a Brownian motion crossing a curved boundary (Durbin, 1992), we obtain a density function

$$r_\theta(t) = \frac{1}{\sqrt{2\pi t}}\left(\frac{a_\theta(t)}{t} - a'_\theta(t)\right)\exp\left(-\frac{a_\theta(t)^2}{2t}\right)$$

(The error made with the approximation is lower than $\sqrt{\pi}/\Gamma(0.5) \times e(t)$ where $e(t)$ represents the maximum of $|\frac{a(r)-a(s)}{r-s} - a'(r)|$ pour $0 < s < r \leq t$. Hence when the slope of the function a varies slowly, the convergence to the true density will be quickly.)

Therefore we deduce an approximated density $p(t)$ for the density of ν_h using (2.40) :

$$p_\theta(t) = \frac{\beta g'(\beta t)}{\sqrt{2\pi g(\beta t)}}\left[\frac{b_z H(\beta t) + c_z}{g(\beta t)} - \frac{b_z H'(\beta t)}{g'(\beta t)}\right]\exp\left(-\frac{(b_z H(\beta t) + c_z)^2}{2g(\beta t)}\right)$$

◇

We can compute an individual dispersal function :

Proposition 2.1 *Assumptions are :*
- The path $(P_t)_{t\geq 0}$ is defined by (2.35) with $\tau_x = \tau_y = \tau$.
- The pollination time is $T_F = \nu_h$ and its density function is approximated by the function $p_{\theta_z}(t)$ defined above.
Then the distribution of (X_{T_F}, Y_{T_F}) admits the following density :

$$\gamma(\theta; x, y) = \frac{\lambda^2}{\pi} \exp\left(2\lambda(\delta_x x + \delta_y y)\right)$$

$$\int_0^{+\infty} \frac{1}{t} \exp\left(-(\delta_x^2 + \delta_t^2)t - \frac{\lambda^2(x^2+y^2)}{t}\right) p_1(t, \theta_z) dt$$

where $p_1(t, \theta_z) = p_{\theta_z}(\frac{t}{\beta})$ *and* $\theta_z = (b_z, c_z)$, $\theta = (\delta_x, \delta_y, \lambda, b_z, c_z)$, *with* $\delta_x^2 = \frac{f_x^2}{2\tau^2\beta}, \delta_y^2 = \frac{f_y^2}{2\tau^2\beta}$, $\lambda^2 = \frac{\beta}{2\tau^2}$ *and* $\theta \in \mathbb{R}^2 \times (\mathbb{R}^+)^3$.

Proof : The joined distribution of (X_t, Y_t) is a normal distribution with mean $(f_x t, f_y t)$ and covariance matrix $\begin{pmatrix} \tau_x^2 t & 0 \\ 0 & \tau_y^2 t \end{pmatrix}$.
Moreover according to Theorem 2.1, the ν_h density is approximated by $p_{\theta_z}(t)$.
And since (B_t^3) is independent of $(B_t^1, B_t^)$, it is clear that (X_t, Y_t) is independent of ν_h (which only depends of the vertical component (Z_t)).
Therefore applying the formula (2.33), (X_{T_F}, Y_{T_F}) has a density on $\mathbb{R}^2$ with

$$\gamma(x, y) = \int_0^{+\infty} g_u(x, y) p_{\theta_z}(u) du$$

Doing the variable change $t = \beta u$, we obtain the result.

b) Model 5

For the last proposed model, the path is modeled in a more precise way. In fact, in the plane (x, y), the velocity vector is still modeled rather than the position. As a first approximation the velocity vector components in the plane (x, y) are assumed to be two independent Ornstein-Uhlenbeck processes. It permits to introduce a minimal wind speed for pollen grains emission (assumption done by biologists). Moreover the vertical component Z_t is still a Brownian motion with drift as in first models.
The pollination time T_F is defined by the first-passage time at the level h. Hence T_F is again a GIG (see Model 2).
So the path $(P_t)_{t\geq 0}$ is initially written in the form :

$$\begin{cases} dX_t = V_t^x dt \\ dY_t = V_t^y dt \\ dZ_t = f_z dt + \tau_z dB_t^3 \end{cases} \tag{2.41}$$

with $dV_t^x = -c_x V_t^x dt + \tau_x dB_t^1$, $V_0^x = v_0^x$ and $dV_t^y = -c_y V_t^y dt + \tau_y dB_t^2$, $V_0^y = v_0^y$ and $(X_0, Y_0, Z_0) = (0, 0, 0)$; τ_x, τ_y, τ_z positive, c_x, c_y positive and f_z negative.
$(B_t^i)_{i=1,2,3}$ are three independent Brownian motions.
So the processes $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ are integrated Ornstein-Uhlenbeck processes. Contrary to Tufto *et al* (1997), both Ornstein-Uhlenbeck processes (V_t^x) and (V_t^y), are not supposed strictly stationary which permits to introduce a minimal wind speed for pollen grains emission.

However, if the path $(P_t)_{t\geq 0}$ is defined by equation (2.41) with $T_F = T_h = inf\{t > 0, Z_t = h\}$, the computed individual dispersal function is written as a integral. And under this form, the resolution for parameters estimation is not easily resolvable. In particular, there are numeric issues for some integrals.

This leads to define a new individual dispersal function based on approximation of the expectation and variance functions of the processes (X_t) and (Y_t).
According to Lemma 2.1, we have

$$E(X_t) = m_t(c_x, v_0^x) = v_0^x \frac{1 - e^{-c_x t}}{c_x}$$

$$Var(X_t) = \sigma_t^2(c_x, \tau_x) = \frac{\tau_x^2}{c_x^2}\left(t + \frac{4e^{-c_x t} - 3 - e^{-2c_x t}}{2c_x}\right)$$

And it is the same for Y_t with parameters (c_y, τ_y, v_0^y).

By analogy with the previous models GTM, NIG and GHD (cf section a)) it is assumed that the expectation is approximated by $\bar{m}_t(v_0) = v_0 t$.
By applying a limited development of order 3 on the variance function, the variance function is approximated by $\bar{\sigma_t}^2(\tau) = \tau^2 t^3$. So in this case the function is proportional to t^3 whereas for previous models the term is proportional to t. Also parameter c no appears in this model.

Thus we obtain a new model, called Model 5 and a new individual dispersal function :

Proposition 2.2 *Assuming that the path $(P_t)_{t\geq 0}$ is defined by*

(i) *X_t, a Gaussian distribution with mean $\bar{m}_t(v_0^x)$ and variance $\bar{\sigma_t}^2(\tau_x)$, with $v_0^x \in \mathbb{R}$ and $\tau_x > 0$.*

(ii) *Y_t, a Gaussian distribution with mean $\bar{m}_t(v_0^y)$ and variance $\bar{\sigma_t}^2(\tau_y)$, with $v_0^y \in \mathbb{R}$ and $\tau_y > 0$.*

(iii) *$Z_t = f_z t + \tau_z B_t$ with $f_z < 0\,,\ \tau_z > 0$.*

(iv) *X, Y and Z are independent.*

And moreover the pollination time is $T_F = T_h = inf\{t > 0, Z_t = h\}$.
Then the process (X_{T_F}, Y_{T_F}) admits the following density

$$\tilde{\gamma}(\theta; x, y) = \tilde{C}(\theta) \int_0^{+\infty} \frac{\exp\left(-\frac{\lambda_z}{u} - u\right) \exp\left(-\frac{(x - w_x u)^2}{2a_x^2 u^3} - \frac{(y - w_y u)^2}{2a_y^2 u^3}\right)}{2\pi a_x a_y u^{9/2}}\, du \qquad (2.42)$$

with $\theta = (\lambda_z, a_x, a_y, w_x, w_y)$ and $\lambda_z = \rho\eta$, $a_x^2 = \frac{\tau_x^2}{\rho^3}$, $a_y^2 = \frac{\tau_y^2}{\rho^3}$, $w_x = \frac{v_0^x}{\rho}$, $w_y = \frac{v_0^y}{\rho}$.
Moreover θ lies in $\Theta = (\mathbb{R}^+)^3 \times \mathbb{R}^2$ and $\tilde{C}(\theta) = \frac{\sqrt{\lambda_z} e^{2\sqrt{\lambda_z}}}{\sqrt{\pi}}$.

Proof : It is known that T_h is a GIG with parameters $\frac{3}{2}, \frac{f_z^2}{2\tau_z^2}$ and $\frac{h^2}{2\tau_z^2}$.
As before, under (i)-(iv)we can use the formula (2.33) and (X_{T_F}, Y_{T_F}) has a density on $\mathbb{R}^2$.

If $\omega = (\rho, \eta, \tau_x, \tau_y, f_x, f_y, v_0^x, v_0^y)$ with $\rho = \dfrac{f_z^2}{2\tau_z^2}$, $\eta = \dfrac{h^2}{2\tau_z^2}$, then

$$\begin{aligned}\tilde{\gamma}(\omega; x, y) &= \int_0^{+\infty} f_{(X_{T_h}, Y_{T_h})|T_h=t}(x, y) f_{T_h}(t) dt \\ &= C(\omega) \int_0^{+\infty} \frac{\exp\left(-\frac{\eta}{t} - \rho t\right)}{t^{3/2}} \frac{\exp\left(-\frac{(x-v_0^x t)^2}{2\tau_x^2 t^3} - \frac{(y-v_0^y t)^2}{2\tau_y^2 t^3}\right)}{2\pi\tau_x\tau_y t^3} dt\end{aligned}$$

with $C(\omega) = \dfrac{\sqrt{\eta} e^{2\sqrt{\eta\rho}}}{\sqrt{\pi}}$.

Then the result is obtained by carrying out a variable change $u = \rho t$ in the integral, with notations introduced in the proposition.

For computation of the backward individual function $\mu(x, y)$, the continuous shape is used. This leads to computation of a single integral.

After modeling different parametric individual dispersal functions, it is necessary to estimate parameters and then to compare the results to choose the most fitted model compared to the observed data.

2.11.5 Statistical analysis

a) Statistical method

Let us recall the considered statistical model :

$$N_k = n_t \mu(\theta; x_k, y_k) + \varepsilon_k \text{ with } E(\varepsilon_k) = 0 \text{ and } Var(\varepsilon_k) = \sigma^2 n_t v(\theta, b; (x_k, y_k))$$

where the $(\varepsilon_k)_k$ are assumed independent, n_t represents the total number of grains on an ear (and is taken constant) and $\sigma_k^2 > 0$.

Let the dispersion parameter $\sigma^2 = (1 + d(n_t - 1))$. σ^2 must be strictly positive.
The parameter d represents the correlation between the genotypes of two sampled descendants on a same ear (Collett 1991). If d is positive, it is called an overdispersion parameter, else it is called an underdispersion parameter. In case of overdispersion ($\delta > 0$), $Var(\varepsilon_k) > n_t v(\theta, b; (x_k, y_k))$. Hence positive correlation amongst the observations leads to greater variation in the numbers of successes (blue grains) that would be expected if they were independent. Moreover the effects of correlation between binary responses and variation between the response probabilities can not be distinguished.

A quasi-likelihood method is used to estimate parameters of the different families of individual dispersion functions $\{\gamma(\theta; x, y), \theta \in \Theta \; and \; (x, y) \in \mathbb{R}^2\}$ (Wedderburn 1974, Huet *et al* 1996).

We note $g(\theta, b, n_t; x_k, y_k) = \sigma^2 n_t v(\theta, b; (x_k, y_k))$.
Assuming that n is the number of observations, p the dimension of θ parameters vector and that the parameter b is one dimension, then quasi-likelihood equations

are given by, for $i = 1, ..., p$:

$$U_i(\theta, b) = \sum_{k=1}^{n} \frac{\partial \mu}{\partial \theta_i}(\theta; x_k, y_k) \frac{N_k - n_t \mu(\theta; x_k, y_k)}{g(\theta, b, n_t; x_k, y_k)}$$

and

$$U_{p+1}(\theta, b) = \sum_{k=1}^{n} \frac{\partial g}{\partial b}(\theta, b, n_t; (x_k, y_k)) \frac{(N_k - n_t \mu(\theta; x_k, y_k))^2 - g(\theta, b, n_t; (x_k, y_k))}{g^2(\theta, b, n_v; (x_k, y_k))} \tag{2.43}$$

The quasi-likelihood estimator (θ, b) , called $(\hat{\theta}, \hat{b})$, is then defined by $U_i(\hat{\theta}, \hat{b}) = 0$ for all $i = 1, ..., p$ and $U_{p+1}(\hat{\theta}, \hat{b}) = 0$.

The σ^2 parameter is estimated by the residual variance :

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} \frac{(N_i - n_t \mu(\theta; x_i, y_i))^2}{g(\theta, b, n_t; (x_k, y_k))} \tag{2.44}$$

The quasi-likelihood estimator $(\hat{\theta}, \hat{b})$ is consistent and asymptotically converges in distribution to a centered Gaussian distribution (Huet *et al*, 1996).

The σ^2 parameter is also consistent and asymptotically converges in distribution to a $\chi^2(n - p)$ (Collett (1991)) where n is the data number and p the θ parameter vector dimension.

Proposed variance functions

1. Binomial type function :
   As Klein *et al* (2003), in most cases it is assumed that the random variables N_k have binomial distributions of parameters $(n_k, \mu(x_k, y_k))$ (as the sum of independent variables). However it is not necessary realistic for biological and environmental reasons. In fact, the experimental conditions may vary during the experiment (variation of wind speed or of the period of ovules fertility for example). As a result the grains genotypes of a same plant are in fact correlated. Hence a dispersion parameter is introduced as in Collett 1991 or McCullagh and Nelder 1989.
   This leads to choose a binomial type variance function : $v_1(\mu) = \mu(1 - \mu)$.
2. Linear type function :
   Studying experimental data shows that the μ observed values are regrouped between 0 and 0.2. It is therefore more accurate to model the variance function on this interval, rather than on the whole interval $[0, 1]$. Moreover, it seems appropriate to use a variance function being not null at 0. Indeed, it is necessary to set weight where blue grains are observed.
   This leads to choose a linear type variance function : for $\mu \in [0, 0.5]$,
   $v_2(\mu, a) = (a + \mu)$ where $a \in \mathbb{R}^+$.

Remark : an exponential type function has also been considered, plotting the empirical variance versus expectation graphic (and having regrouped data by paquets of consequent size). However, the results are not so good, in particular because it creates visible structures on residuals graphics.

	Binomial variance		Linear variance	
Parameters	Estimation	Std error	Estimation	Std error
δ	11.04	64.85	10.18	2.55
λ_x	-0.0001	0.0602	-0.0004	0.0004
λ_y	0.0096	0.0094	0.0113	0.003
λ_z	0.0156	0.0965	0.0175	0.004
a	-	-	$1.093\ 10^{-5}$	$3.567\ 10^{-7}$
σ^2	9588		14.64	
	[9115 ; 10098]		[13.90 ; 15.37]	
d	24.39		0.0347	
$AICc$	9.172		2.6876	

TAB. 2.12 – Parameters estimations for model 1 : GTM

b) Results

Results of parameters estimations, for models 1 to 5, are given in the tables hereafter (2.12, 2.13, 2.14, 2.15 and 2.16).
For each proposed variance function type parameters estimations and associated standard errors are computed. Moreover the parameter d corresponding to the correlation and the AICc are displayed.

The estimated value of α for the third model (GHD) is 1.40 for the binomial type variance and 1.41 for the linear type variance. As a reminder first model (GTM) and second model (NIG) are both particular cases of the third model (GHD) with parameter α equal to 1/2 and 3/2 respectively. A quasi-likelihood ratio test (in the case of a heteroscedastic non linear model, Huet *et al* (1996)) can be done on the parameter α : the null hypothesis H_0 "$\alpha = 3/2$" against the alternative H_1 "$\alpha \neq 3/2$". The H_0 hypothesis is rejected at asymptotic 5% level for the binomial type variance but is accepted for the linear type variance.

The study of the standard errors for different models shows that the standard errors are significant and better with the linear type variance compared to the binomial type variance. In fact the standard errors for the binomial type variance are not always significant : for the first model for example, $\hat{\delta} = 11.04$ with a standard error $\hat{e} = 64.85$) ; and $\hat{\lambda_z} = 0.0156$ with a standard error $\hat{e} = 0.0965$.
On the other hand, for the fifth model, results are similar for the both types of variances.

The figure (2.27) (a) represents the curves of the individual dispersal functions in the wind direction for the first four models and for the linear type variance. The GTM curve decreases very quickly compared to the others. Moreover the fifth model function decreases quickly at the beginning, and after for large distance has similar values as the GHD function.
The figure (2.27) (b) represents the curves of the individual dispersal functions following the wind axis for models NIG and GHD and for the linear type variance.

	Binomial variance		Linear variance	
Parameters	Estimation	Std error	Estimation	Std error
δ	0.5176	0.0837	0.5177	0.0225
λ_x	- 0.0056	0.0509	-0.0096	0.0065
λ_y	0.1808	0.0244	0.1914	0.0143
λ_z	0.0561	0.0835	0.0669	0.0204
a	-	-	$1.175\ 10^{-5}$	$3.589\ 10^{-7}$
σ^2	145.9 [138.7 ; 156.7]		9.261 [8.801 ; 9.755]	
d	0.3687		0.0210	
$AICc$	4.986		2.2299	

TAB. 2.13 – Parameters estimations for model 2 : NIG

	Binomial variance		Linear variance	
Parameters	Estimation	Standard error	Estimation	Standard error
δ	0.5985	0.2334	0.5850	0.0571
λ_x	-0.0046	0.0217	-0.0082	0.0057
λ_y	0.1555	0.0746	0.1683	0.0205
λ_z	0.0799	0.0573	0.0853	0.0151
α	1.3986	0.1804	1.4133	0.0461
a	-	-	$1.166\ 10^{-5}$	$3.534\ 10^{-7}$
σ^2	155.6 [147.9 ; 163.8]		9.263 [8.808 ; 9.752]	
d	0.3934		0.0210	
$AICc$	5.051		2.2308	

TAB. 2.14 – Parameters estimations for model 3 : GHD

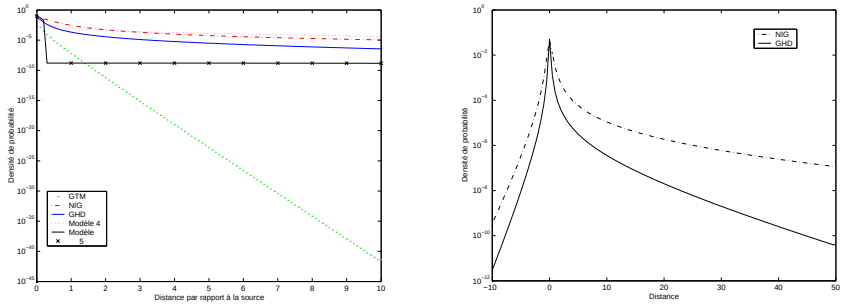
	Binomial variance		Linear variance	
Parameters	Estimation	Standard error	Estimation	Standard error
δ_x	-0.1822	0.3221	-0.1903	0.1010
δ_y	1.4359	0.4796	1.3340	0.1620
λ	0.1571	0.0179	0.1820	0.0078
b_z	0.4683	0.0532	0.4417	0.0163
c_z	0.1208	0.0647	0.1299	0.0216
a	-	-	$8.948\ 10^{-6}$	$4.803\ 10^{-8}$
σ^2	153.6 [146.03 ; 161.76]		18.81 [17.88 ;19.82]	
d	0.3883		0.0453	

TAB. 2.15 – Parameters estimations for model 4

Parameters	Binomial variance		Linear variance	
	Estimation	Standard error	Estimation	Standard error
λ_z	0.3034	0.0470	0.3058	0.0480
a_x	7.514	0.9423	7.498	0.9423
a_y	10.13	1.233	10.07	1.234
w_x	-0.7779	0.1826	-0.8749	0.1910
w_y	4.873	0.5155	4.969	0.5173
a	-	-	$1.006\ 10^{-5}$	$2.630\ 10^{-7}$
σ^2	7.674		7.089	
	[7.296 ; 8.082]		[6.7406 ; 7.466]	
d	0.0170		0.0155	

TAB. 2.16 – Parameters estimations for model 5

It can be remarked that these functions are not isotropic as suggested by the corn dispersion type.



(a) (b)

FIG. 2.27 – Individual dispersal functions in the wind direction for the fourth models (a) and individual dispersal functions following the wind axis for models NIG and GHD (b), in the case of the linear type variance.

These results lead to select the linear type variance model. The analysis and validation of these results in the below section will confirm this choice. In fact, the aim is now to choose the most fitted model according to the experiment data. For this several tools are used.

Remark : Because of numerical issues, the functions $\frac{H(u)}{g(u)}$ and $\frac{H'(u)}{g'(u)}$ have been approximated by -1 in the density function p_{θ_z} defined in (2.39) for Model 4.

2.11.6 Discussion

Standardized residuals

The study of standardized residuals permits to select the most fitted model according

to the experiment data. In fact a model is suitable if the majority of residuals lie between -2 and +2 and if the graphic does not have a particular structure. However, here there is a visible structure due to the nature of the data (in particularly because of a dispersion based on a dominant wind direction). The graphics of standardized residuals on the field are given hereafter. The middle square of blue grains corn is represented in white. On the graph of positive values, negative values are in white, and vice-versa on the other graph.

Figures (2.28), (2.29), (2.30), show that :
Regarding the binomial type variance function (Fig (2.28)) : the majority of residuals lie between -1 and 1 for 1 models 1 to 4. For the fifth model, they are bigger and in particularly there are strong positive residuals higher than two spread everywhere on the right side of the field (so in wind direction).
For the linear type variance function (Fig (2.30)), generally graphics are better than for the binomial type variance. The GTM model graphics present strong residuals on the right border of the field that are not found on the NIG model graphics. For the fourth model, strong residuals on the right border of the field appear. For the fifth model, the graphics are similar to those with binomial type variance.
Hence, this graphical interpretation leads to select the NIG model with a linear type variance function.

Remark : A selection criterion of Akaïke type (Hurvich and Tsai, 1995) can be used to choose the most fitted model for each variance function type. It is defined by $AICc = \log(\sigma^2)+1+\frac{2(p+1)}{n}$ where p is the parameter dimension and n the number of observations. Thus, having estimated parameters with a discrete sum for the global dispersal function μ (2.32), models GTM,NIG and GHD can be compared (see last line of tables of section 2.11.5). Using this test, the NIG model with a linear type variance, is selected again but the AICc value of GHD model is almost similar.
Moreover, the obtained values for σ^2 and d show that d is positive. Hence it is an overdispersion case and the variables N_k have correlated positively.

Comparison to biological and physical parameters

The estimated parameters are compared with biological parameters (female flower height h, settling velocity f_z) and physical parameters based on meteorological data (wind mean intensity and direction f_x, f_y, turbulence parameters τ, τ_z) used in the description of different models. This is also interesting in order to choose the most realistic model. The results are in Table (2.17).

We find again that, for the first three models, the most satisfying results are found using the NIG model with a linear type variance : f_x, f_y and τ_z are estimated in a satisfactory way, but τ is over estimated. Moreover these results are correct contrary to those of Klein *et al* (2003) who used a binomial variance without a dispersion parameter. (They had to multiply per two the meteorological data to obtain correct results.)

For Model 4, the estimated physical parameters are in Table (2.18). The introduced parameter β is equal to 0.3456.

For the fifth model, the estimated physical parameters are in Table (2.19). The

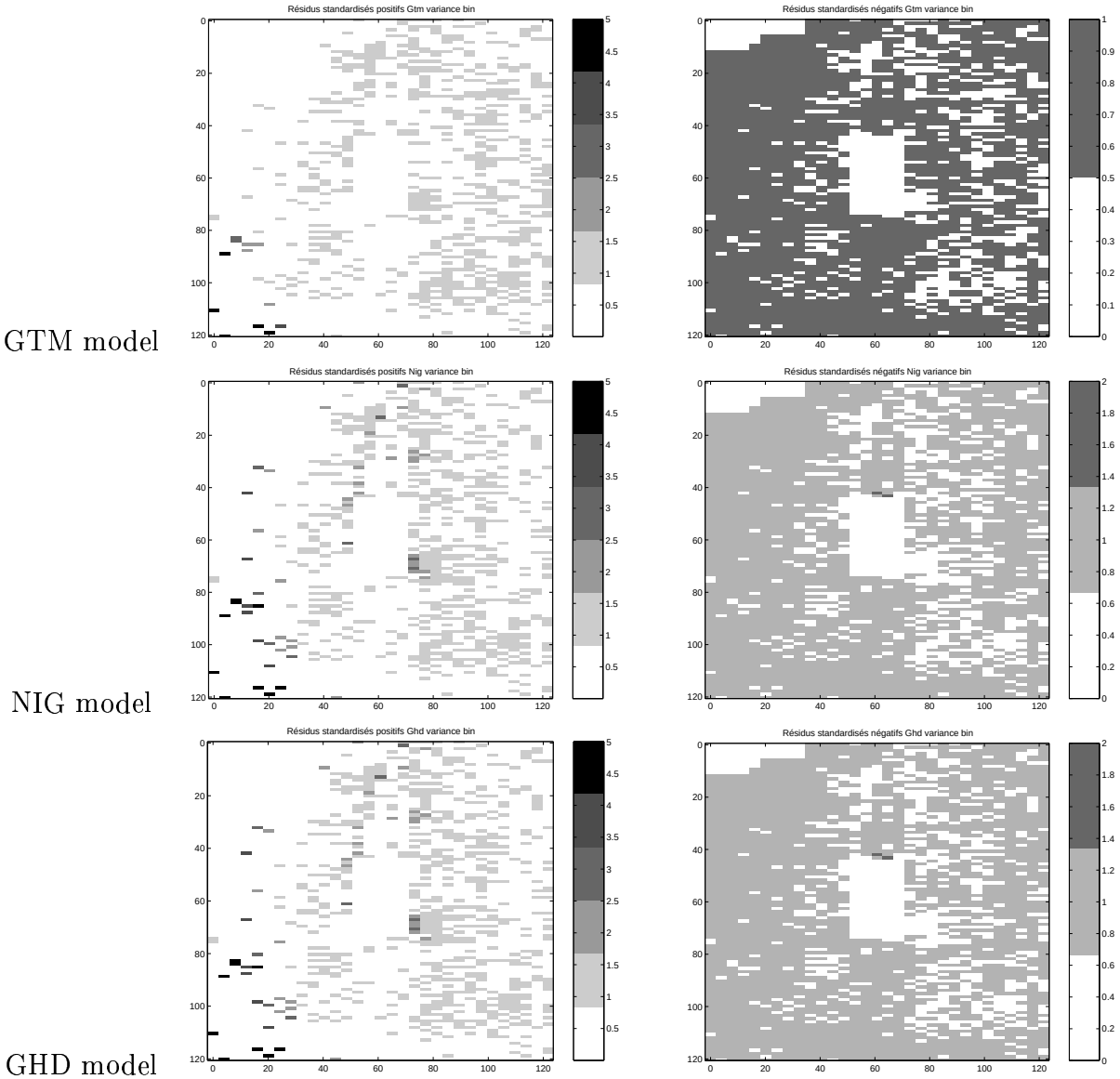


FIG. 2.28 – Positive standardized residuals (a) and negative standardized residuals (b) for the first four models in the case of the binomial type variance.

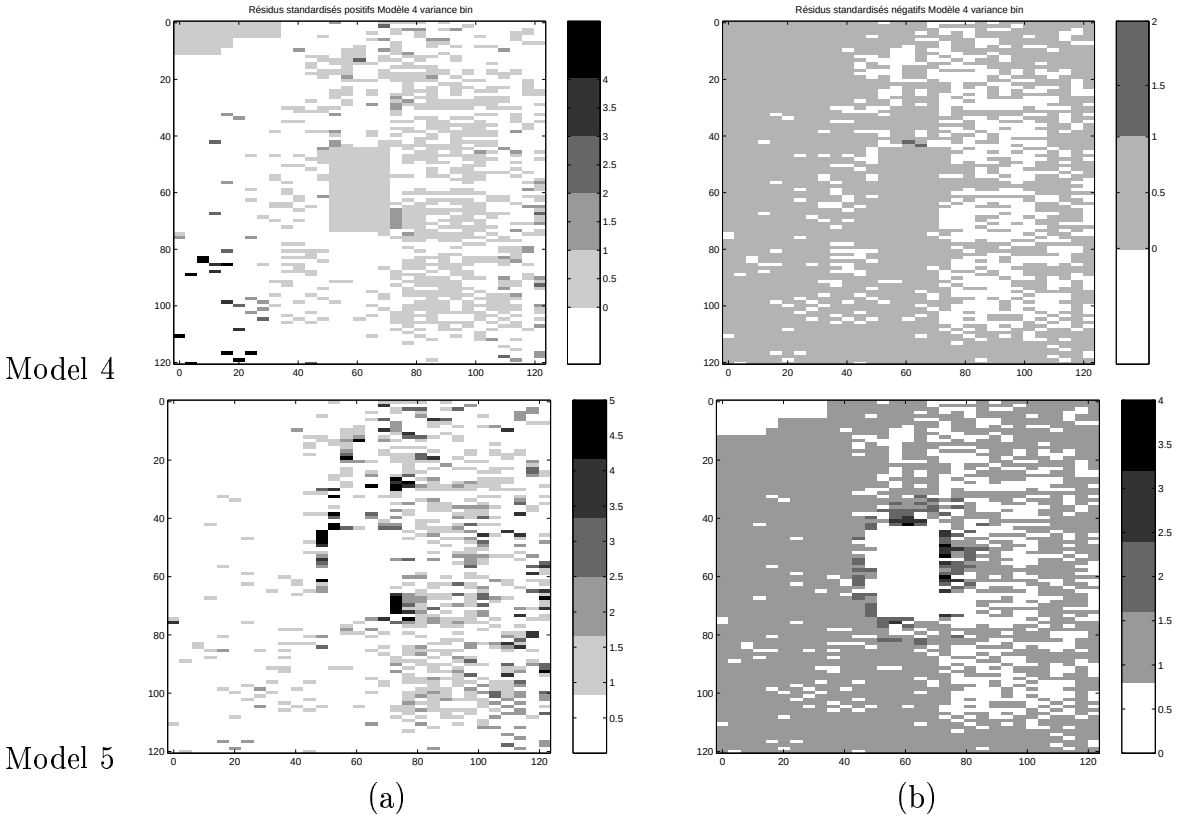


FIG. 2.29 – Positive standardized residuals (a) and negative standardized residuals (b) for models 4 et 5 in the case of the binomial type variance.

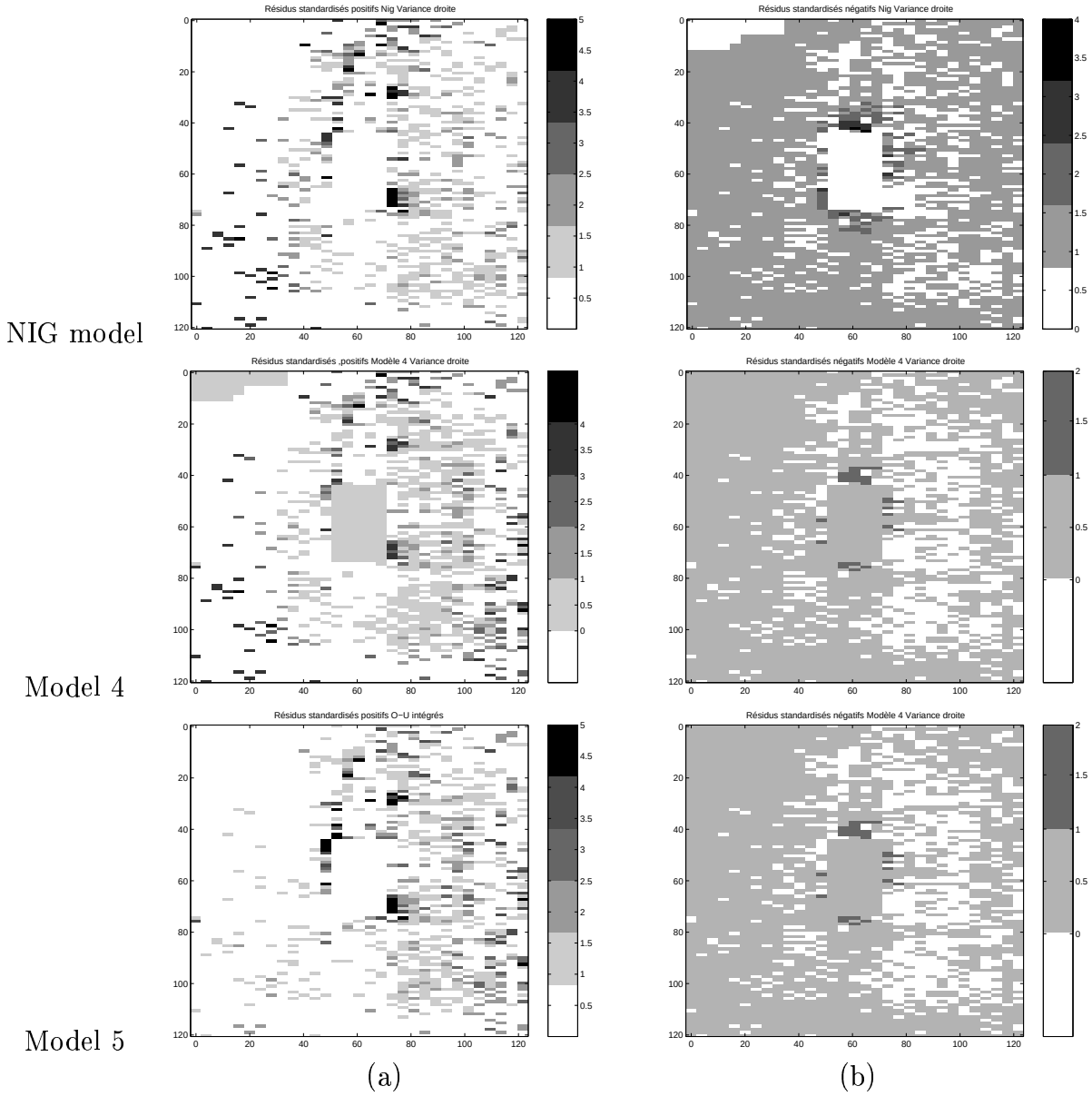


FIG. 2.30 – Positive standardized residuals (a) and negative standardized residuals (b) for the first four models in the case of the linear type variance.

minimal wind speed v_0 is thus equal to 0.6 meters per second.

This study consolidates the choice of the NIG individual dispersal function with a linear type variance.

Parameters	Min	Mean	Max	NIG Klein *et al*	NIG Bin	NIG Lin	GHD Lin
Horizontal drift : f_x $(m.s^{-1})$	-	-0.056	-	-0.074	-0.042	-0.061	-0.0362
f_y $(m.s^{-1})$	-	0.998	-	1.74	1.37	1.22	0.742
Vertical variance, τ_z $(m.s^{-1})$	0.35	1.175	2	2.37	1.65	1.51	1.33
Horizontal variances $\tau_x = \tau_y$ $(m.s^{-1})$	0.65	1.325	2	5.70	3.83	3.51	2.75
Vertical drift : f_z $(m.s^{-1})$	-	0.183	-	-	-	-	-
Height difference : h (m)	-	0.831	-	-	-	-	-
Wind velocity $(m.s^{-1})$	-	1	-	-	-	-	-

TAB. 2.17 – Comparison of estimated parameters with biological and physical parameters

Parameters $(m.s^{-1})$	f_x $(m.s^{-1})$	f_y $(m.s^{-1})$	τ_z $(m.s^{-1})$	$\tau_x = \tau_y$ $(m.s^{-1})$	β
Estimations	-0.1542	1.0807	0.1047	0.9744	0.3456

TAB. 2.18 – Estimated physical parameters for model 4

Parameters $(m.s^{-1})$	τ_z	τ_x	τ_y	v_0^x	v_0^y	Norme de v_0
Estimations	0.371	0.318	0.427	0.106	0.604	0.613

TAB. 2.19 – Estimated physical parameters for model 5

Conclusions

In these models the existing relations between physical parameters and models parameters permit to make prediction. This is interesting because for a given field and a given wind, it is possible to compute a pollination rate.

The next step regarding this work is to try to apply these results to the heterogeneous environment, i.e. when two corn fields are separated by another culture or a nude ground.

REFERENCES

Abramowitz, M. et Stegun, A.I. editors (1972). *Handbook of Mathematical Functions : with formulas, graphs and mathematical tables.* Dover Books on Advanced Mathematics. Dover Publications.

Barndorff-Nielsen, O.E. (1997). Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling. *Scandinavian Journal of Statistics* **24**, 1-13.

Bateman, A.J. (1947). Contamination of seed crops. II, Wind pollination. *Heredity* **1**, 235-246.

Collett, D. (1991). *Modelling binary data.* Chapman and hall, London.

Durbin, J. (1992) The first-passage density of the brownian motion process to a curved boundary. *J. Appl. Prob.* **29**, 291-304.

Huet S., Bouvier A., Gruet M.A. and Jolivet E. (1996). *Statistical tools for nonlinear regression.* Springer-Verlag, New-York, USA.

Hurvich, C.M. and Tsai, C.L. (1995). Model selection for extended quasi-likelihood in small samples. *Biometrics* **51**, 1077-1084.

Klein, E. (2000) *Estimation de la fonction de dispersion du pollen. Application à la dissémination de transgènes dans l'environnement.* Thèse, Université Paris XI, Orsay.

Klein, E.K., Lavigne, C., Foueillassar, X., Gouyon, P.H., Laredo, C. (2003) Corn pollen dispersal : quasi-mechanistic models and field experiments. *Ecological Monographs* **73**, 131-150.

Lavigne, C., Klein, E.K., Vallée, P., Pierre, J., Godelle, B. et Renard, M. (1998). A pollen-dispersal experiment with transgenic oilseed rape. Estimation of the average pollen dispersal of an individual plant within a field. *Theoretical and Apllied Genetics* **96**, 886-896.

McCartney, H.A. et Fitt, B.D.L. (1998). Dispersal of foliar fungal plant pathogens : mechanisms, gradients and spatial patterns. Pages 138-160 dans *The epidemiology of plant diseases.* Kluwer, Dordrecht, The Netherlands.

McCullagh P. et Nelder J.A. (1989). *Generalized Linear Models.* 2nd Edition Chapman and Hall, London.

Morris, W. F., Kareiva, P.M. et Raymer P.L. (1994). Do barren zones and pollen traps reduce gene escape from transgenic crops ? *Ecological Applications* **4**, 157-165.

Nurminiemi M., Tufto J., Nilsson O., Rognli O.A. (1998). Spatial models of pollen dispersal in the forage grass meadow fescue. *Evolutionary Ecology* **12**, 487-502.

Poilleux-Milhem, H. (2002) *1) Test de validation adaptatif dans un modèle de régression. 2) Modélisation et estimation de l'effet d'une discontinuité du couvert végétal sur la dispersion du pollen de colza.* Thèse, Université Paris XI, Orsay.

Portnoy, S. et Willson, M.F. (1993). Seed dispersal curves : behaviour of the tail of the distribution. *Evolutionary Ecology* **7**, 25-44.

Protter, P. (1992). *Stochastic Integration and Differential Equations. Applications of Mathematics*. New-York : Springer.

Rogers, L.C.G. et Williams, D. (1994). *Diffusions, Markov processes and martingales, Volume 2 Itô Calculus*. Cambridge University Press, Second edition.

Tufto, J., Engen, S., Hindar, K. (1997). Stochastic Dispersal Processes in Plant Populations. *Theoretical Population Biology* **52**, 16-26.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

Chapitre 3

Modélisation et estimation de la dispersion du flux de pollen en milieu hétérogène

Sommaire

3.1 Introduction

Dans le chapitre précédent, nous nous sommes intéressés à la modélisation de la dispersion du flux de pollen du maïs en milieu homogène à l'aide de modèles dits "mécanistes". Ici on s'intéresse à la modélisation de la trajectoire et de la fonction de dispersion individuelle d'un grain de pollen de maïs en milieu hétérogène, c'est-à-dire lorsqu'il y a une discontinuité du couvert végétal (sol nu ou autre culture). Plus précisément, on considèrera le cas où un champ de maïs est entouré d'une certaine culture puis à nouveau d'un deuxième champ de maïs.

Ce problème est important pour la gestion de l'espace dans un paysage agricole. En effet, à l'heure actuelle, la séparation entre les différents champs est le principal moyen employé pour minimiser la pollution génétique d'un champ à un autre. Mais il est rapidement apparu que l'isolation totale des cultures n'était pas réalisable.

En effet, il existe deux modes de dispersion du pollen par le vent : une grande partie du pollen reste à petite altitude et a une durée de vie de une à deux heures. Une petite partie est emportée par des vents ascensionnels à haute altitude et vit plus longtemps (les basses températures et le taux d'humidité le rendent plus viable). Ainsi, pour le maïs, des études suggèrent qu'il existe du pollen viable susceptible d'être transporté à des distances de plusieurs kilomètres (Loubet *et al*, 2004). La question de savoir si ce pollen peut effectivement polliniser est à ce jour encore ouverte.

Les modèles présentés dans la suite considèrent uniquement le premier mode de dispersion. La difficulté essentielle est que nous ne pouvons plus considérer que toutes les plantes dispersent leur pollen suivant la même fonction de dispersion individuelle. En effet, pour chaque plante du champ, la discontinuité se trouve à une distance différente et il faut donc tenir compte de la position initiale du grain de pollen dès lors que l'on considère que la discontinuité a un effet sur la fonction de dispersion individuelle.

Ainsi, pour commencer, on utilisera les résultats obtenus au chapitre précédent, en particulier le fait que la trajectoire d'un grain de pollen la plus adaptée aux données est modélisée par trois mouvements browniens avec drift indépendants pour les trois coordonnées. La fonction de dispersion individuelle associée à cette modélisation est alors une NIG (Normal Inverse Gaussian, Barndorff-Nielsen, 1997).

A partir de ce résultat, une nouvelle modélisation de la trajectoire du grain de pollen est proposée, basée sur le fait qu'une discontinuité dans le couvert végétal a deux effets principaux : c'est une zone où il n'y a pas d'émission de pollen ni de pollinisation et le vecteur de dispersion (vent) peut varier. Cependant cette modélisation ne permet pas d'exploiter numériquement la fonction de dispersion individuelle obtenue (car composée de plusieurs intégrales doubles). Cela a conduit à proposer d'autres modélisations de la discontinuité du couvert végétal, partant de la fonction (paramétrique) de dispersion individuelle NIG.

Des paramètres de translation, liés à la discontinuité, ont alors été introduit. Le

maïs étant une espèce anémophile, ces paramètres sont choisis de façon à prendre en compte le fait que la dispersion du pollen n' est pas isotrope. De tels paramètres avaient été utilisés par Poilleux-Milhem (2002), pour l'étude de l'effet d'une discontinuité sur la dispersion pour le colza, qui par contre est une espèce entomophile.
Pour finir, ces modélisations ont été confrontées avec deux expériences réalisées par l'AGPM où la discontinuité du couvert est soit du trèfle (considéré comme du sol nu) soit du tournesol (d'une hauteur de 2 mètres environ). Un modèle de régression a été considéré pour les observations qui représentent le nombre de grains bleus par épi sur le champ étudié. Les paramètres des différents modèles ont alors été estimés à l'aide d'une méthode de quasi-vraisemblance (introduite par Wedderburn, 1974). Pour finir, les différents modèles sont comparés à l'aide de méthodes graphiques (études des résidus réduits).

3.2 Description des expériences

Deux expériences ont eu lieu dans la région de Montargis en 1999 et 2000 sur deux champs d'environ 200 $m \times 160$ m.
Pour chacun des champs, une parcelle de 20×20 m de maïs homozygote pour la coloration du grain en bleu a été semée. On rappelle que ce marqueur se comporte comme un allèle dominant et s'exprime sur le grain de maïs après fécondation (cf chapitre précédent). Autour des parcelles de maïs à grains bleus, une autre culture a été plantée d'une largeur d'environ 50 m.

Pour la première expérience, il s'agit de trèfle et pour la seconde, il s'agit de tournesol. Dans le reste du champ, du maïs à grains jaunes (homozygote pour l'absence du marqueur) de la même variété a été semé. Il est à noter que le trèfle est une culture très basse qui est équivalente à un sol nu. En revanche, le tournesol est une plante d'une hauteur comprise entre $1.5m$ et $2m$. Il a donc une hauteur de l'ordre de celle des plants de maïs.

On a mesuré, dans le champ de maïs jaune, le nombre de grains bleus sur des épis échantillonnés sur un maillage régulier du champ de pas 2×1.6 m. Pour le premier champ, 8195 épis ont été échantillonnés et pour le second champ 5351 épis. De plus, pour avoir une image plus précise de la dispersion du pollen dans les premiers mètres d'un champ, un second échantillonnage avec un pas de 0.45×1.6 m a été réalisé dans la bordure face à la parcelle de maïs bleu et dans la direction du vent. (pour le premier champ, 990 épis dans les 5 premiers mètres et pour le second champ 1216 épis dans les 7 premiers mètres)

Ci-dessous, sont représentées les proportions observées de grains bleus sur chaque épi échantillonné. La parcelle de maïs bleu est représentée en bleu au centre et la discontinuité (trèfle ou tournesol) est représentée en blanc.

Remarques :
Le début de l'étape de pollinisation a lieu quand le pollen est libéré de la panicule mâle. Ensuite a lieu une phase de transport puis de dépôt (fécondation). Un grain de pollen est viable en moyenne une heure (au maximum une heure et demie). Cela

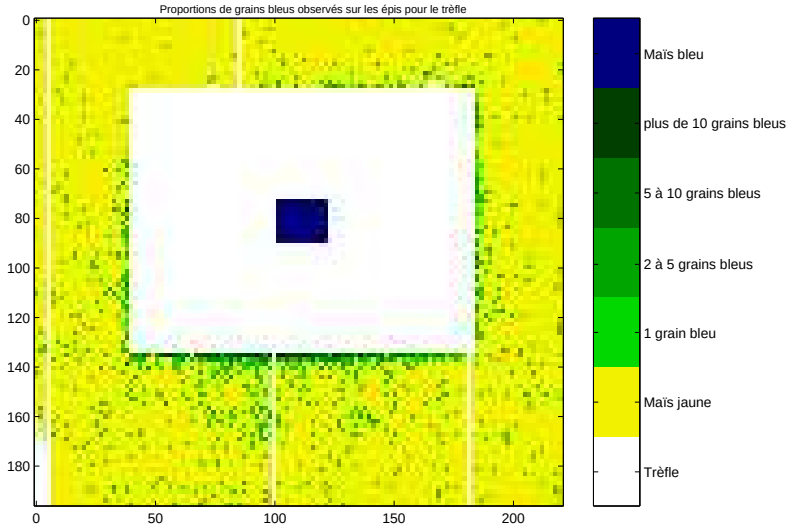


FIG. 3.1 – Proportions observées des grains bleus sur les épis échantillonnés pour le premier champ avec du trèfle.

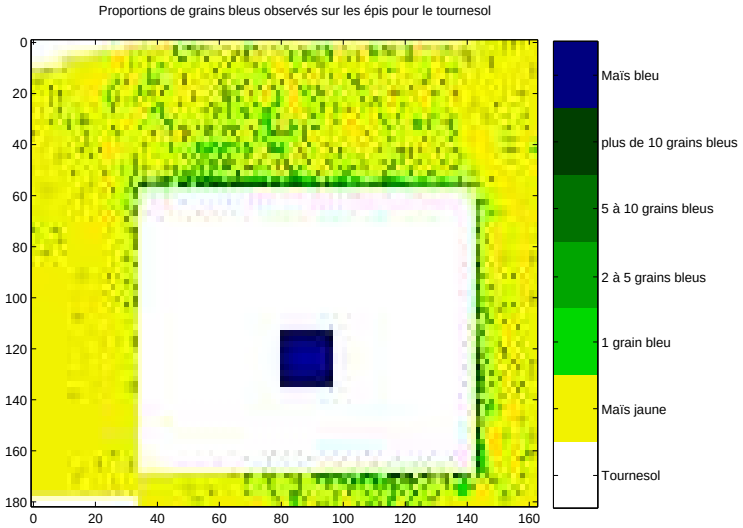


FIG. 3.2 – Proportions observées des grains bleus sur les épis échantillonnés pour le second champ avec du tournesol.

dépend des conditions climatiques : température, humidité de l'air par exemple. Lorsque l'on observe sur les épis le nombre de grains bleus, cela donne une image de toutes les pollinisations ayant eu lieu pendant la période de floraison, qui s'étale sur deux semaines environ.

Au cours des deux expériences, les données météorologiques ont indiqué qu'il y avait eu deux directions prédominantes pour le vent (un vent d'est/Nord-Est (vers le bas du champ pour le trèfle et vers le haut du champ pour le tournesol) de $1.3m.s^{-1}$ en moyenne la première semaine de floraison et un vent de Sud/Sud-Ouest (vers la gauche pour le trèfle et vers le bas droit du champ pour le tournesol) de $0.5m.s^{-1}$ en moyenne la deuxième semaine). L'observation des données sur le champ, en particulier pour le tournesol, confirme les deux directions principales de vent. Cela amène à envisager de prendre en compte le changement de direction de vent pendant cette période lors de l'étude de la dispersion du flux de pollen.

D'autre part, on constate que dans les bordures de la discontinuité, c'est-à-dire à l'entrée du champ de maïs à grains jaunes, on observe beaucoup de grains bleus. Et l'essentiel des grains bleus se trouve dans les premiers mètres de ce second champ. Plus exactement, dans les cinq premiers mètres de bordure dans la direction prédominante du vent face à la parcelle de maïs bleu, le champ est plus pollué dans le cas du trèfle par rapport au cas du tournesol (0.77 % pour le trèfle contre 0.55 % pour le tournesol).
Cependant, on observe des grains bleus en bordure du deuxième champ à plus de 100 m de la parcelle émettrice. Il semble que pour le tournesol, la pollution s'"étende" plus que pour le trèfle. Par contre la "pollution" globale des champs est similaire : 0.12 % pour le trèfle et 0.14 % pour le tournesol.

3.3 Modélisation de la dispersion du pollen

On a vu au chapitre précédent que deux fonctions de dispersion sont utilisées pour étudier la dispersion du flux de pollen : la fonction de dispersion individuelle et la fonction de dispersion globale.
Dans la suite, on notera $\gamma_{(x_0,y_0)}$ la fonction de dispersion efficace individuelle du pollen pour une plante émettrice située en (x_0, y_0). $\gamma_{(x_0,y_0)}(x, y)dxdy$ représente donc la probabilité qu'un grain de pollen émis en (x_0, y_0) tombe et féconde une plante dans le rectangle $((x, y), (x + dx, y + dy))$.

Lors de l'émission des grains de pollen, se crée un nuage pollinique au dessus de chaque plante à féconder. De plus, le pollen est supposé surabondant, c'est-à-dire que toutes les fleurs de tous les épis sont fécondées. Les pollinisations sur une plante reflètent donc exactement la constitution du nuage pollinique situé au dessus de cette plante.
La fonction de dispersion globale du pollen, représentant la composition du nuage pollinique, est la probabilité pour qu'un grain de maïs situé au point (x, y) soit bleu. Elle est notée $\mu(x, y)$.
Cette fonction est la plus naturelle et la plus utilisée par les biologistes. Cepen-

dant elle a le désavantage de dépendre des dispositifs expérimentaux utilisés : tailles respectives des champs et formes des sources. A l'inverse, la fonction de dispersion individuelle γ est intrinsèque et presque indépendante du dispositif. C'est pourquoi c'est avec cette fonction que l'on va travailler, comme l'ont fait précédemment Nurminiemi *et al* (1998), E. Klein (2000) et en particulier Klein *et al* (2003).

Dans le cas de la modélisation de la dispersion du flux de pollen en milieu homogène, on avait fait les hypothèses suivantes :
(H1) : *Toutes les plantes dispersent leur pollen suivant la même fonction de dispersion individuelle γ.*
(H2) : *Toutes les plantes produisent le même nombre de grains de pollen, et ceci quel que soit leur génotype.*
(H3) : *Il n'y a pas de différences intrinsèques entre les plantes marquées et non marquées (même viabilité, même taux de fécondation).*

L'hypothèse (H1) n'est plus valable dans le cas d'une discontinuité du couvert végétal. En effet, pour chaque plante des deux champs de maïs (bleu et jaune), la discontinuité du couvert végétal (trèfle ou tournesol) se trouve à une distance différente. La position initiale du grain de pollen rentre donc en compte lors de la modélisation de sa trajectoire et lors de l'introduction des fonctions de dispersion individuelles proposées. On fait donc désormais l'hypothèse :
(H1') : *Une plante émettrice située en (x_0, y_0) disperse son pollen suivant la fonction de dispersion individuelle $\gamma_{(x_0,y_0)}$, qui dépend de sa position.*

Sous les hypothèses (H1'), (H2) et (H3), la relation entre les fonctions μ et γ est encore valable :

$$\mu(x,y) \quad = \quad \frac{\sum_{k=1}^{S_A} \gamma_{(x_k,y_k)}(x,y)}{\sum_{k=1}^{S_A} \gamma_{(x_k,y_k)}(x,y) + \sum_{k=1}^{S_B} \gamma_{(x_k,y_k)}(x,y)} \tag{3.1}$$

où $(x_k, y_k)_{k=1,\ldots,S_A}$ représentent la localisation des plantes de maïs à grains bleus constituant la source S_A et $(x_k, y_k)_{k=1,\ldots,S_B}$ la position des plantes de maïs à grains jaunes constituant la source S_B.

Les données dont on dispose correspondent au nombre de grains bleus sur un épi localisé au point (x, y). Ce sont des observations bruitées de la fonction de dispersion globale μ. Le but est donc d'estimer les paramètres des fonctions de dispersion individuelles γ envisagées à partir de ces observations.

3.3.1 Modèle statistique

On note n_k le nombre total de grains de maïs sur un épi localisé en (x_k, y_k) et N_k le nombre de grains bleus sur cet épi.

Le modèle statistique considéré est le même que celui du chapitre précédent :

$$N_k = n_k \mu(\theta; x_k, y_k) + \varepsilon_k$$

avec $E(\varepsilon_k) = 0$ et $\mathrm{Var}(\varepsilon_k) = \sigma_k^2 n_k v(\theta, b; (x_k, y_k))$.
Et les variables aléatoires $(\varepsilon_k)_k$ sont supposées indépendantes.

Cette hypothèse d'indépendance des observations entre différents capteurs est réaliste : elle provient du fait que les grains de pollen dans le nuage pollinique sont très nombreux et non limitants. Le fait d'observer beaucoup de grains dans un capteur n'affecte donc pas le nombre de grains trouvés dans les autres capteurs.

Le paramètre σ_k^2 est appelé paramètre de dispersion et on doit avoir $\sigma_k > 0$.
Il représente la corrélation entre les génotypes de deux descendants échantillonnés sur un même capteur (Collett 1991, Huet *et al* 1996). Pour plus de détails, on se reportera au Chapitre 2, Paragraphe 2.3.2).

Enfin, on constate, d'après (3.1), que l'on a affaire à un problème de déconvolution non linéaire qui est difficile à étudier du point de vue statistique. C'est pourquoi, par la suite, on considérera des fonctions de dispersion individuelles paramétriques de la forme $\{\gamma(\theta; x, y), \theta \in \Theta \; et \; (x, y) \in \mathbb{R}^2\}$ avec Θ sous-ensemble de $\mathbb{R}^p$ $(p \geq 1)$.

3.3.2 Modélisation de la trajectoire

Lors de l'étude de la dispersion du pollen de maïs en milieu homogène à l'aide de modèles mécanistes, on avait conclu que la fonction de dispersion individuelle qui s'adaptait le mieux aux données, était une fonction de type NIG :

$$f_{NIG}(\lambda_z, \lambda_x, \lambda_y, \delta; x, y) = \frac{\delta^2 e^{\lambda_z}}{2\pi} \frac{(q(x,y)^{-1/2} + p^{1/2})}{q(x,y)} e^{-\sqrt{pq(x,y)}} e^{\delta(\lambda_x x + \lambda_y y)} \tag{3.2}$$

avec $p = \lambda_z^2 + \lambda_x^2 + \lambda_y^2$ et $q(x, y) = 1 + \delta^2(x^2 + y^2)$.

Dans ce cas, la trajectoire du grain de pollen, $P_t = (X_t, Y_t, Z_t)_{t \geq 0}$, peut être modélisée par trois mouvements browniens avec drift indépendants :

$$\begin{cases} dX_t = f_x dt + \tau_x dB_t^1 \\ dY_t = f_y dt + \tau_y dB_t^2 \\ dZ_t = f_z dt + \tau_z dB_t^3 \end{cases} \tag{3.3}$$

où $f_x, f_y > 0$, $f_z < 0$ et $\tau_x, \tau_y, \tau_z > 0$.

Rappelons que les composantes horizontales du drift représentent les vitesses moyennes du vent et que la composante verticale du drift représente la vitesse de chute du grain de pollen due à la gravité. Les paramètres de variance du processus stochastique représentent les turbulences dues au vent.
On a les relations suivantes reliant les paramètres de la fonction de dispersion individuelle, f_{NIG}, et les paramètres "physiques" de la modélisation (3.3) de la trajectoire d'un grain de pollen :

$$\delta = \frac{\tau_z}{\tau |h|}, \; \lambda_x = \frac{f_x h}{\tau \tau_z}, \; \lambda_y = \frac{f_y h}{\tau \tau_z}, \; \lambda_z = \frac{f_z h}{\tau_z^2}$$

Une discontinuité dans le couvert végétal a deux effets :

– C'est une zone où il n'y a ni émission de pollen, ni pollinisation.
– Le vecteur de dispersion (vent) peut varier et donc modifier la fonction de dispersion individuelle du pollen.

Il est donc nécessaire d'apporter des modifications au modèle ci-dessus.

3.4 Modélisation stochastique de la trajectoire

On suppose que le vent a une direction prédominante lors de la période de pollinisation et que la discontinuité est constituée de plantes moins hautes que la hauteur des fleurs femelles.

Dans les champs où il y a du maïs, c'est-à-dire les domaines D_1 et D_3, la trajectoire est modélisée par trois mouvements browniens avec drift indépendants. Dans la partie cultivée avec une autre culture, domaine noté D_2, on suppose que la trajectoire est déterministe suivant la direction du vent et qu'elle descend. En effet, dans cette partie, on considère que le grain de pollen est soumis seulement à la force du vent et ne subit pas de turbulences atmosphériques dues à la végétation.

D'autre part, il existe trois possibilités pour un grain de pollen émis :

1. Il se fixe sur une fleur avant de sortir du premier champ de maïs D_1.
2. Il sort du premier champ et meurt au sol dans la discontinuité D_2 ou dans le deuxième champ D_3.
3. Il sort du premier champ, va jusqu'au deuxième champ et se fixe sur une fleur.

Dans la suite, on s'intéressera à une trajectoire en dimension 2 pour simplifier les notations et les calculs.

On suppose donc que le grain de pollen est une particule partant d'un point $(x_0, 0)$ appartenant au domaine D_1. On suppose que le vent a une direction prédominante suivant l'orientation de l'axe des abscisses. Soit $P_t = (X_t, Z_t)$ la position du grain de pollen à l'instant $t > 0$.
On définit la trajectoire du grain de pollen par :
– Dans D_1 et D_3, on suppose que :

$$\begin{cases} dX_t = f_x dt + \tau_x dB_t^1 \\ dZ_t = f_z dt + \tau_z dB_t^2 \end{cases} \tag{3.4}$$

où (B_t^1) et (B_t^2) sont deux mouvements browniens indépendants ; $f_x > 0$, $f_z < 0$ et $\tau_x, \tau_z > 0$.

– Dans D_2, on suppose donc la trajectoire déterministe :

$$\begin{cases} dX_t = \bar{f}_x dt \\ dZ_t = \bar{f}_z dt \end{cases} \tag{3.5}$$

avec $\bar{f}_x > 0$ (par hypothèse) et $\bar{f}_z < 0$.

Ci-dessous, à la Figure 3.3, une représentation de la modélisation proposée de la trajectoire d'un grain de pollen.

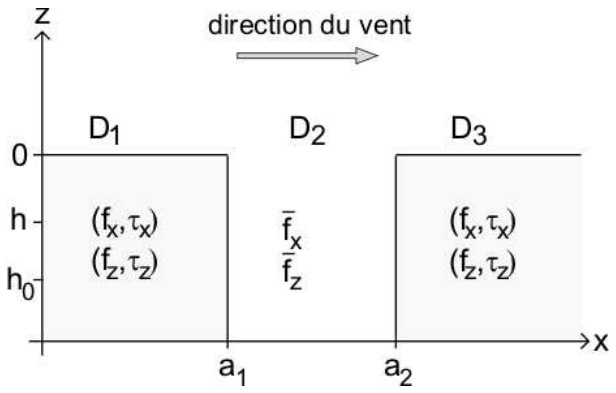


FIG. 3.3 – Représentation de la modélisation de la trajectoire d'un grain de pollen proposée en dimension 2.

En envisageant ces différents cas, cela conduit à définir le temps de fécondation T_F, instant auquel a lieu la pollinisation, de la façon suivante :

On introduit les temps d'atteinte T_h et $\nu_{a_i}(X)$ de la façon suivante :

- $T_h = \inf\{t > 0, Z_t = h\}$.
- Pour $i = 1, 2$; $\nu_{a_i}(X) = \inf\{t > 0, X_t = a_i\} = \nu_{a_i}$.
  Comme $f_x > 0$, on a $\nu_{a_1}(X) < +\infty$ *p.s.*

Le temps de fécondation T_F est alors défini par :

- Si $T_h \wedge \nu_{a_1} = T_h$, on prend $T_F = T_h$ et la pollinisation a lieu dans D_1 en (X_{T_h}, Y_{T_h}).
- Si $T_h \wedge \nu_{a_1} = \nu_{a_1}$, Trois cas sont à distinguer. On introduit une hauteur seuil h_0 en-dessous de laquelle il ne peut pas y avoir pollinisation.
  1. Si $Z_{\nu_{a_2}} < h_0$, le grain de pollen est perdu.
  2. Si $h_0 \leq Z_{\nu_{a_2}} \leq h$, il y a pollinisation sur la frontière de D_2 (en a_2).
  3. Si $Z_{\nu_{a_2}} > h$, la trajectoire continue en repartant de $(a_2, Z_{\nu_{a_2}})$ et on pose $T_F = \inf\{t > \nu_{a_2}, Z_t = h\}$.

  Dans le cas 3., pour $\nu_{a_1} \leq t \leq \nu_{a_2}$, on peut écrire d'après (3.5)

$$X_t = \bar{f}_x(t - \nu_{a_1}) + a_1 \text{ et } Z_t = \bar{f}_z(t - \nu_{a_1}) + Z_{\nu_{a_1}}.$$

  Alors, on a $\nu_{a_2} = \bar{a} + \nu_{a_1}$ où on a posé $\bar{a} = \dfrac{a_2 - a_1}{\bar{f}_x}$. On en déduit que

$$Z_{\nu_{a_2}} = \bar{f}_z + Z_{\nu_{a_1}} \tag{3.6}$$

Dans tous les cas, T_F est fini $p.s.$ puisque $f_z < 0$ et $Z_{\nu_2} > h$.

Le but est de calculer, pour un grain de pollen allant **effectivement** polliniser une fleur (donc n'étant pas perdu), la loi de la variable aléatoire X_{T_F} qui représente une fonction de dispersion efficace individuelle du pollen d'après la Propriété 2.1 du chapitre 2 (paragraphe 2.4).

Proposition 3.1 *On pose $u = x - x_0$ et $u_1 = a_1 - x_0$. La loi de la variable aléatoire X_{T_F}, pour un grain de pollen allant effectivement polliniser une fleur, est :*

$$\gamma(x, x_0, \theta) = I(u, h, \theta)\mathbf{1}_{\{x \in D_1\}} + J(u_1, \theta)\mathbf{1}_{\{x = a_2\}} + G(x, x_0, a_1, a_2, h_2, \theta)\mathbf{1}_{\{x > a_2\}} \quad (3.7)$$

avec $\theta = (f_x, f_z, \sigma_x, \sigma_z, \bar{f}_x, \bar{f}_z)$, et

1. $I(u, h, \theta) = 2C(u, h, \theta) \times \left(\dfrac{p(u, h, \theta)}{q(\theta)}\right)^{1/2} \times \mathcal{K}_{-1}\left(2\sqrt{q(\theta)p(u, h, \theta)}\right)$

   avec $p(u, h, \theta) = \dfrac{h^2}{2\tau_z^2} + \dfrac{u^2}{2\tau_x^2}$, $q(\theta) = \dfrac{f_z^2}{2\tau_z^2} + \dfrac{f_x^2}{2\tau_x^2}$ *et*

   $C(u, h, \theta) = \dfrac{|h|}{2\pi\tau_x\tau_z} \exp\left(\dfrac{f_z h}{\tau_z^2} + \dfrac{f_x u}{\tau_x^2}\right)$

2. $J(u_1, \theta) = \displaystyle\int_{h_1}^{h_2} J(z, u_1, \theta)\, dz$ *avec*

   $J(z, u_1, \theta) = \tilde{C}(u_1, \theta) \exp\left(\dfrac{f_z}{\tau_z^2} z\right) \times 2 \left(\dfrac{r(z, u_1, \theta)}{s(\theta)}\right)^{1/2} \times \mathcal{K}_{-1}\left(2\sqrt{s(\theta)r(z, u_1, \theta)}\right)$

   $r(z, u_1, \theta) = \dfrac{u_1^2}{2\tau_x^2} + \dfrac{z^2}{2\tau_z}$, $s(\theta) = \dfrac{f_x^2}{2\tau_x^2} + \dfrac{f_z^2}{2\tau_z^2}$ *et* $\tilde{C}(u_1, \theta) = \dfrac{u_1}{2\pi\tau_x\tau_z} \exp\left(\dfrac{f_x u_1}{\tau_x^2}\right)$.

3. $G(x, x_0, a_1, a_2, h_2, \theta) = \displaystyle\int_{z > h_2} I(x - a_2, h_2 - z, \theta)\, J(z; a_1 - x_0, \theta)\, dz$

Démonstration : Elle se trouve à l'annexe (3.8).

Remarque :
Le terme J est une intégrale définie et le terme G est une double intégrale avec une borne infinie. De plus, pour calculer la fonction de dispersion globale associée, il faut effectuer une somme discrète sur tous les x_0 appartenant au domaine D_1 et D_3.

Cas de la dimension 3 :
Le grain de pollen est une particule partant d'un point $(x_0, y_0, 0)$ appartenant au domaine D_1. On suppose que le vent a une direction prédominante et on choisit l'orientation des axes des abscisses et des ordonnées de telle sorte que la direction principale du vent se trouve dans le premier quadrant.
On définit

- D_1 comme le carré : $D_1 = \{(x,y) \in \mathbb{R}^2, -a_1 \leq x \leq a_1 \text{ et } -b_1 \leq y \leq b_1\}$.
- D_2 comme l'ensemble $D = \{(x,y) \in \mathbb{R}^2, -a_2 \leq x \leq a_2 \text{ et } -b_2 \leq y \leq b_2\}$ auquel on retire D_1.
- D_3 représente le grand champ auquel on retire le carré D.

On a affaire à la même problématique qu'en dimension 2. Cependant, si il n'y a pas pollinisation dans le premier champ, le grain de pollen sort de D_1 par l'un des quatre côtés du champ et de même entre dans D_3 par l'un des quatre côtés. Il faut donc tenir compte de ces différents cas dans la modélisation. Les calculs pour déterminer la loi du couple $(X_{T_F} Y_{T_F})$ s'effectuent alors sur le même principe qu'en dimension 2. L'expression de la fonction de dispersion individuelle est donc composée de la somme de plusieurs termes (correspondant aux différents cas envisagés) incluant des intégrales simples ou doubles dont les bornes peuvent être infinies.

Conclusion :
Malgré les simplifications faites sur les hypothèses de la trajectoire d'un grain de pollen, la fonction de dispersion efficace individuelle calculée est une expression compliquée et n'est donc pas numériquement exploitable pour estimer les paramètres du modèle introduit comme cela avait été fait au chapitre précédent à l'aide d'une méthode de quasi-vraisemblance. On a donc envisagé de modéliser l'effet d'une discontinuité du couvert végétal d'une autre manière.

Cependant, on peut noter que la forme de la fonction individuelle obtenue en (3.7) nous apporte des informations. Elle est constituée de la somme de trois termes : le premier est une fonction pour les points appartenant au champ bleu, le troisième est une fonction différente pour les points appartenant au champ jaune. Enfin le second terme est une contribution en bordure de la discontinuité. Cela est en accord avec les représentations graphiques des données (Figures 3.1 et 3.1) qui ont montré que à l'entrée du champ de maïs jaune, c'est-à-dire à la bordure de la discontinuité, il y avait une forte pollution. Cela nous servira pour le troisième modèle envisagé au paragraphe suivant.

3.5 Introduction de l'effet de la discontinuité dans les fonctions de dispersion individuelles

Dans cette section, des fonctions de dispersion individuelles sont proposées, basées sur les résultats obtenus au chapitre précédent dans le cas du milieu homogène .

On rappelle que le domaine D_1 représente le champ de maïs à grains bleus, le domaine D_2 est constitué d'une autre culture et D_3 représente le champ de maïs à grains jaunes.

3.5.1 Modèle A : Normale Inverse Gaussienne, NIG

Le premier modèle envisagé est simple : on suppose juste qu'il n'y a pas de pollinisation dans le domaine D_2, c'est-à-dire que la fonction de dispersion individuelle est nulle pour les points appartenant à ce domaine.

On pose

$$f_{NIG}(x,y;\theta) = \frac{\delta^2 e^{\lambda_z}}{2\pi}\frac{(q(x,y)^{-1/2}+p^{1/2})}{q(x,y)}e^{-\sqrt{pq(x,y)}}e^{\delta(\lambda_x x+\lambda_y y)}$$

avec $p=\lambda_x^2+\lambda_y^2+\lambda_z^2$, $q(x,y)=1+\delta^2(x^2+y^2)$ et $\theta=(\delta,\lambda_x,\lambda_y,\lambda_z)$.

On définit donc la fonction de dispersion individuelle pour une plante émettrice située en un point (x_0,y_0) appartenant à D_1 par

$$\gamma_{(x_0,y_0)}(x,y) = \begin{cases} f_{NIG}(x-x_0,y-y_0;\theta) & \text{si } (x,y)\in D_1 \\ 0 & \text{si } (x,y)\in D_2 \\ f_{NIG}(x-x_0,y-y_0;\theta) & \text{si } (x,y)\in D_3 \end{cases} \tag{3.8}$$

On définit de la même façon $\gamma_{(x_0,y_0)}(x,y)$ pour une plante émettrice située en un point (x_0,y_0) appartenant à D_3.

3.5.2 Modèle B : NIG "translatée"

Ici, l'idée est de supposer que la trajectoire du grain de pollen est déterministe et rectiligne dans le domaine D_2. Le grain de pollen parcourt donc une certaine distance d. Suivant la hauteur de la culture se trouvant dans la zone de transition, on peut penser que le grain de pollen va changer de vitesse (accélération ou ralentissement). C'est pourquoi, on introduit des paramètres de translation α_i et on s'intéresse au terme $\alpha_i d$ en prenant en compte le fait que la dispersion n'est pas isotrope.
(Dans le cas de l'étude de la dispersion du pollen de colza, H. Poilleux-Milhem (2002) avait également introduit un tel paramètre, fonction de la largeur de la discontinuité.)

On considère deux points $P_0=(x_0,y_0)$, position de la plante émettrice, et $P_1=(x_1,y_1)$, position de la plante réceptrice, se situant dans les champs D_1 ou D_3. On envisage deux possibilités :

- Si la droite (P_0P_1) n'intercepte pas le domaine D_2, on pose :

$$D_{x_0,x_1}=0 \text{ et } D_{y_0,y_1}=0.$$

- Si la droite (P_0P_1) intercepte le domaine D_2, ou bien les deux points sont dans D_3, ou bien un point est dans D_1 et l'autre est dans D_3. Alors, on définit D_{x_0,x_1} et D_{y_0,y_1} de la façon suivante (voir la figure 3.4 pour plus de précisions) :
  On considère le segment reliant les deux points P_0 et P_1.
  On définit D_{x_0,x_1} comme la distance entre les deux points P_0 et P_1 dans le domaine D_2 suivant l'axe des abscisses.
  De même, on définit D_{y_0,y_1} comme la distance entre les deux points P_0 et P_1 dans le domaine D_2 suivant l'axe des ordonnées.

Ainsi, on considère que, dans le domaine D_2, le grain de pollen parcourt une distance D_{x_0,x_1} suivant l'axe des x et une distance D_{y_0,y_1} suivant l'axe des y.

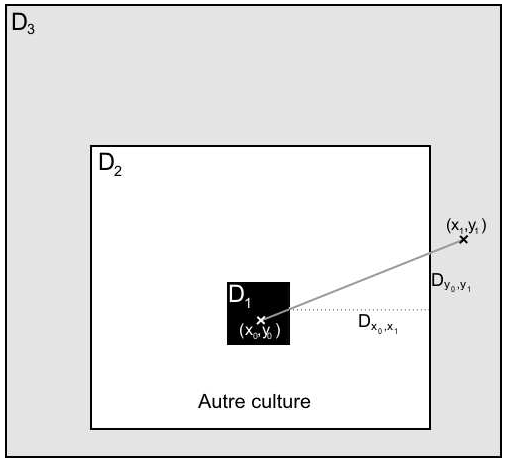


FIG. 3.4 – Modélisation de la discontinuité à l'aide de paramètres de translation

On modélise l'effet de la discontinuité du couvert végétal dans le domaine D_2 en introduisant des paramètres de translation $\alpha_{1,i}$ suivant l'axe des x et $\alpha_{2,i}$ suivant l'axe des ordonnées de la façon suivante :
Pour $(x, y) \in \mathbb{R}^2$, on considère ses coordonnées polaires associées (r, θ). Alors pour $1 \leq i \leq m_i$, on définit l'ensemble $S_i = \{(x, y) \in \mathbb{R}^2, u_i \leq \theta \leq u_{i+1}\}$. Les (u_i) sont tels que $\cup_i S_i = \mathbb{R}^2$.
Pour $(x, y) \in D_1$ ou D_3 et $(x, y) \in S_i$, on remplace donc $(x - x_0)$ par $(x - x_0 - \alpha_{1,i} D_{x_0,x})$ et $(y - y_0)$ par $(y - y_0 - \alpha_{2,i} D_{y_0,y})$.
Dans la pratique les ensembles S_i sont choisis par rapport aux directions du vent lors de l'étape de pollinisation, et sont au nombre de deux ou trois.

On peut remarquer que $\alpha_i > 0$ signifie qu'il y a un effet d'accélération sur la dispersion du pollen et inversement, si $\alpha_i < 0$, il y a un effet de ralentissement. Ci-dessous, à la figure 3.5 , la représentation de cet effet en dimension 1.

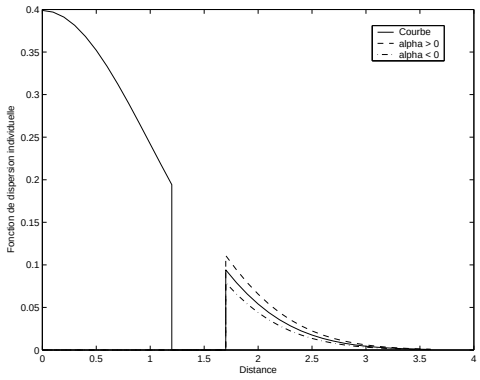


FIG. 3.5 – Effet d'un paramètre de translation sur la courbe de dispersion

Finalement, on définit la fonction de dispersion individuelle pour une plante émettrice située en (x_0, y_0) par

$$\gamma_{(x_0,y_0)}(x, y) = \begin{cases} f_{NIG}(x - x_0 - \alpha_1 D_{x_0,x}, y - y_0 - \alpha_2 D_{y_0,y}; \theta) & \text{si } (x, y) \in D_1 \\ 0 & \text{si } (x, y) \in D_2 \\ f_{NIG}(x - x_0 - \alpha_{1,i} D_{x_0,x}, y - y_0 - \alpha_{2,i} D_{y_0,y}; \theta) & \text{si } (x, y) \in D_3 \cap S_i \end{cases} \quad (3.9)$$

On note que si $P_0 = (x_0, y_0)$ et $P = (x, y)$ appartiennent à D_3 et si la droite (P_0P) n'intercepte pas le domaine D_2 alors on a $D_{x_0,x} = D_{y_0,y} = 0$.

3.5.3 Modèle C :

Le troisième modèle considéré prend en compte le fait de la forte pollinisation en bordure de la discontinuité comme on peut le voir sur les représentations graphiques

des données (Figures 3.1 et 3.2). De plus, il s'appuie sur la forme de la fonction individuelle obtenue en (3.7) qui contient un terme constant à la frontière du domaine D_2. On note $\tilde{D}_3$ l'ensemble constitué des points de D_3 situés à au plus 1 mètre de D_2.
On va donc considérer que si l'épi de maïs appartient à l'ensemble $\tilde{D}_3$ et si le grain de pollen passe dans la discontinuité, alors la fonction de dispersion individuelle est égale à la somme d'un paramètre $q \in]0,1[$ et d'un terme de la forme NIG. Sinon on prend l'expression de la fonction individuelle proposée pour le modèle B.

Ainsi, avec les notations introduites à la section précédente, pour une plante émettrice de coordonnées (x_0, y_0) se situant dans D_1, la fonction de dispersion individuelle est définie par :

$$\gamma_{(x_0,y_0)}(x,y) = \begin{cases} f_{NIG}(x-x_0, y-y_0; \theta) & \text{si } (x,y) \in D_1 \\ 0 & \text{si } (x,y) \in D_2 \\ q + f_{NIG}(x - x_0 - \alpha_{1,i} D_{x_0,x}, y - y_0 - \alpha_{2,i} D_{y_0,y}; \theta) & \text{si } (x,y) \in \tilde{D}_3 \cap S_i \\ f_{NIG}(x - x_0 - \alpha_{1,i} D_{x_0,x}, y - y_0 - \alpha_{2,i} D_{y_0,y}; \theta) & \text{si } (x,y) \in D_3 \cap S_i \end{cases} \tag{3.10}$$

3.6 Estimation des paramètres

3.6.1 Rappel de la méthode statistique

Pour les modèles étudiés, nous avons fait l'hypothèse que le nombre de grains de maïs par épis est constant, c'est-à-dire que les n_k sont constants, égaux à $n = 394$ (nombre moyen par épi).
Au paragraphe 3.3.2, lors de la modélisation de la dispersion, on a vu que le modèle statistique considéré était un modèle de régression de la forme :

$$N_k = n\mu(\theta; x_k, y_k) + \varepsilon_k$$

$$\text{avec } E(\varepsilon_k) = 0 \text{ et } Var(\varepsilon_k) = \sigma^2 g(\theta, n; (x_k, y_k))$$

où les $(\varepsilon_k)_k$ sont supposées indépendantes.
La fonction de variance utilisée est de type binomiale :

$$g(\theta, n; (x_k, y_k)) \;=\; n\mu(\theta; x_k, y_k)(1 - \mu(\theta; x_k, y_k))$$

Les paramètres ont été estimés en utilisant une méthode de quasi-vraisemblance (Wedderburn (1974)) comme au chapitre précédent.

Remarque : Au chapitre 2, lors de l'étude de la dispersion du pollen de maïs en milieu homogène, on avait conclu qu'une fonction de variance de type linéaire était la mieux adaptée aux données étudiées. Cependant, ici, avec une telle fonction, l'estimation des paramètres des différentes fonctions de dispersion proposées, conduit à de moins bons résultats. En effet, les résidus réduits obtenus sont élevés dès que l'on s'éloigne d'environ 20 mètres de la discontinuité. Ainsi, pour des distances "longues", le nombre de grains bleus sur les épis est sous-estimé. Cela peut s'expliquer par le

fait que l'on donne un poids trop important aux observations situées au bord de la discontinuité et on néglige celles non nulles qui se trouvent à plusieurs dizaines de mètres de la discontinuité.

3.6.2 Expérience 1 : trèfle

On rappelle que la zone de discontinuité constitué de trèfle correspond à une zone de sol nu.

a) Résultats

Les résultats des estimations des paramètres, pour l'expérience avec le trèfle, se trouvent dans le tableau (3.1) pour le modèle A (NIG) ; et dans le tableau (3.2) pour le modèle B (NIG "translatée") et le modèle C.

Pour les modèles B et C, on a pris deux ensembles S_1 et S_2, S_1 étant le demi-plan contenant la direction prédominante du vent lors de la première semaine de floraison et S_2 l'autre demi-plan.
De plus, aux vues des résultats obtenus pour le deuxième modèle, l'ensemble $\tilde{D}_3$ est constitué seulement de la bordure en bas.

On remarque que les valeurs des paramètres δ et λ_z varient significativement entre le modèle A, décrit en (3.5.1), et les modèles B et C,décrits en (3.5.2) et (3.5.3) (0.043, 0.24 et 0.28 pour δ ; 1.45, 0.36 et 0.27 pour λ_z).
D'autre part, les paramètres de translation introduits dans les modèles B et C sont positifs. Cela signifie que la zone de discontinuité a un effet d'accélération sur la dispersion du pollen. Il apparaît qu'il y a un plus grand phénomène d'accélération avec le modèle C.
La bordure de la discontinuité, dans la direction du vent, est donc plus polluée que dans le cas où le maïs aurait été cultivé en milieu homogène pour une même distance.
Enfin, pour le modèle C, on obtient un paramètre q valant environ 0.01.

	Modèle A : NIG	
Paramètres	Estimation	Ecart-type
δ	0.0435	0.0036
λ_x	0.5413	0.0408
λ_y	-0.1096	0.0237
λ_z	1.4542	0.2045
σ^2	63.59	
	[61.87 ;65.47]	

TAB. 3.1 – Résultats des estimations des paramètres pour le modèle A dans le cas du trèfle

	Modèle B : NIG "translatée"		Modèle C	
Paramètres	Estimation	Ecart-type	Estimation	Ecart-type
δ	0.2455	0.1537	0.2862	0.2137
λ_x	0.2578	0.0652	0.3236	0.1646
λ_y	-0.1500	0.0401	-0.0287	0.0830
λ_z	0.3627	0.1197	0.3262	0.2160
q	-	-	0.0104	$9.168\ 10^{-3}$
$\alpha_{1,1}$	0.6010	0.1297	0.7877	0.1228
$\alpha_{1,2}$	0.4635	0.0839	0.7077	0.1195
$\alpha_{2,1}$	0.6349	0.0955	0.7533	0.0600
$\alpha_{2,2}$	0.7286	0.1139	0.6835	0.1576
σ^2	35.68 [33.89 ;37.49]		96.79 [91.96 ;101.69]	

TAB. 3.2 – Résultats des estimations des paramètres pour les modèles B et C dans le cas du trèfle

b) Etude des résidus réduits

Ci-dessous, se trouvent les graphiques des résidus réduits sur le champ (figures 3.6, 3.7 et 3.8).

Les résidus réduits négatifs apportent peu de renseignements sur les différents modèles (car pour les trois modèles, ils sont compris entre -1 et 0) contrairement aux résidus réduits positifs. Rappelons que des résidus positifs correspondent à une sous estimation des données, et inversement pour les résidus négatifs.

- Pour le modèle A, NIG, on constate qu'il y a un grand nombre de forts résidus un peu partout, et en particulier en bordure de la discontinuité et à la limite du champ.
- Pour le modèle B, NIG "translatée", la plupart des forts résidus en bordure de la discontinuité ont disparu. Cependant il reste des résidus supérieurs à 2 au bord de tout le champ (pour des valeurs faibles de μ, la dispersion globale). Il s'avère donc que ce modèle sous-estime légèrement la dispersion à "longue" distance.
- Pour le modèle C, les forts résidus en bordure du bas du champ ont disparu mais il reste quelques forts résidus dans la direction opposée à celle prédominante du vent (la première semaine de floraison).

D'après ces résultats, le modèle A est exclu. Les modèles B et C sont assez semblables, avec une préférence pour le modèle B pour lequel les écarts-types des paramètres $\delta, \lambda_x, \lambda_y$ et λ_z sont plus petits que ceux du modèle C.

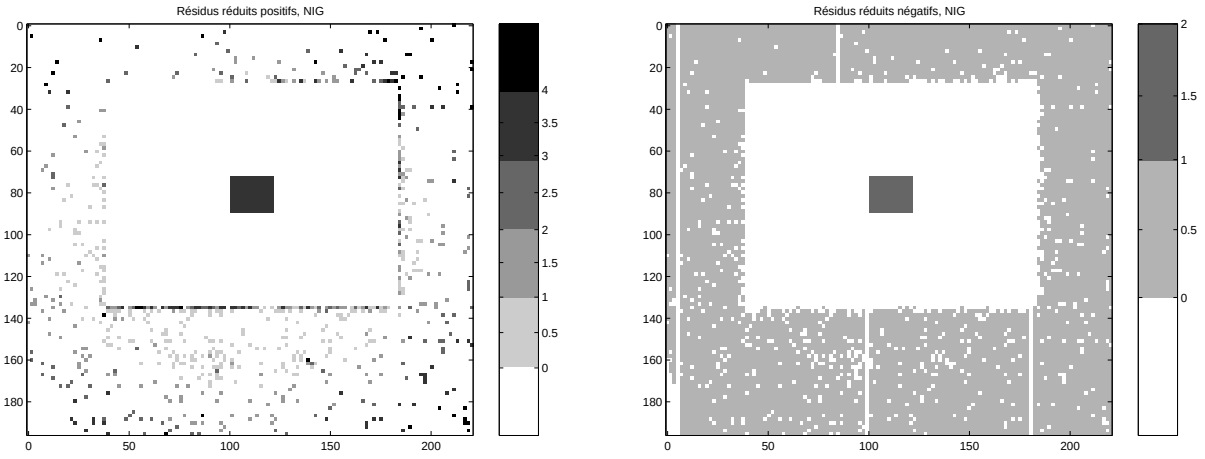


FIG. 3.6 – Résidus réduits sur le champ pour le modèle A dans le cas du trèfle

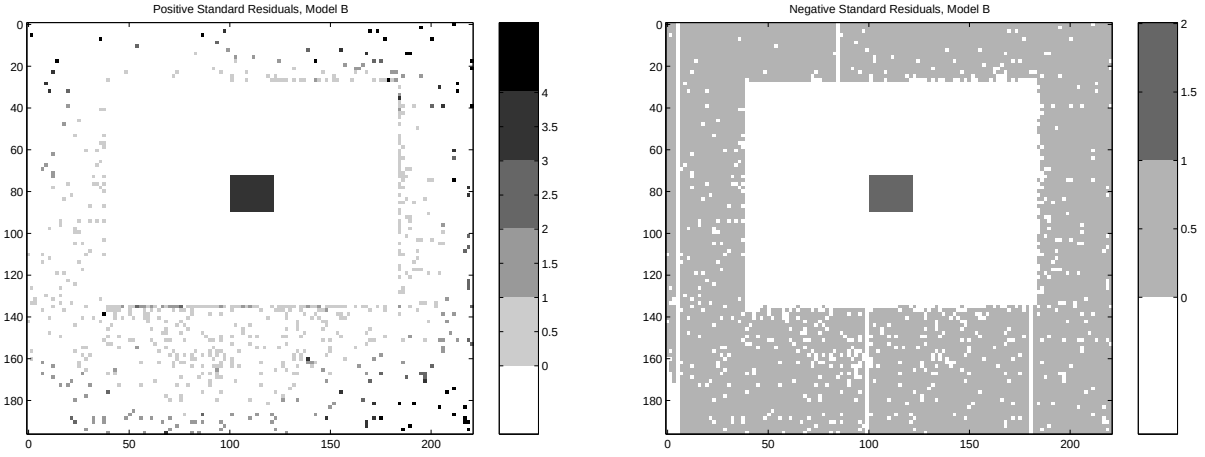


FIG. 3.7 – Résidus réduits sur le champ pour le modèle B dans le cas du trèfle

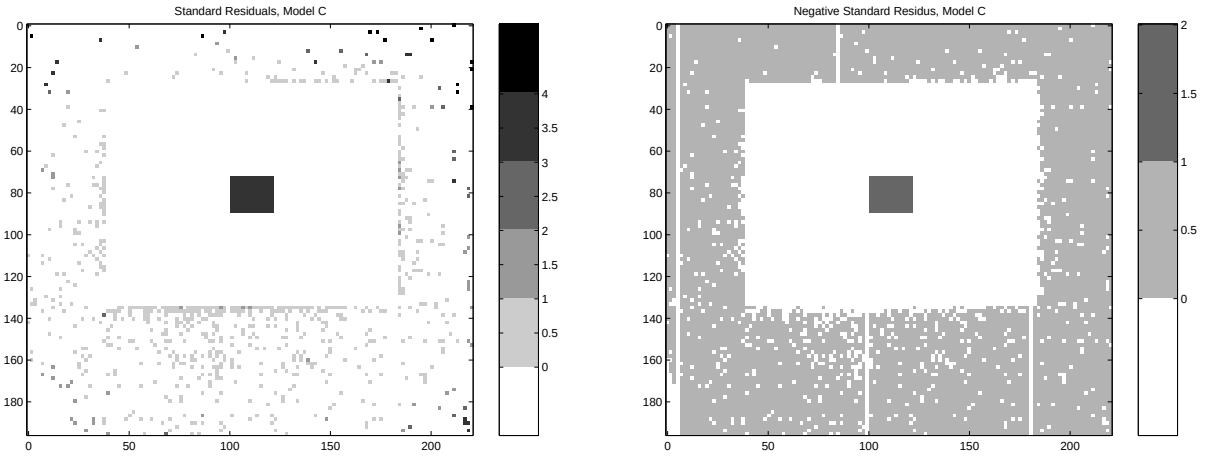


FIG. 3.8 – Résidus réduits sur le champ pour le modèle C dans le cas du trèfle

c) Courbes des fonctions de dispersion individuelles

Ci-dessous se trouvent les courbes des trois fonctions de dispersion individuelles étudiées dans trois directions : vers le bas du champ (sens 1) , vers la gauche du champ (sens 2) et enfin vers le haut du champ.
Dans tous les cas, la fonction de dispersion du modèle A (modèle NIG venant de la modélisation du milieu homogène) sous-estime la dispersion du pollen.
Ces graphiques confirment les résultats obtenus avant. En particulier, dans le sens opposé à la direction prédominante (vers le haut du champ), les fonctions de dispersion, associées aux modèles B et C, sont assez semblables ce qui coïncide avec le fait de la sous-estimation dans les deux cas.

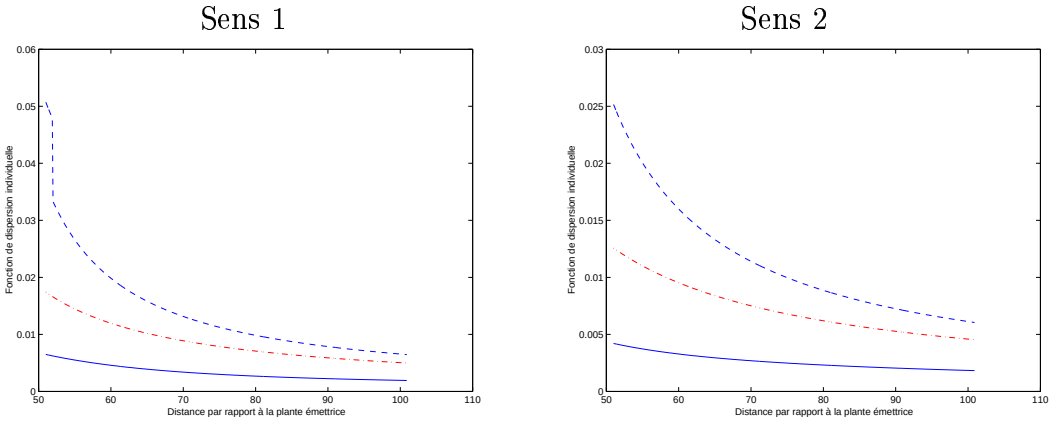


FIG. 3.9 – Courbes des trois fonctions de dispersion individuelles, le modèle A est en trait continu, le modèle B en rouge et en trait discontinu et le dernier modèle en bleu en pointillé régulier.

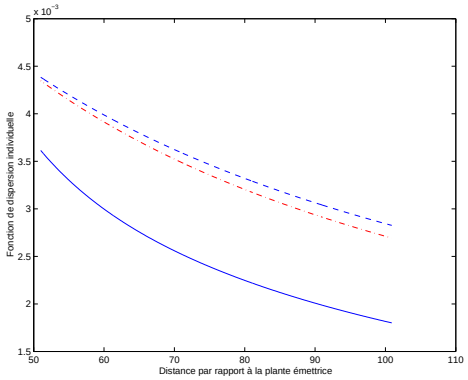


FIG. 3.10 – Courbes des trois fonctions de dispersion individuelles, le modèle A est en trait continu, le modèle B en rouge et en trait discontinu et le dernier modèle en bleu en pointillé régulier.

3.6.3 Expérience 2 : tournesol

On rappelle que le tournesol est une plante d'une hauteur plus haute (ou semblable) que le maïs.

a) Résultats

Les résultats des estimations des paramètres, pour l'expérience avec le tournesol, se trouvent dans le tableau (3.3) pour le modèle A, NIG, et dans le tableau (3.4) pour le modèle B, NIG "translatée" et le modèle C.

Pour les modèles B et C, le champ a été divisé en trois sous-ensembles S_1, S_2 et S_3. S_1 contient la direction prédominante du vent lors de la première semaine de floraison et S_2 la direction prédominante du vent lors de la deuxième semaine de floraison.
De plus, l'ensemble $\tilde{D}_3$ est constitué ici de deux côtés de la bordure : celui du haut et celui de droite.

On constate que la valeur du paramètre δ varie entre le modèle A et les modèles B et C (0.117, 0.389 et 0.394 respectivement). D'autre part le paramètre de sur-dispersion dans le modèle C est élevé (197) ce qui explique sans doute les écart-types élevés pour les paramètres δ, λ_x et λ_z.

Les paramètres de translation introduits dans les modèles B et C sont positifs dans les ensembles S_1 et S_2 mais négatifs dans S_3. Cela signifie que la zone de discontinuité a un effet d'accélération sur la dispersion dans le sens du vent. Il semble par contre que dans l'autre direction, il y ait un léger ralentissement. Dans le cas du modèle C, ces paramètres sont plus petits que dans le cas du trèfle ce qui signifie une accélération moindre de la dispersion.

Pour le modèle C, on trouve une valeur du paramètre q égale à 1.0510^{-3}.

	Modèle A : NIG	
Paramètres	Estimation	Ecart-type
δ	0.1175	0.0105
λ_x	-0.2012	0.0249
λ_y	0.0212	0.0105
λ_z	0.2742	0.0572
σ^2	6.914	
	[6.703 ;7.159]	

TAB. 3.3 – Résultats des estimations des paramètres pour le modèle A dans le cas du tournesol

	Modèle B : NIG "translatée"		Modèle C	
Paramètres	Estimation	Ecart-type	Estimation	Ecart-type
δ	0.3888	0.1827	0.3945	0.5411
λ_x	-0.1315	0.0561	-0.1282	0.1490
λ_y	-0.0403	0.0302	-0.0371	0.0543
λ_z	0.2291	0.0711	0.2069	0.3462
q	-	-	$1.0476\ 10^{-3}$	$9.156\ 10^{-4}$
$\alpha_{1,1}$	0.6014	0.0738	0.4957	0.2470
$\alpha_{1,2}$	0.5656	0.0617	0.3287	0.2157
$\alpha_{2,1}$	0.5632	0.0468	0.5509	0.2458
$\alpha_{2,2}$	0.8698	0.0628	0.8476	0.1028
$\alpha_{3,1}$	-0.4264	0.0941	-0.4966	0.2516
$\alpha_{3,2}$	-0.2234	0.1862	-0.6545	0.2364
σ^2	50.43		197.72	
	[47.91 ;52.98]		[187.85 ; 207.74]	

TAB. 3.4 – Résultats des estimations des paramètres pour les modèles B et C dans le cas du tournesol

b) Etude des résidus réduits

Ci-dessous, se trouvent les graphiques des résidus réduits sur le champ (figures 3.11, 3.12 et 3.13).

- Pour le modèle A, NIG, on constate de forts résidus positifs en bordure de la discontinuité.
- Pour le modèle B, NIG "translatée", la plupart des forts résidus en bordure de la discontinuité ont disparu. Et il n'y a pas de forts résidus à la limite du champ contrairement au cas du trèfle.
- Pour le modèle C, il reste quelques résidus élevés dans le coin en bas à gauche. Cependant le fait que le paramètre de sur-dispersion soit élevé entraîne des valeurs de résidus plutôt petites.

D'après ces résultats, le modèle ajustant le mieux les données est le modèle B dans le cas du tournesol.
Il semble que la direction des vents au cours de la pollinisation ait eu plus d'impact sur la dispersion du pollen dans le cas du tournesol que dans celui du trèfle. Une modélisation envisageable est de définir le paramètre q de manière plus précise. Par exemple, on pourrait considérer que c'est une fonction $q(\theta)$ dépendant de la direction θ qui semble être un facteur à ne pas négliger.

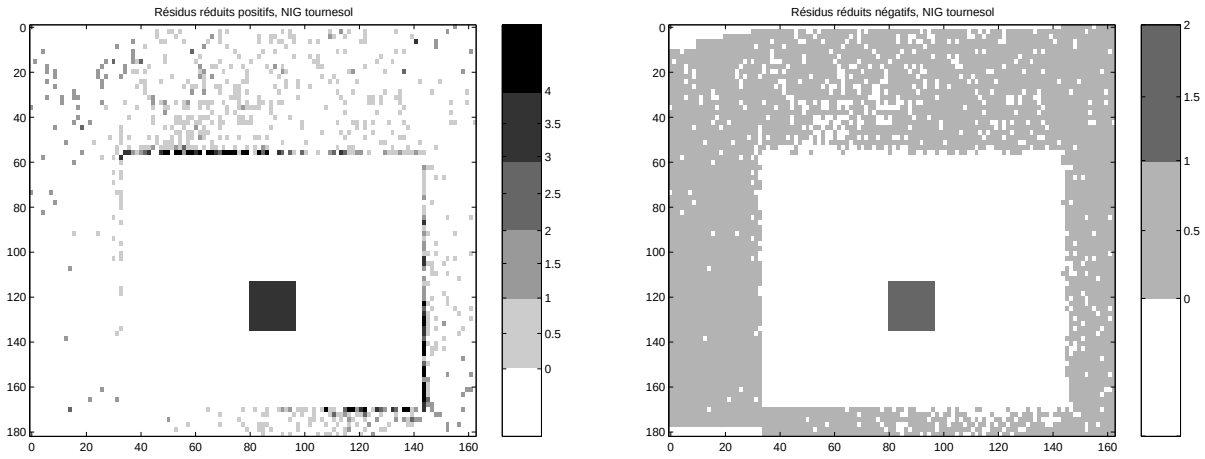


FIG. 3.11 – Résidus réduits sur le champ pour le modèle A dans le cas du tournesol

c) Courbes des fonctions de dispersion

Ci-dessous se trouvent les courbes des trois fonctions de dispersion individuelles étudiées dans trois directions comme précédemment : vers le haut du champ (sens

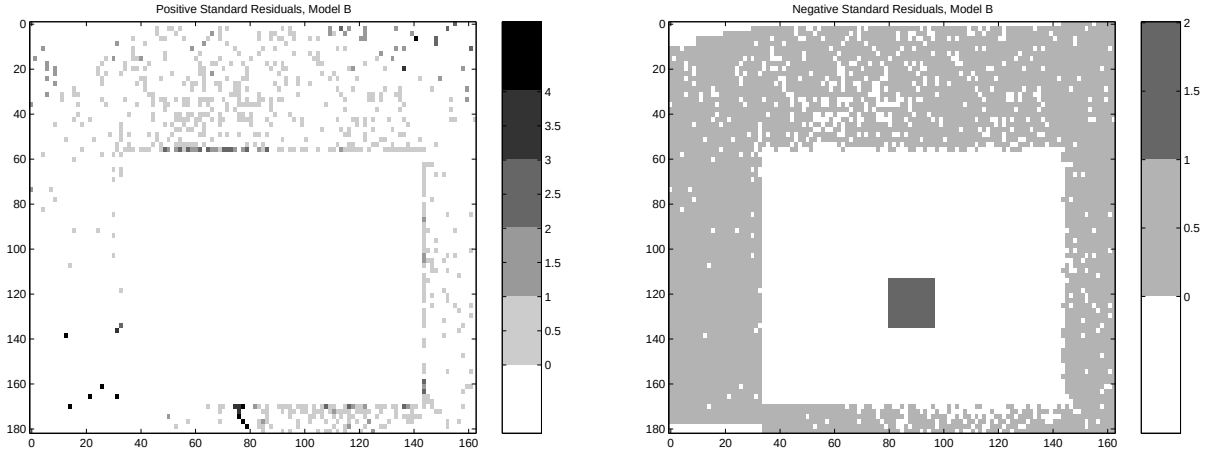


FIG. 3.12 – Résidus réduits sur le champ pour le modèle B dans le cas du tournesol

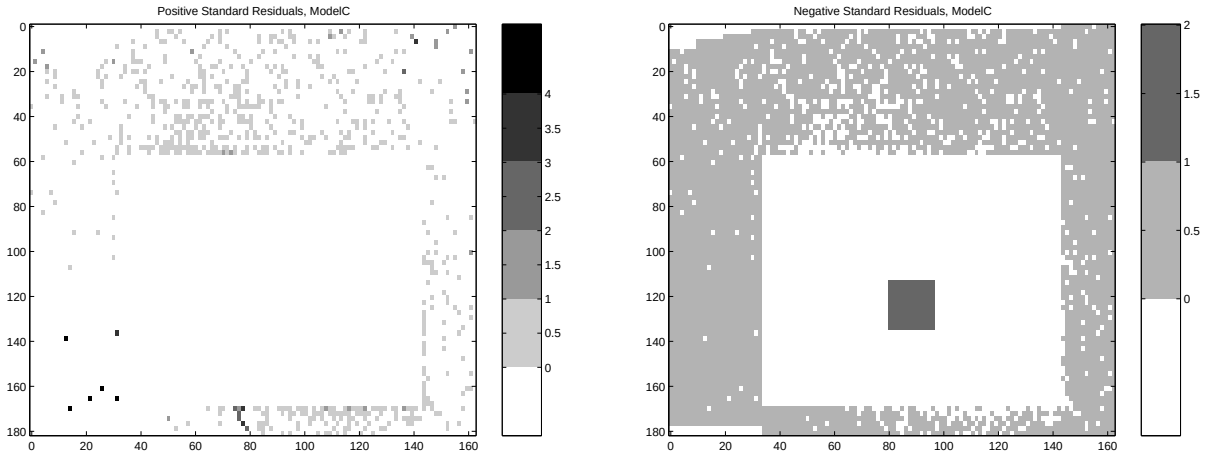


FIG. 3.13 – Résidus réduits sur le champ pour le modèle C dans le cas du tournesol

1), vers la droite du champ (sens 2) et vers le bas du champ.
Dans tous les cas, la fonction de dispersion du modèle A (modèle NIG venant de la modélisation du milieu homogène) sous-estime la dispersion du pollen (plus ou moins suivant les directions).
On peut constater que les courbes de dispersion ont des valeurs plus élevées dans le sens 2 et non le sens 1. C'est donc le sens 2 qui "prédomine" dans le cas du tournesol contrairement au trèfle. Pour le sens 3, la fonction associée au modèle B estime de façon légèrement plus élevée la dispersion par rapport au modèle C.

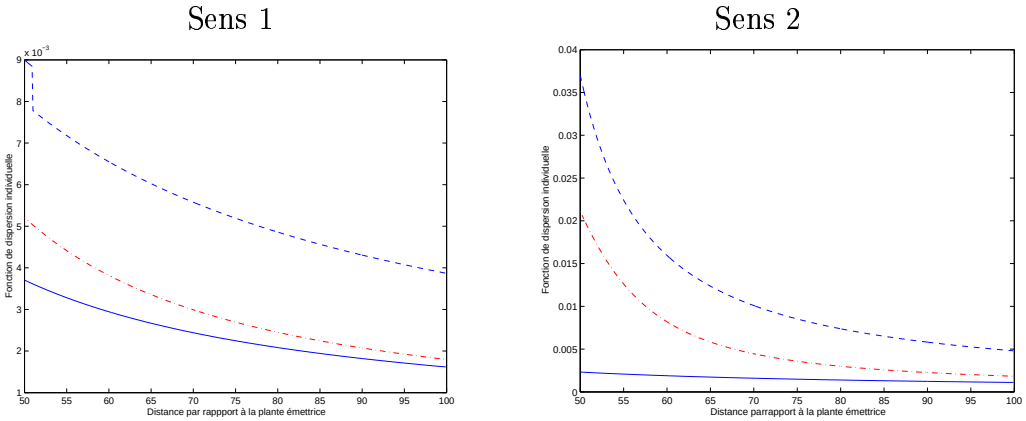


FIG. 3.14 – Courbes des trois fonctions de dispersion individuelles pour le tournesol, le modèle A est en trait continu, le modèle B en rouge et en trait discontinu et le dernier modèle en bleu en pointillé régulier.

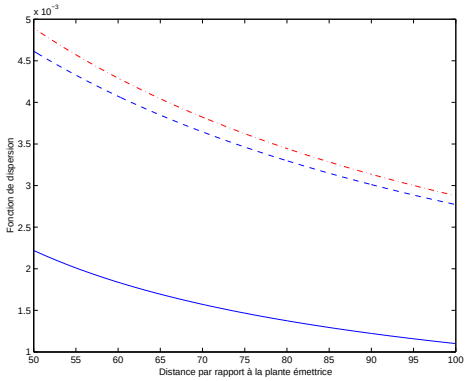


FIG. 3.15 – Courbes des trois fonctions de dispersion individuelles pour le tournesol, le modèle A est en trait continu, le modèle B en rouge et en trait discontinu et le dernier modèle en bleu en pointillé régulier.

3.7 Conclusion et discussion

L'estimation des paramètres pour les différents modèles a conduit au choix du modèle B, NIG "translatée", dans les deux cas (trèfle et tournesol).
Les résultats obtenus confirment l'analyse des données. Ils montrent qu'une zone de discontinuité accélère la dispersion du pollen, et ceci de façon plus forte à la frontière de la discontinuité, en particulier pour le trèfle.
Les résultats obtenus pour le modèle C le confirment : le paramètre q est plus grand pour le trèfle que pour le tournesol.
On avait vu que la bordure de la discontinuité, dans les cinq premiers mètres de bordure dans la direction prédominante du vent face à la parcelle de maïs bleu, était plus polluée dans le cas du trèfle par rapport au cas du tournesol (0.77 % pour le trèfle contre 0.55 % pour le tournesol). Le dépôt de pollen est donc plus important pour le trèfle. Cet accroissement du dépôt de pollen dans la zone située juste en aval de la parcelle pourrait s'expliquer par le fait que le maïs représente un obstacle après la zone de sol nu qui tend à faire croître les paramètres de turbulence (vertical ou horizontaux) à l'entrée du champ, avant de revenir à sa valeur initiale (Loubet *et al* (2004)).
Par contre la "pollution" globale des champs est similaire : 0.12 % pour le trèfle et 0.14 % pour le tournesol. Ainsi pour le tournesol, la pollinisation a tendance à plus se répartir sur l'ensemble du champ comparé au trèfle.

Les résultats obtenus nous apportent donc des informations utiles sur le comportement des grains de pollen de maïs traversant une zone de discontinuité de deux sortes : une constituée du trèfle et une autre constituée de tournesol. Cependant, ces modélisations ne nous permettent pas de faire des prédictions directement. En effet, l'introduction de paramètres de translation permet l'obtention d'un meilleur modèle. Mais comme nous n'avions à notre disposition qu'une seule expérience, nous ne pouvons pas savoir de quoi dépendent exactement ces paramètres. On peut penser qu'ils sont proportionnels à l'intensité du vent. D'autre part, ces modèles sont-ils encore valables pour des zones de discontinuité plus larges ou plus étroites ?

Il faudrait donc envisager de réaliser deux expériences simultanées, dans les mêmes conditions expérimentales : une en milieu homogène et l'autre en milieu hétérogène. Cela permettrait d'une part d'estimer les paramètres de la fonction de dispersion individuelle liés aux conditions météorologiques et aux caractéristiques du pollen grâce à l'expérience en milieu homogène, et d'autre part d'estimer les paramètres des modèles lors de l'expérience en milieu hétérogène et pouvoir ainsi regarder précisément quel est l'effet d'une zone de discontinuité sur ces paramètres, en particulier comment varient les paramètres de turbulence.

D'autre part, une modélisation plus précise du paramètre q introduit dans le modèle C pourrait être envisagée : une idée serait de définir q comme une fonction dépendant de la direction θ, facteur apparemment essentiel dans la modélisation.
Enfin, il serait intéressant d'essayer d'intégrer ces résultats au logiciel Mapod qui actuellement ne prend pas en compte les effets du paysage (discontinuité due à l'isolement ou une autre culture).

3.8 Annexe : démonstration de la proposition 3.4

Démonstration de la Proposition 3.1 (Paragraphe 3.4.1) :

Soit φ une fonction mesurable positive. Le but est de calculer $E[\varphi(X_{T_F})]$.
Trois cas sont possibles :

1. Ou bien X_{T_F} appartient à D_1 (c'est-à-dire qu'il y a fécondation dans D_1).
2. Ou bien $X_{T_F} = a_2$, c'est-à-dire que $\nu_{a_1} < T_h$ et $h_0 \leq Z_{\nu_{a_2}} \leq h$.
3. Ou bien X_{T_F} appartient à D_3 et $Z_{\nu_{a_2}} > h$.

Ainsi,

$$E[\varphi(X_{T_F})] = E[\varphi(X_{T_F})\mathbf{1}_{\{X_{T_F}\in D_1\}}]+E[\varphi(X_{T_F})\mathbf{1}_{\{X_{T_F}=a_2\}}]+E[\varphi(X_{T_F})\mathbf{1}_{\{X_{T_F}\in D_3\}\cap\{Z_{\nu_{a_2}}>h\}}].$$

Soit $T_1 = E[\varphi(X_{T_F})\mathbf{1}_{\{X_{T_F}\in D_1\}}]$, $T_2 = E[\varphi(X_{T_F})\mathbf{1}_{\{X_{T_F}=a_2\}}]$ et
$T_3 = E[\varphi(X_{T_F})\mathbf{1}_{\{X_{T_F}\in D_3\}\cap\{Z_{\nu_{a_2}}>h\}}]$.

Calcul du terme T_1 :

Si $X_{T_F} \in D_1$, alors $T_F = T_h$. Soit f_{X_t,x_0} la fonction de densité du processus X_t (partant du point x_0) et f_{T_h} la densité du temps d'atteinte, T_h, du niveau h par le processus (Z_t) partant de 0. On a

$$\begin{aligned} T_1 &= E[\varphi(X_{T_h})\mathbf{1}_{0\leq X_{T_h}\leq a_1}] \\ T_1 &= \int_0^{a_1} \varphi(x)\left\{\int_0^{+\infty} f_{X_t,x_0}(x) f_{T_h}(t)\,dt\right\} dx \end{aligned}$$

en utilisant une formule de Bayes car les variables X_t et T_h sont indépendantes.

Dans D_1, X_t suit une loi normale de moyenne $x_0 + f_x t$ et de variance $\tau_x^2 t$.
De plus, on sait que T_h suit une GIG de paramètres $\left(3/2, \frac{f_z^2}{2\tau_z^2}, \frac{h^2}{2\tau_z^2}\right)$ (voir Paragraphe 2.4.2 du chapitre 2 pour plus de précisions), de fonction de densité

$$f_{T_h}(t) = \frac{|h|\exp\left(\frac{f_z h}{2\tau_z^2}\right)}{\sqrt{2\pi}\tau_z}\frac{1}{t^{3/2}}\exp\left(-\frac{h^2}{2\tau_z^2 t} - \frac{f_z^2}{2\tau_z^2}t\right)\mathbf{1}_{\{t>0\}}$$

Alors, on a en posant $u = x - x_0$:

$$\begin{aligned} I(x-x_0,h,\theta) &= \int_0^{+\infty} f_{X_t,x_0}(x) f_{T_h}(t)\,dt \\ I(u,h,\theta) &= C(u,h,\theta)\int_0^{+\infty}\frac{1}{t^2}\exp\left(-\frac{p(u,h,\theta)}{t} - q(\theta)t\right)dt \end{aligned}$$

avec $p(u,h,\theta) = \dfrac{h^2}{2\tau_z^2} + \dfrac{u^2}{2\tau_x^2}$, $q(\theta) = \dfrac{f_z^2}{2\tau_z^2} + \dfrac{f_x^2}{2\tau_x^2}$ et

$$C(u,h,\theta) = \frac{|h|}{2\pi\tau_x\tau_z}\exp\left(\frac{f_z h}{\tau_z^2} + \frac{f_x u}{\tau_x^2}\right).$$

D'où, d'après la propriété 2 du chapitre 2 paragraphe 2.5.1 (Prudnikov *et al* 1986), on obtient

$$I(u,h,\theta) = C(u,h,\theta) \times 2\left(\frac{p(u,h,\theta)}{q(\theta)}\right)^{1/2} \mathcal{K}_{-1}\left(2\sqrt{q(\theta)p(u,h,\theta)}\right)$$

Finalement, on a

$$T_1 = \int_{\mathbb{R}} \varphi(x)[I(x-x_0,h,\theta)\mathbf{1}_{0\leq x\leq a_1}]\,dx \tag{3.11}$$

Calcul du terme T_2 :

Pour une trajectoire d'un grain de pollen qui va effectivement polliniser,on a $\{X_{T_F} = a_2\} = \{\nu_{a_1} < T_h\} \cap \{h_0 \leq Z_{\nu_{a_2}} \leq h\}$ avec

$$\begin{aligned}
\{\nu_{a_1} < T_h\} \cap \{h_0 \leq Z_{\nu_{a_2}} \leq h\} &= \{Z_{\nu_{a_1}} > h \text{ et } h_0 \leq Z_{\nu_{a_2}} \leq h\} \\
&= \{Z_{\nu_{a_1}} > h \text{ et } h_0 \leq Z_{\nu_{a_1}} + \bar{f}_z\bar{a} \leq h\} \text{ d'après (3.6)} \\
&= \{\max(h, h_0 - \bar{f}_z\bar{a}) \leq Z_{\nu_{a_1}} \leq h - \bar{f}_z\bar{a}\} \\
&= \{h_1 \leq Z_{\nu_{a_1}} \leq h_2\}
\end{aligned}$$

où on a posé $h_1 = \max(h, h_0 - \bar{f}_z\bar{a})$ et $h_2 = h - \bar{f}_z\bar{a}(> 0)$.

On obtient donc

$$\begin{aligned}
T_2 &= \varphi(a_2)E[\mathbf{1}_{\{h_1\leq Z_{\nu_{a_1}}\leq h_2\}}] \\
&= \varphi(a_2)\int_{h_1}^{h_2}\left\{\int_0^{+\infty} f_{Z_t}(z)f_{\nu_{a_1}}(t)\,dt\right\}dz
\end{aligned}$$

où f_{Z_t} représente la fonction de densité de Z_t et $f_{\nu_{a_1}}$ la densité du temps d'atteinte ν_{a_1}.
Comme $Z_{\nu_{a_1}} \in D_1$, alors $Z_{\nu_{a_1}}|\nu_{a_1} = t$ suit une loi normale de moyenne $f_z t$ et de variance $\tau_z^2 t$. De plus, on sait que ν_{a_1} suit une GIG de paramètres $\left(3/2, \frac{f_x^2}{2\tau_x^2}, \frac{(a_1-x_0)^2}{2\tau_x^2}\right)$ (voir Paragraphe 2.4.2 du chapitre 2 pour plus de précisions).
Ainsi, les calculs s'effectuant comme pour le terme T_1, on obtient en posant $u_1 = a_1 - x_0$

$$\begin{aligned}
J(z;u_1,\theta) &= \int_0^{+\infty} f_{Z_t}(z)f_{\nu_{a_1}}(t)\,dt \\
&= \tilde{C}(u_1,\theta)\exp\left(\frac{f_z}{\tau_z^2}z\right) \times 2\left(\frac{r(z,u_1,\theta)}{s(\theta)}\right)^{1/2} \\
&\qquad \times\mathcal{K}_{-1}\left(2\sqrt{s(\theta)r(z,u_1,\theta)}\right)
\end{aligned}$$

avec $r(z,u_1,\theta)=\dfrac{(u_1)^2}{2\tau_x^2}+\dfrac{z^2}{2\tau_z}$, $s(\theta)=\dfrac{f_x^2}{2\tau_x^2}+\dfrac{f_z^2}{2\tau_z^2}$ et

$\tilde{C}(u_1,\theta)=\dfrac{(u_1)}{2\pi\tau_x\tau_z}\exp\left(\dfrac{f_x u_1)}{\tau_x^2}\right)$.

En posant $J(u_1,\theta)=\displaystyle\int_{h_1}^{h_2} J(z,u_1,\theta)\,dz$, on obtient finalement

$$T_2=\varphi(a_2)\times J(a_1-x_0,\theta)=\int_{\mathbb{R}}\varphi(x)J(a_1-x_0,\theta)\mathbf{1}_{\{x=a_2\}}\,dx \tag{3.12}$$

Calcul du terme T_3 :

Dans ce cas, la pollinisation a lieu dans D_3.
$T_3=E[\varphi(X_{T_F})\mathbf{1}_{\{X_{T_F}>a_2\}\cap\{Z_{\nu_{a_2}}>h\}}]$ avec $T_F=\inf\{t>\nu_{a_2},Z_t=h\}$.
On a $\{Z_{\nu_{a_2}}>h\}=\{Z_{\nu_{a_1}}>h_2\}=\{\nu_{a_1}(X)<T_{h_2}(Z)\}$.

$$\begin{aligned}T_3 &= E\left[E^{\mathcal{F}_{\nu_{a_2}}}\left[\varphi(X_{T_F})\mathbf{1}_{\{X_{T_F}>a_2\}\cap\{\nu_{a_1}(X)<T_{h_2}(Z)\}}\right]\right]\\ &= E\left[\mathbf{1}_{\{\nu_{a_1}(X)<T_{h_2}(Z)\}}E\left[\varphi(X_{\tilde{T}_F})\mathbf{1}_{\{X_{\tilde{T}_F}>a_2\}}\right]\right]\end{aligned}$$

avec $\tilde{T}_F=\inf\{t>0,Z_t=h-Z_{\nu_{a_2}}\}=\inf\{t>0,Z_t=h_2-Z_{\nu_{a_1}}\}$. (et en utilisant le fait que les variables $T_{h_2}(Z)$ et $\nu_{a_1}(X)$ sont indépendantes.)

On introduit les notations suivantes :
- f_{X_t,a_2} la densité du processus de diffusion X_t partant de a_2, défini par $dX_t=f_x dt+\tau_x dB_t^1$.
- f_l la densité du premier temps d'atteinte d'un niveau l pour le processus de diffusion $Z_t=f_z t+\tau_z B_t^2$.
- $f_{\nu_{a_1},x_0}$ la densité du premier temps d'atteinte du niveau a_1 pour le processus de diffusion $X_t=f_x t+\tau_x B_t^1+x_0$.
- f_{Z_u} la densité du processus de diffusion Z_t.

Alors, on peut écrire :

$$T_3 = E\left[\mathbf{1}_{\{Z_{\nu_{a_1}}>h_2\}}E_{a_2}\left[\int_{t>0}\varphi(X_t)\mathbf{1}_{\{X_t>a_2\}}f_{h_2-Z_{\nu_{a_1}}}(t)\,dt\right]\right]$$

Alors, en utilisant le théorème de Fubini et une formule de Bayes (les variables $\nu_{a_1}(X)$ et Z_t étant indépendantes), on obtient

$$T_3=\int_{x>a_2}dx\varphi(x)\left\{\int_{z>h_2}\int_{t>0}\int_{u>0}f_{X_t,a_2}(x)f_{h_2-z}(t)f_{\nu_{a_1}}(u)f_{Z_u}(z)\,dz\,dt\,du\right\}$$

On pose alors

$$G(x, x_0, a_1, a_2, h_2, \theta) = \int_{z>h_2} \int_{t>0} \int_{u>0} f_{X_t,a_2}(x) f_{h_2-z}(t) f_{\nu_{a_1},x_0}(u) f_{Z_u}(z)\, dz\, dt\, du.$$

En utilisant une nouvelle fois le théorème de Fubini, on a

$$\begin{aligned} G(x, x_0, a_1, a_2, h_2, \theta) &= \int_{z>h_2} \left[\int_{u>0} f_{\nu_{a_1},x_0}(u) f_{Z_u}(z)\, du \right] \times \left[\int_{t>0} f_{X_t,a_2}(x) f_{h_2-z}(t)\, dt \right] dz \\ &= \int_{z>h_2} I(x - a_2, h_2 - z, \theta)\, J(z; a_1 - x_0, \theta)\, dz \end{aligned}$$

Finalement, on a

$$T_3 = \int_{\mathbb{R}} \varphi(x) [G(x, x_0, a_1, a_2, h_2, \theta \mathbf{1}_{x \le a_2}]\, dx \tag{3.13}$$

En sommant les trois termes calculés en (3.11), (3.12) et (3.13), on obtient le résultat souhaité.

Deuxième partie

Estimation dans des modèles à volatilité stochastique

Chapitre 4

Estimation for mean-reverting stochastic volatility models using Whittle methods alternatives

Sommaire

4.1 Introduction

The continuous-time stochastic volatility models have been the subject of many recent articles. These models, introduced by Hull and White (1987), have an important role in the financial and statistic fields (see e.g. Ghysels *et al* (1996) for a survey). An example is the so-called Black and Scholes model which models the log of an asset price by the solution to the stochastic differential equation : $dY_t = \mu(\sigma_t^2)dt + \sigma_t dW_t\,, \quad t \in [0, S]$, where σ_t^2 is the instantaneous volatility, (W_t) a standard Brownian motion and μ a real function.

In this document, we study the case where $V_t = \sigma_t^2$ is a mean-reverting diffusion process, and we consider the two-dimensional stochastic process (Y_t, V_t) given by :

$$(*) \qquad \begin{cases} dY_t = \sigma_t dB_t, \quad Y_0 = 0 \\ dV_t = \alpha(\beta - V_t)dt + a(\theta, V_t)dW_t \\ V_t = \sigma_t^2, \quad V_0 = \eta \end{cases}$$

where (B_t, W_t) is a standard Brownian motion of $\mathbb{R}^2$, η a variable independent of $(B_t, W_t)_{t\geq 0}$ and α, β, θ are unknown real parameters.

We don't observe the process (V_t) but a discrete sampling of the integrated process at regularly spaced times $Z_i = \dfrac{1}{\sqrt{\Delta}} \displaystyle\int_{(i-1)\Delta}^{i\Delta} \sigma_s \, dB_s$ with Δ fixed. This is equivalent to the observations of the increments $(Y_{i\Delta} - Y_{(i-1)\Delta})$.
The aim is to estimate the unknown parameter vector θ from the observations $(Z_i, 1 \leq i \leq n)$.

In the literature, a first approach is to make tend the number of observations n to infinity while the sampling interval $\Delta = \Delta_n$ tends to zero. The length of the observation time $n\Delta_n$ can be fixed or tends to infinity. For example, for a strictly stationary and ergodic process (V_t), Genon-Catalot *et al* (1999) propose an explicit method based on functions of the observations $(Y_{i\Delta_n}, 1 \leq i \leq n)$ (and empirical moment estimators (proposed by Genon-Catalot *et al* 1998) are included in this approach).

Another case is when the sampled Δ is fixed : this is what will be developed in this document. Different kinds of estimating functions have been studied (for example Sørensen (2000) considering a class of prediction-based estimating functions). In particular, various and simple to use "moment based methods" are proposed to estimate unknown parameters (for instance Hansen (1982) for generalized moment method for stationary and ergodic processes, Gallant *et al* (1997) used the efficient moment method). Genon-Catalot *et al* (2000) built also empirical moment estimators using the fact that discretely observed stochastic volatility models can be viewed as hidden Markov model (with a non-compact state space). A Kalman filter method can also used (Harvey, 1989).

However, the exact distribution of the integrated process is generally not explicit excepted for very few models. So the exact likelihood is difficult to compute expli-

citly. Then Genon-Catalot *et al* (2003) proposed a new contrast function for hidden Markov models, based on the conditional likelihood method. (This has the advantage to only need the existence of the first stationary distribution moment, contrary to moment methods which often require at least two moment conditions.)

In this document we are interested in mean-reverting hidden diffusion models for which the drift function b is equal to $b(\theta, V_t) = \alpha(\beta - V_t)$ in (*). These models have an specific structure used in the following. In particular, under good assumptions, the process $(Z_i, 1 \leq i \leq n)$ is strictly stationary and admits moments of order two. Moreover the process (Z_i^2) has a structure of ARMA(1, 1) (Genon-Catalot *et al*, 2003).
Then we have chosen to use the Whittle approximation to compute the Whittle estimator. However, in the small sample situation, the Whittle estimator may be bad and the Tapered Whittle estimator may be better.
If the structure of ARMA is not used, moments methods are often used. Hence we propose a two steps statistical method to estimate parameters : one parameter is estimated with an empirical moment method and the other parameters with a Whittle method.
Finally for two examples of mean-reverting models, we simulate several paths to compare the efficiency of these estimators (Whittle and Tapered Whittle estimators, empirical moments estimators and two steps estimator).

4.2 Mean reverting Stochastic Volatility Models

4.2.1 Definitions and assumptions

In the following, we take notations used by Genon-Catalot *et al* (2000).

We consider the model defined by a two-dimensional process (Y_t, V_t) with

$$dY_t = \sigma_t \, dB_t \text{ and } V_t = \sigma_t^2$$

V_t is called the volatility and is an unobserved Markov process independent of the Brownian motion (B_t).
We observe a discrete sampling of the integrated process (Z_i) (with a regular sampling interval Δ) defined by :

$$Z_i = \frac{1}{\sqrt{\Delta}}(Y_{i\Delta} - Y_{(i-1)\Delta}) = \frac{1}{\sqrt{\Delta}} \int_{(i-1)\Delta}^{i\Delta} \sigma_s \, dB_s \tag{4.1}$$

Conditionally on $(V_s, s \geq 0)$, the random variables (Z_i) are independent and Z_i has a distribution $\mathcal{N}(0, \bar{V}_i)$ with

$$\bar{V}_i = \frac{1}{\Delta} \int_{(i-1)\Delta}^{i\Delta} V_s \, ds \tag{4.2}$$

Such models have been first introduced by Hull and White (1987) with a diffusion process (V_t). Barndorff-Nielsen and Shepard (2001) propose analogous models with (V_t) being an Ornstein-Uhlenbeck Levy process.

In the following, we will consider (V_t) as a mean-reverting hidden diffusion process :

$$\begin{cases} dY_t = \sigma_t dB_t, \quad Y_0 = 0 \\ dV_t = \alpha(\beta - V_t)dt + a(V_t)dW_t \\ V_t = \sigma_t^2, \quad V_0 = \eta \end{cases} \tag{4.3}$$

with $\alpha > 0$, $\beta > 0$ and the real function a may also depend on unknown parameters.

Moreover we assume that $(B_t, W_t)_{t>0}$ is a standard Brownian motion of $\mathbb{R}^2$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$.
We make the following standard assumptions on the function $a(\theta, x)$ in order to ensure that equation (4.3) admits an unique strictly stationary and ergodic solution (V_t) with state space (l, r) included in $(0, \infty)$.

Let θ the parameters vector and assume that $\theta \in \Theta \subset \mathbb{R}^p$.

(A0) : η is a real random variable defined on Ω and independent of (W_t).

(A1) : For all $\theta \in \Theta$, $a(\theta, x)$ is a continuous (in x) real function on $\mathbb{R}$, and C^1 on (l, r) such that

$$\exists\, k > 0,\ \forall\, x \in (l, r),\ a^2(\theta, x) \leq k(1 + x^2) \text{ and } \forall\, x \in (l, r),\ a(\theta, x) > 0$$

For $x_0 \in (l, r)$, the scale and speed densities of diffusion (V_t) are for the mean-reverting case on the form :

$$s(\theta, x) = \exp\left(-2\int_{x_0}^{x} \frac{\alpha(\beta - u)}{a^2(\theta, u)}\, du\right), \quad \text{and } m(\theta, x) = \frac{1}{a^2(\theta, x)s(\theta, x)} \tag{4.4}$$

(A2) : For all $\theta \in \Theta$, it is assumed that

$$\int_{l^+} s(\theta, v)\, dv = +\infty, \quad \int^{r^-} s(\theta, v)\, dv = +\infty, \quad \int_l^r m(\theta, v)\, dv = M_\theta < +\infty$$

Hence, under (A0)-(A2), the process (V_t) lies in (l, r) included in $(0, \infty)$ and the stationary distribution of (V_t) is $\pi_\theta(dv) = \pi(\theta, v)dv$ with

$$\pi(\theta, v) = \frac{1}{M_\theta} \frac{1}{a^2(\theta, v)s(\theta, v)} \text{ for } v \in (l, r) \tag{4.5}$$

(A3) The initial random variable V_0 has a distribution $\pi_\theta(dv)$ and $E(V_0^2)$ is assumed finite.

Under (A0)-(A3), the process (Z_i) is strictly stationary and ergodic.
Indeed the process (Z_i) can be seen as a hidden Markov process. We define

$$U_i = (\bar{V}_i, V_{i\Delta}) \quad \text{for} i \geq 1$$

with introduced notations in (4.2).

Then , $(Z_i, i \geq 1)$ is a hidden Markov model with hidden chain $(U_i, i \geq 1)$. Moreover, under assumptions (A0)-(A3), $(U_i, i \geq 1)$ is a strictly stationary Markov chain with state space $(l, r)^2$ and ergodic. Genon-Catalot *et al* (2000) show that it is the same for the process (Z_i).

4.2.2 Specific properties for mean-reverting hidden diffusions

The unknown parameters vector θ is written

$$\theta = (b, \beta, c) \text{ with } b = \alpha\Delta > 0, \beta > 0 \text{ and } c \in \mathbb{R}^{(p-2)}$$

The mean-reverting hidden diffusion processes are specific stochastic volatility models, commonly used in finance. They possess some special properties due to the drift function form (Genon-Catalot *et al* (2000) ; Sørensen (2000)). In particular, assuming that the diffusion (V_t) satisfies (A1)-(A3), then the covariance structure of the process $(\bar{V}_i)$ has the following shape (Genon-Catalot *et al*, 2003) :

(i) $\mathrm{E}(\bar{V}_1) = \mathrm{E}(V_0) = \beta$ and

$$\mathrm{E}(\bar{V}_1{}^2) = \beta^2 + \mathrm{Var}(V_0)\, \frac{2(\alpha\Delta - 1 + \exp(-\alpha\Delta))}{\alpha^2\Delta^2} \tag{4.6}$$

(ii) For $k \geq 1$,

$$\mathrm{E}(\bar{V}_1\bar{V}_k) = \beta^2 + \mathrm{Var}(V_0)\, \frac{(1 - \exp(-\alpha\Delta))^2}{\alpha^2\Delta^2} \exp(-\alpha\Delta(k-1)) \tag{4.7}$$

These previous results allow to obtain properties of the process (Z_i^2) used in the following.

We define the quantities :

$$\gamma_0(\theta) = (1 + e^{-2b})(2\beta^2 + 6\,\mathrm{Var}(V_0)\, \frac{(b - 1 + e^{-b})}{b^2}) - 2e^{-b}\,\mathrm{Var}(V_0)\, \frac{(1 - e^{-b})^2}{b^2} \tag{4.8}$$

$$\gamma_1(\theta) = \frac{\mathrm{Var}(V_0)}{b^2}\left(1 - 6be^{-b} - 5e^{-2b} + 4e^{-b}\right) - 2e^{-b}\beta^2 \tag{4.9}$$

We make the assumption **(A4)** : θ is such that $\gamma_1(\theta) \neq 0$.

Under (A4), we define

$$\psi(\theta) = \frac{\gamma_0^2(\theta) - \sqrt{\gamma_0^2(\theta) - 4\gamma_1^2(\theta)}}{-2\gamma_1(\theta)} \text{ and } \sigma^2(\theta) = \frac{-\gamma_1(\theta)}{\psi(\theta)} \tag{4.10}$$

If (A4) is not satisfied, we define

$$\psi(\theta) = 0 \text{ and } \sigma^2(\theta) = \gamma_0(\theta) \tag{4.11}$$

To conclude we assume

(A5) : α is chosen such that $e^{-\alpha\Delta} \neq \psi(\theta)$.

Proposition 4.1 *:*
Assuming that the diffusion (V_t) defined in (4.3) satisfies (A0)-(A3), then the process $((Z_i^2-\beta), i\geq 1)$ is centered and ARMA(1,1). More precisely, we have

$$Z_i^2-\beta-e^{-b}(Z_{i-1}^2-\beta)=\varepsilon_i-\psi(\theta)\varepsilon_{i-1}$$

where (ε_i) is a centered white noise with $\mathrm{Var}(\varepsilon_i)=\sigma^2(\theta)$ and $|\psi(\theta)|<1$ (defined in 4.10 and 4.11).
Under (A5), the process is a causal and invertible ARMA(1,1).
Moreover the spectral density is given , for $\lambda\in[0,2\pi]$, by :

$$f(\lambda,\theta)=\frac{\sigma^2(\theta)}{2\pi}\frac{1+\psi(\theta)^2-2\psi(\theta)\cos\lambda}{1+e^{-2b}-2e^{-b}\cos\lambda} \tag{4.12}$$

Proof :

First, we compute $\mathrm{Var}(Z_i^2-\beta)$ and $\mathrm{Cov}(Z_i^2-\beta, Z_{i+k}^2-\beta)$ for $k\geq 1$ and $i\geq 1$.
The result is obtained easily using the above formulae (4.6), (4.7) and the fact that, conditionally on $(V_s, s\geq 0)$, the random variables (Z_i) are independent and Z_i has distribution $\mathcal{N}(0,\bar{V}_i)$. Then we have

$$\mathrm{Var}(Z_i^2-\beta)=2\,\mathrm{E}(\bar{V_1}^2)+\mathrm{Var}(\bar{V_1})=2\beta^2+6\,\mathrm{Var}(V_0)\,\frac{(\alpha\Delta-1+\exp(-\alpha\Delta))}{\alpha^2\Delta^2} \tag{4.13}$$

And for $k\geq 1$ and $i\geq 1$

$$\mathrm{Cov}(Z_i^2-\beta, Z_{i+k}^2-\beta)=\mathrm{Cov}(\bar{V_1},\bar{V_{1+k}})=\mathrm{Var}(V_0)\,\frac{(1-\exp(-\alpha\Delta))^2}{\alpha^2\Delta^2}\exp(-\alpha\Delta(k-1)) \tag{4.14}$$

Second, according to Genon-Catalot *et al* (2003), the process $((Z_i^2-\beta), i\geq 1)$ is centered and ARMA(1,1). We can write

$$X_i=Z_i^2-\beta-e^{-\alpha\Delta}(Z_{i-1}^2-\beta)=\varepsilon_i-\psi(\theta)\varepsilon_{i-1}$$

which is the canonical representation of X_i as a MA(1) process.

Let set γ the auto-covariance function of X_i. We know that for a MA(1) process, $\gamma(0)=\sigma^2(\theta)(1+\psi^2(\theta))$ and $\gamma(1)=-\sigma^2(\theta)\psi(\theta)$ (for $k\geq 1$, we have $\gamma(k)=0$).

Moreover we have $\gamma^2(0)-4\gamma^2(1)>0$ (since the spectral density is strictly positive).

If $\gamma(1)\neq 0$, we obtain (see also Genon-Catalot *et al* (2003))

$$\psi(\theta)=\frac{\gamma^2(0)-\sqrt{\gamma^2(0)-4\gamma^2(1)}}{-2\gamma(1)}\quad\text{and}\quad\sigma^2(\theta)=\frac{-\gamma(1)}{\psi(\theta)}$$

And if $\gamma(1)=0$, then $\psi(\theta)=0$ and $\sigma^2(\theta)=\gamma_0(\theta)$.

On the other hand, we can compute these two terms as following :

For the term $\gamma(0)$, we have

$\gamma(0) = \mathrm{Var}(X_i) = (1 + e^{-2\alpha\Delta})\,\mathrm{Var}(Z_1^2) - 2e^{-\alpha\Delta}\,\mathrm{Cov}(Z_1^2 - \beta, Z_2^2 - \beta)$.

And according to (4.13) and (4.14), we obtain $\gamma(0) = \gamma_0(\theta)$ (defined in 4.8).

For the term $\gamma(1)$, using the fact that $e^{-2\alpha\Delta}\,\mathrm{Cov}(Z_1^2, Z_2^2) - e^{-\alpha\Delta}\,\mathrm{Cov}(Z_1^2, Z_3^2) = 0$, we have with (4.9)

$$\gamma(1) = \mathrm{Cov}(Z_1^2, Z_2^2) - e^{-\alpha\Delta}\,\mathrm{Var}(Z_1^2) = \gamma_1(\theta)$$

Finally the polynomials $\phi(z) = 1 - e^{-\alpha\Delta}z$ and $\psi(z) = 1 - \psi(\theta)z$ have no zero in the unit circle ; and no common zeros under (A5) : $e^{-\alpha\Delta} \neq \psi(\theta)$. Then the process is causal and invertible.
Moreover the spectral density of an ARMA process has the expression

$$f(\lambda, \theta) = \frac{\sigma^2(\theta)}{2\pi}\frac{|1 - \psi(\theta)e^{-i\lambda}|^2}{|\phi(e^{-i\lambda})|^2} = \frac{\sigma^2(\theta)}{2\pi}\frac{1 + \psi(\theta)^2 - 2\psi(\theta)\cos\lambda}{1 + e^{-2b} - 2e^{-b}\cos\lambda}$$

Remark on assumption (A4) : A numerically study shows that
If $\alpha\Delta < 2.35$ then the first term of $\gamma_1(\theta)$ is negative and we have $\gamma_1(\theta) < 0$ for all $\theta \in \Theta$.
If $\alpha\Delta \geq 2.35$, it is possible that $\gamma_1(\theta) = 0$. With the additional assumption $\beta^2/\,\mathrm{Var}(V_0) > 0.2$, then we also have $\gamma_1(\theta) < 0$.

4.3 Whittle and Tapered Whittle Estimators

The aim is to estimate the parameters vector θ from a sample $Z^2(n)' = (Z_1^2, ..., Z_n^2)$. For a causal and invertible ARMA process, a natural approach is to maximize the Gaussian likelihood function. The associated maximum likelihood estimator has good asymptotic properties in particular when the white noise is a sequence of i.i.d. random variables (Brockwell and Davis 1991). If the process is an AR then the Durbin-Levinson algorithm can be also used. For a MA process, it is the innovations algorithm which can be used.

However, the considered process is not Gaussian. So this leads to use the likelihood approximation suggested by Whittle (1953) and the Whittle minimum contrast function defined by :

$$U_n(\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\ln(2\pi f(\lambda, \theta)) + \frac{I_n(\lambda)}{f(\lambda, \theta)}\right] d\lambda \tag{4.15}$$

where $I_n(\lambda)$ is the periodogram of the vector $Z^2(n)$ defined by

$$I_n(\lambda) = \frac{1}{2\pi n}\sum_{k,l=1}^{n}(Z_k^2 - \bar{Z}^2)(Z_l^2 - \bar{Z}^2)e^{i(k-l)\lambda} \tag{4.16}$$

with $\bar{Z}^2 = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} Z_j^2$.

The Whittle estimator is then defined by $\hat{\theta}_n = arg \inf_{\theta\in\Theta} U_n(\theta)$.

Remark : If the spectral density f is bounded and continuous on $\lambda \in \Pi$, we have $\lim_{n\to+\infty} E[I_n(\lambda)] = f(\lambda)$. However, $I_n(\lambda)$ is not generally a consistent estimator of $f(\lambda)$ (see Brockwell and Davis 1991 for example).

4.3.1 Whittle minimum contrast function computation

With notations introduced in section 4.2.2, we have

Proposition 4.2 *:*
We assume that (V_t) satisfies (A0)-(A3). Then, the minimum contrast function has the following form
– Under (A4) (i.e. $\gamma_1(\theta) \neq 0$),

$$U_n(\theta) = \frac{1}{2}\ln(\sigma^2(\theta)) + \frac{1}{2n\sigma^2(\theta)} \sum_{k,l=0}^{n-1} c(|k-l|,\theta)\tilde{Z}_k\tilde{Z}_l$$

with $\tilde{Z}_i = Z_i^2 - \bar{Z}^2$ for all $i \geq 1$ and

$$c(0,\theta) = \frac{e^{-b}}{\psi(\theta)} + \frac{-e^{-b}(1+\psi^2(\theta)) + (1+e^{-2b})\psi(\theta)}{\psi(\theta)(1-\psi^2(\theta))} \tag{4.17}$$

$$c(k,\theta) = \psi^{k-1}(\theta)\, \frac{-e^{-b}(1+\psi^2(\theta)) + (1+e^{-2b})\psi(\theta)}{(1-\psi^2(\theta))} \quad for\ k \geq 1 \tag{4.18}$$

– If $\gamma_1(\theta) = 0$, then $U_n(\theta) = \frac{1}{2}\ln\sigma^2(\theta) + \frac{1}{2n\sigma^2(\theta)} \sum_{k,l=0}^{n-1} d(k-l,\theta)\tilde{Z}_k\tilde{Z}_l$
with $d(0,\theta) = 1+e^{-2b}$, $d(1,\theta) = d(-1,\theta) = -e^{-b}$ and $d(k,b) = 0$ for $|k| \geq 2$.

Proof :

We use the spectral density expression calculated in (4.12) and the definition of $U_n(\theta)$ given in (4.15). We have three terms to compute. When $\gamma_1(\theta) \neq 0$, we follow the same way as Gloter (2000).

- The first term to compute is $\frac{1}{4\pi}\int_{-\pi}^{\pi} \ln(2\pi f(\lambda,\theta))\, d\lambda$.

We have

$$\ln(2\pi f(\lambda,\theta)) = \ln(\sigma^2(\theta)) + \ln(1+\psi(\theta)^2 - 2\psi(\theta)\cos\lambda) - \ln(1+e^{-2b} - 2e^{-b}\cos\lambda).$$

Using the following equality, (for example Rudin (1966), p.299), valid for all $|x| \leq 1$:

$$\int_{-\pi}^{\pi} \ln(1+x^2-2x\cos\lambda)\, d\lambda = 0$$

we obtain (since $|e^{-b}| < 1$ and $|\psi(\theta)| < 1$) that

$$\frac{1}{4\pi}\int_{-\pi}^{\pi} \ln(2\pi f(\lambda,\theta))\, d\lambda = \frac{1}{2}\ln(\sigma^2(\theta))$$

- The second term to compute is $T_2 = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} \frac{1+e^{-2b}-2e^{-b}\cos\lambda}{1+\psi^2(\theta)-2\psi(\theta)\cos\lambda}\, d\lambda$.

If $\psi(\theta) \neq 0$, we write

$$T_2 = \frac{1}{2i\pi}\int_S \frac{-e^{-b}z^2+(1+e^{-2b})z-e^{-b}}{z[-\psi(\theta)z^2+(1+\psi^2(\theta))z-\psi(\theta)]}\, dz$$

where S represent the unit circle.
Hence, we have to compute the integral of a rational fraction, called $F(z)$, in which the denominator doesn't nullify in S. The function F admits two simple poles in 0 and $\psi(\theta)$ in the interior of S and we have

$$Res(F,0) = \frac{e^{-b}}{\psi(\theta)}\ , \quad Res(F,\psi(\theta)) = \frac{-e^{-b}(1+\psi^2(\theta))+(1+e^{-2b})\psi(\theta)}{\psi(\theta)(1-\psi^2(\theta))}$$

Then the residue Theorem (for holomorphic functions) gives $T_2 = c(0,\theta)$ with $c(0,\theta)$ defined in (4.17).

If $\psi(\theta) = 0$, we have $T_2 = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi}(1+e^{-2b}-2e^{-b}\cos\lambda)\, d\lambda = 1+e^{-2b}$

- The last term to compute is, for $|k| \geq 1$:

$$T_3 = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\lambda}\frac{1+e^{-2b}-2e^{-b}\cos\lambda}{1+\psi^2(\theta)-2\psi(\theta)\cos\lambda}\, d\lambda = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i|k|\lambda}\frac{1+e^{-2b}-2e^{-b}\cos\lambda}{1+\psi^2(\theta)-2\psi(\theta)\cos\lambda}\, d\lambda$$

If $\psi(\theta) \neq 0$, by using the same way as for the second term we obtain :

$$T_3 = \frac{1}{2i\pi}\int_S z^{|k|-1}\frac{-e^{-b}z^2+(1+e^{-2b})z-e^{-b}}{(-\psi(\theta)z^2+(1+\psi^2(\theta))z-\psi(\theta)} = c(k,\theta) \qquad (4.19)$$

the introduced rational fraction having a single simple pole in $\psi(\theta)$ in the interior of S with the associated residue equal to $c(k,\theta)$.

If $\psi(\theta) = 0$, we have $T_3 = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi} e^{i|k|\lambda}(1+e^{-2b}-2e^{-b}\cos\lambda)\, d\lambda$

We deduce that $T_3 = -e^{-b}$ if $|k| = 1$ and $T_3 = 0$ otherwise.

Using the above computations we obtain the result of the Proposition 4.2.

4.3.2 Asymptotic properties

In this section, we assume that the function a has only one unknown parameter and we will estimate the parameter vector

$$\theta = (b, \beta^2, \mathrm{Var}(V_0)) \quad \text{where} \quad b = \alpha\Delta$$

. We assume that

(H0) θ lies in Θ which is a compact subset of $(\mathbb{R}^+)^3$ such that the process (V_t) satisfies assumptions (A0)-(A3). Moreover the true value of the parameter θ_0 belongs to $\mathring{\Theta}$.

For $\theta \in \mathring{\Theta}$, we define the information matrix for all $(i,j) \in \{1,2,3\}^2$:

$$\Gamma(\theta)_{i,j} = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_i} \left(\ln(2\pi f(\lambda,\theta))\right) \frac{\partial}{\partial \theta_j} \left(\ln(2\pi f(\lambda,\theta))\right) d\lambda \tag{4.20}$$

Let $f_{4,\theta}(x,y,z)$ be the fourth-order spectra (see for example Brillinger 1981 for definition), function of $\mathbb{R}^3$ which represents a degree of non-Gaussianity of the process. We define the matrix $B(\theta)$ by

$$B(\theta)_{i,j} = \frac{1}{8\pi} \int \int_{\Pi^2} \left(\frac{\partial}{\partial \theta_i} \ln(f(\lambda_1,\theta))\right) \left(\frac{\partial}{\partial \theta_j} \ln(f(\lambda_2,\theta))\right) \frac{f_{4,\theta}(\lambda_1,-\lambda_1,\lambda_2)}{f(\lambda_1,\theta) f(\lambda_2,\theta)} \, d\lambda_1 \, d\lambda_2 \tag{4.21}$$

In our case, under (A5), the ARMA(1, 1) process is causal, then we can write $Z_i^2 - \beta = \sum_{j=0}^{\infty} d_j \varepsilon_{i-j}$.

We define the function $D(\lambda) = \sum_{k=0}^{\infty} d_k e^{-i\lambda k}$ for $\lambda \in [0, 2\pi[$.

Then under (A5), according to Taniguchi and Kakizawa (2000) we have

$$B(\theta)_{i,j} = \frac{K(\theta)}{8\pi} \int_{\Pi} \left(\frac{\partial}{\partial \theta_i} \ln(f(\lambda_1,\theta))\right) \frac{D(\lambda_1)D(-\lambda_1)}{f(\lambda_1,\theta)} \, d\lambda_1 \times$$
$$\int_{\Pi} \left(\frac{\partial}{\partial \theta_j} \ln(f(\lambda_2,\theta))\right) \frac{D(\lambda_2)D(-\lambda_2)}{f(\lambda_2,\theta)} \, d\lambda_2$$

where $K(\theta)$ is a constant depending of the parameter θ.

And we define the functions

$$A(b) = 2(1+e^{-2b}), \quad B(b) = (b-1-e^{-b}+(3+b)e^{-2b}-e^{-3b})/b^2$$
$$C(b) = -2e^{-b} \text{ and } D(b) = (1-6be^{-b}+4e^{-b}-5e^{-2b})/b^2 \tag{4.22}$$

We make the assumption
(H1) : b_0 is such that $A(b_0)D(b_0) - B(b_0)C(b_0) \neq 0$.

A numerical study shows that (H1) is true if and only if $b_0 \neq b_1$ with $b_1 \simeq 2.1$.

Proposition 4.3 *:*
We assume that (V_t) satisfies (A0)-(A3). Moreover we assume (A5) and (H1). Then, we have

1. *The Whittle estimator $(\hat{\theta}_n)$ is consistent in θ_0.*
2. *The matrix $\Gamma(\theta_0)$ is non singular.*

3. $\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, \Gamma^{-1}(\theta_0)(\Gamma(\theta_0) + B(\theta_0))\Gamma^{-1}(\theta_0)\right)$

Proof :
Let set $\mu = (e^{-b}, \sigma^2, \psi)$. Then we have $\theta = (e^{-b}, \beta^2, \mathrm{Var}(V_0)) = g(\mu)$ where g is a function defined below. Indeed according to (4.8) and (4.9)we have

$$\gamma(0) = A(b)\beta^2 + B(b)\,\mathrm{Var}(V_0) = \sigma^2(1+\psi) \text{ and } \gamma(1) = C(b)\beta^2 + D(b)\,\mathrm{Var}(V_0) = -\sigma^2\psi$$

with $A(b), B(b), C(b)$ and $D(b)$ defined in (4.22).

So we have $U \begin{pmatrix} \beta^2 \\ \mathrm{Var}(V_0) \end{pmatrix} = \begin{pmatrix} \gamma_0(\theta) \\ \gamma_1(\theta) \end{pmatrix}$. The matrix U is non singular if and only if $A(b)D(b) - B(b)C(b) \neq 0$, i.e. under (H2).

Hence, if $U^{-1} = \begin{pmatrix} u_1 & u_2 \\ u_4 & u_3 \end{pmatrix}$ then we have $\theta = g(\mu)$ with

$$g(e^{-b}, \sigma^2, \psi) = (e^{-b}, u_1\sigma^2(1+\psi^2) - u_2\sigma^2\psi, u_3\sigma^2(1+\psi^2) - u_4\sigma^2\psi)$$

and the Jacobian matrix J_g is invertible under (H2).

So we have a classical ARMA(1, 1) process (Y_i) such that $Y_i - e^{-b}Y_{i-1} = \varepsilon_i - \psi\varepsilon_{i-1}$ with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
In this case, the estimator $\hat{\mu}$ is consistent and the associated information matrix $I(\mu)$ is non singular according to the fact that $\psi \neq e^{-b}$. Hence Γ is non singular and 1. and 3. of the proposition are obtained by applying Theorem 6.2 of Dahlhaus (1988) for example.

4.3.3 Tapered Whittle estimator

It is known that inference for time series requires great sample sizes for the observations. In particular here stochastic volatility models are seen as time series. And in the small sample situation, the Whittle estimator may be not be the most suited. Therefore, Dahlhaus (1988) suggests to use a tapered version of the periodogram : the ordinary periodogram is replaced by

$$\tilde{I}_T(\lambda) = \frac{1}{2\pi H_{2,T}} \sum_{k,l=0}^{T-1} h_{k,T}h_{l,T}(Z_k^2 - \bar{Z}^2)(Z_l^2 - \bar{Z}^2)e^{i(k-l)\lambda} \tag{4.23}$$

$$\text{where for } k \geq 1, H_{k,T} = \sum_{t=0}^{T-1} (h_{t,T})^k \tag{4.24}$$

Note that $H_{2,T}$ is a normalization coefficient associated to the Tapered periodogram. In the non tapered case, using the function $h(x) = 1$ if $x \in [0,1]$ and $h(x) = 0$ otherwise, we find $H_{2,T} = T$.

We assume that $h_{t,T} = h\left(\frac{t}{T}\right)$ with $h(x) = 0$ for all $x \notin [0,1)$.

Moreover, the function h has a maximum at $t = [T/2]$ and decreases as t tends to 0 or $T-1$. Hence the first and last observations have less weight.

In the following we chose the Tukey-Hanning taper

$$\begin{aligned} h_\rho(x) &= \frac{1}{2}[1-\cos(2\pi x/\rho)] \quad x \in [0, \frac{\rho}{2}) \\ &= 1 \quad x \in \left[\frac{\rho}{2}, 1/2\right] \\ &= h_\rho(1-x) \quad x \in (1/2, 1] \end{aligned} \tag{4.25}$$

According to Dahlhaus (1988), we will take $\rho = T^{-\kappa/3}$ with $\kappa < 0.25$ (see paragraph on simulations).

Then the ordinary periodogram is replaced by the tapered version in the minimum contrast function defined in (4.15). Thus, we have the tapered Whittle estimator always defined by $\tilde{\theta}_n = arg \inf_{\theta\in\Theta} U_n(\theta)$.

Asymptotic properties

Assuming that the assumptions in Proposition 4.3 hold, the Tapered Whittle estimator $(\tilde{\theta}_n)$ is consistent in θ_0. However the asymptotic normal distribution is slightly different.
Assuming that $\lim_{T\to+\infty} \frac{TH_{4,T}}{H_{2,T}^2}$ exists and is equal to a constant C with $H_{2,T}$ and $H_{4,T}$ defined in (4.24), we have, using the previous notations (Dahlhaus 1988)

$$\sqrt{T}(\tilde{\theta_T} - \theta_0) \xrightarrow[T\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, C\Gamma^{-1}(\theta_0)(\Gamma(\theta_0) + B(\theta_0))\Gamma^{-1}(\theta_0)\right)$$

Remarks :
- A polynomial type taper can also be used.
- In the non tapered case, consistency and convergence in distribution is deduced from the previous result.

Simulations results show that the estimation of the parameter β is not good although it is the expectation of the process (Z_i^2). However for stochastic volatility models there exists an empirical moments estimator of $E(V_0) = \beta$, easy to compute and powerful. Therefore the idea is to propose a two steps method : β is estimated using a moments method and the others parameters with the Whittle or Tapered Whittle method.

4.4 A two steps statistical method

In this section, the aim is to prove for (mean-reverting) stochastic volatility models defined in (4.3) the consistency and the convergence in distribution for an estimator of $\theta = (\omega, \beta)$ where the estimator of β, noted $\hat{\beta}_n$, converges almost surely to the true value and the parameter ω is estimated using the Whittle method.

First we recall a result obtained by Genon-Catalot *et al* (2000) and used in the following.

Under assumptions (A0)-(A3), we have

$$\hat{\beta}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i^2 \text{ is a strongly consistent estimator of } \beta \tag{4.26}$$

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right) \quad \text{with } \sigma^2 = \frac{2\,\mathrm{Var}(V_0)}{\alpha\Delta} + 2E(\bar{V}_1^2). \tag{4.27}$$

We assume

(H0') : $\theta = (\omega, \beta)$ with $\omega \in \Theta_1$, compact subset of $\mathbb{R}^s$ $(s \geq 1)$ and $\beta \in \Theta_2$, compact subset of $\mathbb{R}$.

We recall that under assumptions (A0)-(A3) the process (Z_i^2) is strictly stationary with a spectral density $f(\theta;\lambda)$.
Moreover, according to Theorem 2.4 of Dahlhaus (1988), the process (Z_1^2) seen as an ARMA(1, 1) process, satisfies all assumptions of Dahlhaus (1988). In particular under assumptions (A0)-(A3), (A5), (H0') and (H1), we have

(H2) : The functions f, $f_{4,\theta}$ and $\frac{\partial}{\partial\theta_i}f$ are continuous on $\Theta \times \Pi$.
And $\theta_1 \neq \theta_2$ implies $f(\theta_1;\lambda) \neq f(\theta_2;\lambda)$ on a positive Lebesgue measure.

(H3) : For $\theta_0 \in Int(\Theta)$, $0 < c_1 \leq f(\theta_0;\lambda)$ and $f(\theta_0;.)$ is a Lipschitz function.

(H4) : The functions $\dfrac{\partial^2}{\partial\theta_i\partial\theta_j}f(\theta;\lambda)$ are continuous on $\Theta \times \Pi$.

For $\omega \in \Theta_1$ we define $R_n(\omega) = U_n(\omega, \hat{\beta}_n)$ where $U_n(\theta)$ is the Whittle minimum contrast function defined in (4.15). Then we note $\hat{\omega}_n = arg\inf R_n(\omega)$. And we assume

(H5) : $\hat{\omega}_n$ uniquely exists and lies in $\mathring{\Theta}_1$.

Starting from definitions of $\Gamma(\theta_0)$ in (4.20) and $B(\theta_0)$ in (4.21), we introduce the following matrix :

The matrix $\Gamma_s(\theta_0) = (\Gamma(\theta_0)_{ij})_{1\leq i,j\leq s}$ and $B_s(\theta_0) = (B(\theta_0)_{ij})_{1\leq i,j\leq s}$ where s is the dimension of parameter ω.

Finally we define $\phi(\omega,\beta;\lambda) = (\phi_i(\omega,\beta;\lambda))_{1\leq i\leq s}$ with $\phi_i(\omega,\beta;\lambda) = \dfrac{1}{4\pi}\dfrac{\frac{\partial f(\omega,\beta;\lambda)}{\partial\omega_i}}{f(\omega,\beta;\lambda)}$.

And the function $g : \mathbb{R} \mapsto \mathbb{R}^s$, is defined by $g(\beta) = \displaystyle\int_\Pi \phi(\omega_0,\beta;\lambda)\,d\lambda$

Then we have the following theorem :

Theorem 4.1 *Under assumptions (A0)-(A3), (A5), (H0')-(H3) and (H5), the estimator $(\hat{\omega}_n)_n$ is consistent in ω_0.*

Proof : For $\phi : \Pi \to \mathbb{C}$, an integrable function, we define

$$J_n(\phi) = \int_\Pi \phi(\lambda) \frac{I_n(\lambda)}{f(\theta;\lambda)}\, d\lambda \quad \text{and} \quad J(\phi) = \int_\Pi \phi(\lambda)\, d\lambda$$

In the following, when needed, we will write$J_n(\phi(\theta))$ to precise that the function ϕ depends also on a parameter θ.

We also define $w(n,\eta) = \sup\{|R_n(u) - R_n(v)|; |u-v| \leq \eta\}$.

For the proof of the theorem we follow the same way as Dahlhaus (1988) (in the Whittle case the data taper h is taken equal to $h(t) = \mathbf{1}_{[0,1]}(t)$). The used Lemma are recalled in Appendix 4.6 and under assumptions (A0)-(A3), (A5), (H0') and (H1), they can be applied.
Recall that $Z(n)' = (Z_1^2, ..., Z_n^2)$.

a) To begin, we prove that for all $\omega_1 \in \Theta_1$, $\omega_1 \neq \omega_0$, there exists a constant $c(\omega_1) > 0$ with $\lim_{n\to+\infty} E_{\theta_0}\left[R_n(\omega_1) - R_n(\omega_0)\right] \geq c(\omega_1)$.

According to Dahlhaus (1988), for all $\theta_1 \in \Theta$ there exists a constant $c(\theta_1) > 0$ such that $\lim_{n\to+\infty} E_{\theta_0}\left[U_n(\theta_1) - U_n(\theta_0)\right] \geq c(\theta_1)$. So here we take $\theta_1 = (\omega_1, \beta_0)$.

Moreover we can write $R_n(\omega_1) = U_n(\omega_1,\beta_0) + T_{1,n}(\omega_1,\beta_0) + T_{2,n}(\omega_1,\beta_0)$ with

$$T_{1,n}(\omega_1,\beta_0) = \frac{1}{4\pi}\int_\Pi \Big(\ln(2\pi f(\omega_1,\hat\beta_n,\lambda)) - \ln(2\pi f(\omega_1,\beta_0,\lambda))\Big)\, d\lambda \text{ and}$$

$$T_{2,n}(\omega_1,\beta_0) = \frac{1}{4\pi}\int_\Pi I_n(\lambda)\left(\frac{1}{f(\omega_1,\hat\beta_n,\lambda)} - \frac{1}{f(\omega_1,\beta_0,\lambda)}\right) d\lambda$$

Using continuity of the function f and the fact that $(\hat\beta_n)$ converges a.s. to β_0, we deduce that $(T_{1,n}(\omega_1,\beta_0))$ converges a.s. to 0 when n tends to infinity.

Moreover for $\varepsilon > 0$, there exists n_0 such that for all $n \geq n_0$, $\left|\frac{1}{f(\omega_1,\hat\beta_n,\lambda)} - \frac{1}{f(\omega_1,\beta_0,\lambda)}\right| < \varepsilon$ a.s.

Note that $I_n(\lambda) = \dfrac{1}{2\pi n}\left|\sum_{k=1}^n X_k e^{-ik\lambda}\right|^2$ so $I_n(\lambda)$ is positive.
Hence we have, for all $n \geq n_0$,

$$|T_{2,n}(\omega_1,\beta_0)| \leq \frac{\varepsilon}{4\pi}\int_\Pi I_n(\lambda)\, d\lambda \quad a.s. \tag{4.28}$$

$$\mathrm{E}[|T_{2,n}(\omega_1,\beta_0)|] \leq \frac{\varepsilon}{4\pi}\,\mathrm{E}\left[\int_\Pi I_n(\lambda)\, d\lambda\right]$$

and using Lemma A1.3 (a) from Dahlhaus (1988), $\mathrm{E}\left[\int_\Pi I_n(\lambda)\, d\lambda\right] = \mathrm{E}[J_n(f)]$ converges to $J(f)$ when n tends to infinity.

We deduce that $\lim_{n\to+\infty} E_{\theta_0}(|T_{2,n}(\omega_1,\beta_0)|) = 0$ and hence $\lim_{n\to+\infty} E_{\theta_0}(T_{2,n}(\omega_1,\beta_0)) = 0$.

Finally we have $R_n(\omega_1) - R_n(\omega_0) = (U_n(\omega_1, \beta_0) - U_n(\omega_0, \beta_0)) +$

$$(T_{1,n}(\omega_1, \beta_0) - T_{1,n}(\omega_0, \beta_0)) + (T_{2,n}(\omega_1, \beta_0) - T_{2,n}(\omega_0, \beta_0))$$

So using the below results,

$$\lim_{n \to +\infty} E_{\theta_0} \left(R_n(\omega_1) - R_n(\omega_0) \right) = \lim_{n \to +\infty} E_{\theta_0} \left[U_n(\omega_1, \beta_0) - U_n(\omega_0, \beta_0) \right] > c(\omega_1)$$

b) Second, we prove that $\lim_{n \to +\infty} \text{Var}_{\theta_0} \left(R_n(\omega_1) - R_n(\omega_0) \right) = 0$.

We have $R_n(\omega_1) = \frac{1}{4\pi} \int_\Pi \ln(2\pi f(\omega_1, \hat{\beta}_n, \lambda))\, d\lambda + T_{2,n}(\omega_1, \beta_0) + J_n(f(\omega_1, \beta_0))$

We have $\lim_{n \to +\infty} \text{Var}_{\theta_0} \left[\int_\Pi \ln(2\pi f(\omega_1, \hat{\beta}_n, \lambda))\, d\lambda \right] = 0$ (using the Lebesgue theorem)

- Using (4.28) for $\varepsilon > 0$, there exists n_0 such that for all $n \geq n_0$,

$$|T_{2,n}(\omega_1, \beta_0)| \leq \varepsilon J_n(f(\omega_1, \beta_0)) \quad a.s.$$

So $E[(T_{2,n}(\omega_1, \beta_0))^2] \leq \varepsilon^2 E[J_n(f(\omega_1, \beta_0))^2]$.

But $E[J_n(f(\omega_1, \beta_0))^2] = \text{Var}[J_n(f(\omega_1, \beta_0))] + E[J_n(f(\omega_1, \beta_0))]^2$.

Then the first term converges to 0, while the second term to converges to $J(f)^2$.

Hence $\lim_{n \to +\infty} E[(T_{2,n}(\omega_1, \beta_0))^2] = 0$.

As $\lim_{n \to +\infty} E[T_{2,n}(\omega_1, \beta_0)] = 0$ also, we obtain $\lim_{n \to +\infty} Var[T_{2,n}(\omega_1, \beta_0)] = 0$

- Using Lemma A1.4 (a) from Dahlhaus (1988), $\lim_{n \to +\infty} \text{Var}_{\theta_0} \left(J_n(f(\omega_1, \beta_0)) \right) = 0$. So $\lim_{n \to +\infty} \text{Var}_{\theta_0} \left(R_n(\omega_1) \right) = 0$.

(using the fact that $\sqrt{\text{Var}(X+Y)} \leq \sqrt{\text{Var}(X)} + \sqrt{\text{Var}(Y)}$)

Hence we obtain the result, the same reasoning applying to $R_n(\omega_0)$

c) Third, using a Taylor formula, we have for a value $\omega \in \Theta_1$

$$R_n(\omega_2) - R_n(\omega_1) = \sum_{i=1}^{s} (\omega_2 - \omega_1)_i \frac{\partial}{\partial \theta_i} (R_n(\omega))$$

with $\frac{\partial}{\partial \theta_i}(R_n(\omega)) = \frac{1}{4\pi} \int_\Pi \frac{\frac{\partial}{\partial \theta_i} f(\omega, \hat{\beta}_n; \lambda)}{f(\omega, \hat{\beta}_n; \lambda)}\, d\lambda + \frac{\partial}{\partial \theta_i} \left[\frac{1}{4\pi n} Z(n)' * T_n \left(\frac{1}{f(\omega, \hat{\beta}_n; \lambda)} \right) * Z(n) \right]$.

Using Lemma A2.1 (j) from Dahlhaus (1988) upon matrix derivatives,

$$\frac{\partial}{\partial \theta_i} \left[T_n \left(\frac{1}{f(\omega, \hat{\beta}_n; \lambda)} \right) \right] = -T_n \left(\frac{1}{f(\omega, \hat{\beta}_n; \lambda)} \right) * T_n^{-1} \left(-\frac{\frac{\partial}{\partial \theta_i} f(\omega, \hat{\beta}_n; \lambda)}{f^2(\omega, \hat{\beta}_n; \lambda)} \right) * T_n \left(\frac{1}{f(\omega, \hat{\beta}_n; \lambda)} \right)$$

Then, as the function f and its derivative functions are continuous, as in Dahlhaus (1988), there exists a constant K such that

$$w(n,\eta) \leq K\eta \left(1 + \frac{1}{n} Z(n)' * T_n^{-1}(f(\omega_0, \beta_0; \lambda)) * Z(n)\right) \tag{4.29}$$

Using a), b) and 4.29 of c), we obtain Theorem 1 as in Dahlhaus (1988).

For a result of convergence in distribution, we assume the following conjecture :

Conjecture : The estimator $\hat{\omega}_n$ converges asymptotically in distribution to a Gaussian vector at the rate $\sqrt{n}$ under assumptions (A0)-(A3), (A5) and (H0')-(H5).

We have a difficulty here. Proving the asymptotic normality is a going work. However, according to the simulations results, it seems that the method is valid and that the conjecture is true.
In fact, we obtain below that $\sqrt{n}\nabla R_n(\omega_0)$ is the sum of two terms, each converging to a Gaussian distribution. However we don't know how to determinate the distribution of the sum. It is probably necessary to take back the proof of Dahlhaus results from the beginning.

Let us now detail the parts of the proof that we can handle concerning R_n.

We note $\nabla g_\theta = \left(\frac{\partial}{\partial \theta_i} g_\theta\right)_{i=1,\ldots,s}$.

By a Taylor formula, we have for $1 \leq i \leq s$:

$$\nabla R_n(\hat{\omega}_n)_i - \nabla R_n(\omega_0)_i = \sum_{j=1}^{s} \nabla^2 R_n(\omega_n^{(i)})_{ij} (\hat{\omega}_n - \omega_0)_j$$

with $|\omega_n^{(i)} - \omega_0| \leq |\hat{\omega}_n - \omega_0|$.

Since $\nabla R_n(\hat{\omega}_n) = 0$ (by definition) and $(\hat{\omega}_n)$ tends to ω_0 in probability, we prove the two following points :

(1) $\nabla^2 R_n(\omega_n) - \nabla^2 R_n(\omega_0)$ converges to 0 in probability and $\nabla^2 R_n(\omega_0)$ converges in probability to $\Gamma_s(\theta_0)$
(2) $\sqrt{n}\nabla R_n(\omega_0)$ converges in distribution to the sum of M_1 being a $\mathcal{N}(0, \sigma^2 \nabla g(\beta_0)' \nabla g(\beta_0))$ and M_2 being a $\mathcal{N}(0, \Gamma_s(\theta_0) + B_s(\theta_0))$.

Proof of (1) :
We have $\dfrac{\partial^2}{\partial \omega_i \partial \omega_j} R_n(\omega) = S_{1,n}(\omega) - S_{2,n}(\omega) + 2S_{3,n}(\omega) - S_{4,n}(\omega)$ with :

$$S_{1,n}(\omega) = \frac{1}{4\pi} \int_\Pi \frac{\frac{\partial^2 f(\omega, \hat{\beta}_n; \lambda)}{\partial \omega_i \partial \omega_j}}{f(\omega, \hat{\beta}_n; \lambda)} \, d\lambda; \quad S_{2,n}(\omega) = \frac{1}{4\pi} \int_\Pi I_n(\lambda) \frac{\frac{\partial^2 f(\omega, \hat{\beta}_n; \lambda)}{\partial \omega_i \partial \omega_j}}{f^2(\omega, \hat{\beta}_n; \lambda)} \, d\lambda$$

$$S_{3,n}(\omega) = \frac{1}{4\pi} \int_\Pi I_n(\lambda) \frac{\frac{\partial f(\omega, \hat{\beta}_n; \lambda)}{\partial \omega_i} \frac{\partial f(\omega, \hat{\beta}_n; \lambda)}{\partial \omega_j}}{f^3(\omega, \hat{\beta}_n; \lambda)} \, d\lambda; \quad S_{4,n}(\omega) = \frac{1}{4\pi} \int_\Pi \frac{\frac{\partial f(\omega, \hat{\beta}_n; \lambda)}{\partial \omega_i} \frac{\partial f(\omega, \hat{\beta}_n; \lambda)}{\partial \omega_j}}{f^2(\omega, \hat{\beta}_n; \lambda)} \, d\lambda$$

Let set $\psi(\omega,\beta;\lambda)=\dfrac{1}{4\pi}\dfrac{\frac{\partial^2 f(\omega,\beta;\lambda)}{\partial\omega_i\partial\omega_j}}{f(\omega,\beta;\lambda)}$ and $S_2(\omega_0,\beta_0)=\dfrac{1}{4\pi}\displaystyle\int_\Pi \psi(\omega_0,\beta_0;\lambda)\,d\lambda = J(\psi(\omega_0,\beta_0))$.

Using the facts that $(\hat\beta_n)$ converges to β_0 a.s and $(\hat\omega_n)$ converges to ω_0 in probability ; and that the functions f and second derivatives functions are continuous in θ, we have $S_{4,n}(\omega_0)$ and $S_{4,n}(\omega_n)$ converge to $\Gamma_s(\omega_0)$ in probability and $S_{1,n}(\omega_0)$ and $S_{1,n}(\omega_n)$ converge to $S_2(\omega_0,\beta_0)$ in probability.

We have $S_{2,n}(\omega_n)-S_2(\omega_0,\beta_0)=[J_n(\psi(\omega_n,\beta_n))-J_n(\psi(\omega_0,\beta_0))]+$

$$[J_n(\psi(\omega_0,\beta_0))-J(\psi(\omega_0,\beta_0))]$$

The first term tends to 0 in probability for the same reason as in the first part of the proof of Theorem 4.1 (equation (4.28)). The second term tends also to 0 using Lemma A1.3 (a) and Lemma A1.4 (a) from Dahlhaus (1988).
(For $\varepsilon>0$, we have $\mathbb{P}(|J_n(\psi)-J(\psi)|\geq\varepsilon)\leq\frac{1}{\varepsilon^2}(\mathrm{Var}(J_n(\psi))+E\left[((J_n(\psi)-J(\psi))^2\right])$

Hence we deduce that $S_{2,n}(\omega_0)$ and $S_{2,n}(\omega_n)$ converge to $S_2(\omega_0,\beta_0)$ in probability under $\mathbb{P}_{\theta_0}$.

In a same way, $S_{3,n}(\omega_0)$ and $S_{3,n}(\omega_n)$ converge to $S_3(\omega_0,\beta_0)$ in probability under $\mathbb{P}_{\theta_0}$, with

$$S_3(\omega_0,\beta_0)=\frac{1}{4\pi}\int_\Pi \frac{\frac{\partial f(\omega_0,\beta_0;\lambda)}{\partial\omega_i}\frac{\partial f(\omega_0,\beta_0;\lambda)}{\partial\omega_j}}{f^2(\omega_0,\beta_0;\lambda)}\,d\lambda=\Gamma_s(\omega_0)_{i,j}$$

Then we obtain (1).

Proof of (2) :

We can write

$$\sqrt{n}\nabla R_n(\omega_0)=\sqrt{n}[g(\hat\beta_n)-g(\beta_0)]-\sqrt{n}[J_n(\phi(\omega_0,\beta_0))-J(\phi(\omega_0,\beta_0))]+D_n$$

where the function g is defined by $g(\beta)=\displaystyle\int_\Pi \phi(\omega_0,\beta;\lambda)\,d\lambda$

where $\phi(\omega,\beta;\lambda)=(\phi_i(\omega,\beta;\lambda))_{1\leq i\leq s}$ and $\phi_i(\omega,\beta;\lambda)=\dfrac{1}{4\pi}\dfrac{\frac{\partial f(\omega,\beta;\lambda)}{\partial\omega_i}}{f(\omega,\beta;\lambda)}$.

Note that $g(\beta)=J(\phi(\omega_0,\beta))$.

Using the Delta method, the first term converges in distribution to a $\mathcal{N}(0,\nabla g(\beta_0)'\sigma^2\nabla g(\beta_0))$. According to Theorem A1.2 from Dahlhaus (1988) and definitions of $\Gamma_s(\theta_0)$ and $B_s(\theta_0)$, the second term converges in distribution to a $\mathcal{N}(0,\Gamma_s(\theta_0)+B_s(\theta_0))$.

The third term D_n tends in probability to 0. Indeed we have

$$D_n=\sqrt{n}\frac{1}{4\pi}\int_\Pi I_n(\lambda)\left[\frac{\frac{\partial f(\omega_0,\beta_0;\lambda)}{\partial\omega}}{f^2(\omega_0,\beta_0;\lambda)}-\frac{\frac{\partial f(\omega_0,\hat\beta_n;\lambda)}{\partial\omega}}{f^2(\omega_0,\hat\beta_n;\lambda)}\right]d\lambda$$

As previous, we can write $|D_n^i|\leq\varepsilon_n\times(\sqrt{n}\int_\Pi I_n(\lambda)\,d\lambda)$

with (ε_n) converging to 0 a.s., so in probability. And according to Theorem A1.2, $\left(\sqrt{n}\int_\Pi I_n(\lambda)\,d\lambda\right)$ converges in distribution then it is bounded in probability.
Therefore we obtain the result.

The remaining part is to study the joint distribution of these two random variables.

4.5 Study on simulations

In this section, the aim is to compare the qualities of the studied estimators : Whittle and Tapered Whittle estimators, two steps estimators and empirical moments estimators.

Simulation of the sample path (Z_i) :

First we simulate the diffusion process (V_t) using an Euler scheme with sampling interval p (p strictly positive). Thus, we approximate the diffusion process
$V_t = \mu(V_t)dt + a(V_t)dW_t$ by the process $(\tilde{V_n})$ defined by

$$\tilde{V_0} = V_0 \text{ and for } n \geq 0, \tilde{V_{n+1}} = \tilde{V_n} + \mu(\tilde{V_n})p + a(\tilde{V_n})\delta W_n$$

where $\delta W_n = W_{(n+1)p} - W_{np}$ are i.i.d. Gaussian random variables with mean 0 and variance p. Here, we will take $p = \dfrac{\Delta}{50}$.
Second, for each i, we obtain an approximation of the integral $\bar{V_i}$ (using a Riemann sum). Finally, we simulate the variable $(Z_i, 1 \leq i \leq n)$.
We have used 150 simulated paths of (Z_i) for different values of n and Δ. Then we computed the different estimators for each path (as well as empirical variance). [1]

4.5.1 Examples

Hereafter, we will focus in particular on two models often used in the financial field.

First, we consider the Model 1 defined by :

$$\begin{cases} dY_t = \sigma_t dB_t, \quad Y_0 = 0 \\ dV_t = \alpha(\beta - V_t)dt + cV_t dW_t \\ V_t = \sigma_t^2, \quad V_0 = \eta \end{cases}$$

with $\alpha > 0, \beta > 0$, $c > 0$ and $\theta = (\alpha, \beta, c)$.
This model appears to be the diffusion approximation of a GARCH(1,1)-M model (Nelson 1990).

Assumption (A1) is verified with $(l, r) = (0, +\infty)$.
Let $a = 1 + \dfrac{2\alpha}{c^2}$ and $\mu = \dfrac{2\alpha\beta}{c^2}$.

[1] All computations were done using Matlab software.

Assumption (A2) holds if and only if $\mu > 0$ and $a > 0$. Then, the stationary distribution is an inverse Gamma distribution with parameters (a, μ).

We need $E(V_0^2) < +\infty$ for assumption (A3). This implies $a > 2$ or $\frac{2\alpha}{c^2} > 1$.

Then, we have $E(V_0) = \beta$ and $\text{Var}(V_0) = \frac{\beta^2}{(2\alpha/c^2 - 1)}$.

Moreover, we have $\theta \in \Theta$ where Θ is a compact subset of
$E_1 = \{\theta = (\alpha, \beta, c) \in (R^+)^3$ such that $\alpha > 0.5c^2\}$.

For the study, the path (Z_i) is simulated for

$$\beta_0 = 2, \alpha_0 = 3, c_0 = \sqrt{2} \text{ and for } \Delta = 0.1 \text{ or } 1$$

For these values, assumptions (A4)-(A5) and (H1) hold.

The Model 2 was proposed by Heston (1993). We consider for (V_t) the classical square-root process used by Cox *et al.* (1985) for interest rates :

$$dV_t = \alpha(\beta - V_t)dt + c\sqrt{V_t}dW_t$$

Assumption (A1) is always satisfied with $(l, r) = (0, +\infty)$.

Here we set $a = \frac{2\alpha\beta}{c^2}$ and $\mu = \frac{2\alpha}{c^2}$.

Assumption (A2) holds if and only if $\mu > 0$ and $a \geq 1$. Then, the stationary distribution is a Gamma distribution with parameters (a, μ).

The Gamma distribution has moments of order q for all q positive. So (A3) holds.

In particular, we have $E(V_0) = \beta$ and $\text{Var}(V_0) = \frac{\beta^2}{a}$.

Moreover, we have $\theta \in \Theta$ where Θ is a compact subset of
$E_2 = \{\theta = (\alpha, \beta, c) \in (R^+)^3$ such that $2\alpha\beta > c^2\}$.

For the study, the path (Z_i) is simulated for

$$\beta_0 = 1.5, \alpha_0 = 1, c_0 = \sqrt{2} \text{ and for } \Delta = 0.1 \text{ or } 1$$

. For these values, assumptions (A4)-(A5) and (H1) hold.

4.5.2 Results and discussion

a) Results for Whittle and Tapered Whittle methods

For Model 1, the results are in Table (5.1). And for Model 2, the results are in Table (5.2).

The overall comparison shows that results with the Tapered Whittle method are always better than with the Whittle method. It is the same for the empirical standard deviations which decrease when n increases.

We can also notice that the higher is the observations time $T = n\Delta$, the better are the estimates. It is in agreement with the fact that estimation in stochastic volatility models requires large sample sizes for the interval of observations.

For $\Delta = 1$, we remark that the variables (Z_i) are almost uncorrelated according to (4.9). The results are good for both models with a little plus for Model 1.

For $\Delta = 0.1$, the estimate of β is quite satisfying for $n = 2000$.
For model 2, the results are satisfying for $n = 2000$. However for Model 1, the estimates of parameters b and c are not very good even for $n = 2000$. A reason suggested by Genon-Catalot *et al* (1999) could be that the parameter involving the number of degrees of freedom of a Student distribution is badly estimated even for independent observations (Blattberg and Gonedes, 1974).
For Model 2, if the parameters are estimated starting from a discretely observation of the process (V_i), good results are obtained using a method proposed by Kessler (2000) for example, for α and β .

n, Δ	True parameters	Parameters	Whittle estimator		Tapered Whittle estimator	
			Mean	Standard deviation	Mean	Standard deviation
$n = 500$	$\beta_0 = 2$	$\hat{\beta}$	1.768	0.488	1.787	0.475
$\Delta = 0.1$	$b_0 = 0.3$	$\hat{b}$	0.683	0.529	0.499	0.297
$n\Delta = 50$	$c_0 = \sqrt{2}$	$\hat{c}$	2.067	0.969	1.824	0.771
$n = 1000$	$\beta_0 = 2$	$\hat{\beta}$	1.889	0.391	1.917	0.275
$\Delta = 0.1$	$b_0 = 0.3$	$\hat{b}$	0.584	0.454	0.458	0.252
$n\Delta = 100$	$c_0 = \sqrt{2}$	$\hat{c}$	1.829	0.850	1.665	0.583
$n = 2000$	$\beta_0 = 2$	$\hat{\beta}$	1.967	0.154	1.977	0.148
$\Delta = 0.1$	$b_0 = 0.3$	$\hat{b}$	0.451	0.235	0.415	0.183
$n\Delta = 200$	$c_0 = \sqrt{2}$	$\hat{c}$	1.667	0.359	1.592	0.148
$n = 500$	$\beta_0 = 2$	$\hat{\beta}$	1.931	0.148	1.954	0.127
$\Delta = 1$	$b_0 = 3$	$\hat{b}$	3.060	0.143	3.023	0.111
$n\Delta = 500$	$c_0 = \sqrt{2}$	$\hat{c}$	1.525	0.107	1.421	0.085
$n = 1000$	$\beta_0 = 2$	$\hat{\beta}$	1.999	0.132	1.990	0.119
$\Delta = 1$	$b_0 = 3$	$\hat{b}$	2.994	0.136	3.014	0.104
$n\Delta = 1000$	$c_0 = \sqrt{2}$	$\hat{c}$	1.425	0.068	1.424	0.080

TAB. 4.1 – Whittle and Tapered Whittle estimators for Model 1 and for $\beta_0 = 2$, $\alpha_0 = 3$, $c_0 = \sqrt{2}$, $b_0 = \alpha_0\Delta$ and for 150 simulations

n, Δ	True parameters	Parameters	Whittle estimator		Tapered Whittle estimator	
			Mean	Standard deviation	Mean	Standard deviation
$n = 500$	$\beta_0 = 1.5$	$\hat{\beta}$	1.415	0.338	1.422	0.464
$\Delta = 0.1$	$b_0 = 0.1$	$\hat{b}$	0.140	0.106	0.140	0.151
$n\Delta = 50$	$c_0 = \sqrt{2}$	$\hat{c}$	1.459	0.677	1.413	0.347
$n = 1000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.459	0.229	1.469	0.215
$\Delta = 0.1$	$b_0 = 0.1$	$\hat{b}$	0.121	0.112	0.114	0.076
$n\Delta = 100$	$c_0 = \sqrt{2}$	$\hat{c}$	1.394	0.408	1.439	0.384
$n = 2000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.475	0.125	1.489	0.092
$\Delta = 0.1$	$b_0 = 0.1$	$\hat{b}$	0.111	0.041	0.109	0.049
$n\Delta = 200$	$c_0 = \sqrt{2}$	$\hat{c}$	1.450	0.289	1.424	0.273
$n = 500$	$\beta_0 = 1.5$	$\hat{\beta}$	1.479	0.200	1.489	0.214
$\Delta = 1$	$b_0 = 1$	$\hat{b}$	1.386	0.820	1.113	0.611
$n\Delta = 500$	$c_0 = \sqrt{2}$	$\hat{c}$	1.502	0.446	1.471	0.409
$n = 1000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.482	0.130	1.487	0.117
$\Delta = 1$	$b_0 = 1$	$\hat{b}$	1.179	0.495	1.082	0.360
$n\Delta = 1000$	$c_0 = \sqrt{2}$	$\hat{c}$	1.464	0.303	1.421	0.304

TAB. 4.2 – Whittle and Whittle Tapered estimator for Model 2 and for $\beta_0 = 1.5$, $\alpha_0 = 1$, $c_0 = \sqrt{2}$, $b_0 = \alpha_0\Delta$ and for 150 simulations

b) Results for the empirical moments method

We use the next results for empirical moment estimators for mean-reverting hidden diffusions, obtained by Genon-Catalot *et al* (2000).

We assume that the process (V_t) satisfies (A0)-(A3). Then the functions of the observations (Z_i)

$$\hat{m_1} = \frac{1}{n}\sum_{i=1}^{n} Z_i^2, \qquad \hat{m_2} = \frac{1}{3n}\sum_{i=1}^{n} Z_i^4, \qquad \hat{m_{1,2}} = \frac{1}{n}\sum_{i=1}^{n-1} Z_i^2 Z_{i-1}^2 \tag{4.30}$$

are consistent estimators of $\beta, E(\bar{V_1^2})$ and $E(\bar{V_1}\bar{V_2})$ respectively. Moreover these estimators converge in distribution asymptotically normally.

Inverting the above formulae, leads to consistent estimators of α, β and c.

The results are in Tables (4.3) and (4.4).

For the parameter β, the moment method estimate and the standard deviations are better than for the Whittle method.

According to the result given in (4.27), we can compute explicitly the asymptotic variance for the parameter β. For example, for Model 1, we obtain a standard deviation for $\Delta = 0.1$ and $n = 1000$ equal to 0.1581. For $\Delta = 0.1$ and $n = 2000$, it is equal to 0.1118 and for $\Delta = 1$ and $n = 500$, it is equal to 0.1494. So in both cases results are similar to the empirical asymptotic variance computed.

On the other hand, for the parameters b and c, the Whittle Tapered method provides better results than the moment method (even if the results are not completely satisfactory). c is better estimated when $\Delta = 1$. It is the opposite for b.
Moreover there are numerical problems when inverting the formula given in (4.30). For approximately a third of paths, the obtained estimates for b is negative.

This results comparison leads us to propose a two steps statistical method.

n, Δ	True parameters	Parameters	Empirical moments estimator	
			Mean	Standard deviation
$n = 500$	$\beta_0 = 2$	$\hat{\beta}$	2.019	0.232
$\Delta = 0.1$	$b_0 = 0.3$	$\hat{b}$	0.806	1.186
$n\Delta = 50$	$c_0 = \sqrt{2}$	$\hat{c}$	2.859	1.814
$n = 1000$	$\beta_0 = 2$	$\hat{\beta}$	1.989	0.161
$\Delta = 0.1$	$b_0 = 0.3$	$\hat{b}$	0.733	0.999
$n\Delta = 100$	$c_0 = \sqrt{2}$	$\hat{c}$	2.541	1.460
$n = 2000$	$\beta_0 = 2$	$\hat{\beta}$	1.990	0.110
$\Delta = 0.1$	$b_0 = 0.3$	$\hat{b}$	0.481	0.737
$n\Delta = 200$	$c_0 = \sqrt{2}$	$\hat{c}$	2.103	1.328
$n = 500$	$\beta_0 = 2$	$\hat{\beta}$	1.995	0.140
$\Delta = 1$	$b_0 = 3$	$\hat{b}$	1.340	2.673
$n\Delta = 500$	$c_0 = \sqrt{2}$	$\hat{c}$	1.263	1.080
$n = 1000$	$\beta_0 = 2$	$\hat{\beta}$	1.994	0.105
$\Delta = 1$	$b_0 = 3$	$\hat{b}$	1.320	1.902
$n\Delta = 1000$	$c_0 = \sqrt{2}$	$\hat{c}$	1.324	0.931

TAB. 4.3 – Empirical moments estimators for Model 1 and for $\beta_0 = 2$, $\alpha_0 = 3$, $c_0 = \sqrt{2}$, $b_0 = \alpha_0\Delta$

n, Δ	True parameters	Parameters	Empirical moments estimator	
			Mean	Standard deviation
$n = 500$	$\beta_0 = 1.5$	$\hat{\beta}$	1.462	0.263
$\Delta = 0.1$	$b_0 = 0.1$	$\hat{b}$	0.550	0.736
$n\Delta = 50$	$c_0 = \sqrt{2}$	$\hat{c}$	4.614	2.155
$n = 1000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.461	0.202
$\Delta = 0.1$	$b_0 = 0.1$	$\hat{b}$	0.372	0.678
$n\Delta = 100$	$c_0 = \sqrt{2}$	$\hat{c}$	3.755	2.048
$n = 2000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.494	0.147
$\Delta = 0.1$	$b_0 = 0.1$	$\hat{b}$	0.281	0.365
$n\Delta = 200$	$c_0 = \sqrt{2}$	$\hat{c}$	3.175	1.548
$n = 500$	$\beta_0 = 1.5$	$\hat{\beta}$	1.488	0.126
$\Delta = 1$	$b_0 = 1$	$\hat{b}$	1.534	1.887
$n\Delta = 500$	$c_0 = \sqrt{2}$	$\hat{c}$	1.886	1.699
$n = 1000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.488	0.092
$\Delta = 1$	$b_0 = 1$	$\hat{b}$	1.481	1.600
$n\Delta = 1000$	$c_0 = \sqrt{2}$	$\hat{c}$	1.821	1.365

TAB. 4.4 – Empirical moments estimators for Model 2 and for $\beta_0 = 1.5$, $\alpha_0 = 1$, $c_0 = \sqrt{2}$, $b_0 = \alpha_0\Delta$

c) Results for the two step method

The results are in Table 4.5 for Model 1 and in Table 4.6 for Model 2.

In a whole the parameter β is better estimated with the empirical moment method than with the Whittle method. Also the standard deviations are smaller with empirical moment method.

- For $\Delta = 0.1$: For model 1, the results are better for the parameter c while they are similar for the parameter b. For Model 2, it is the parameter b which is better estimated than with the Whittle method. For the parameter c it is slightly better for $n = 2000$ but slightly less good for $n = 1000$.

- For $\Delta = 1$: for Model 1 the estimates of c are better than with the Whittle method. However for the estimate of b and the empirical associated variance, it is less good. For Model 2, the results are almost similar.

So, in the whole, the results with the two steps Tapered Whittle method are better than those obtained with the other studied methods. However they are not totally satisfying depending of the models and the Δ value for b and c parameters. In particular, it seems that when Δ is small it is better and the contrary when Δ is "large".
In all cases, the two step Tapered Whittle method permits a time saving for computation of estimators, even if the obtained estimates are not always better.

n, Δ	True parameters	Parameters	Mixed Whittle estimator		Mixed Tapered Whittle estimator	
			Mean	Standard deviation	Mean	Standard deviation
$n = 1000$	$\beta_0 = 2$	$\hat{\beta}$	1.989	0.161	1.989	0.161
$\Delta = 0.1$	$b_0 = 0.3$	$\hat{b}$	0.570	0.678	0.438	0.297
$n\Delta = 100$	$c_0 = \sqrt{2}$	$\hat{c}$	1.675	0.714	1.610	0.638
$n = 2000$	$\beta_0 = 2$	$\hat{\beta}$	1.990	0.110	1.990	0.110
$\Delta = 0.1$	$b_0 = 0.3$	$\hat{b}$	0.471	0.225	0.431	0.195
$n\Delta = 200$	$c_0 = \sqrt{2}$	$\hat{c}$	1.694	0.413	1.545	0.413
$n = 500$	$\beta_0 = 2$	$\hat{\beta}$	1.995	0.140	1.995	0.140
$\Delta = 1$	$b_0 = 3$	$\hat{b}$	3.555	1.300	3.3213	0.770
$n\Delta = 500$	$c_0 = \sqrt{2}$	$\hat{c}$	1.392	0.231	1.411	0.222
$n = 1000$	$\beta_0 = 2$	$\hat{\beta}$	1.994	0.105	1.994	0.105
$\Delta = 1$	$b_0 = 3$	$\hat{b}$	3.217	0.600	3.138	0.554
$n\Delta = 1000$	$c_0 = \sqrt{2}$	$\hat{c}$	1.437	0.218	1.422	0.199

TAB. 4.5 – Mixed estimators for Model 1, $\beta_0 = 2$, $\alpha_0 = 3$ and $c_0 = \sqrt{2}$ and 150 simulations

n, Δ	True parameters	Parameters	Mixed Whittle estimator		Mixed Tapered Whittle estimator	
			Mean	Standard deviation	Mean	Standard deviation
$n = 1000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.491	0.202	1.491	0.202
$\Delta = 0.1$	$b_0 = 0.1$	$\hat{b}$	0.113	0.053	0.108	0.058
$n\Delta = 100$	$c_0 = \sqrt{2}$	$\hat{c}$	1.346	0.397	1.396	0.432
$n = 2000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.494	0.147	1.494	0.147
$\Delta = 0.1$	$b_0 = 0.1$	$\hat{b}$	0.109	0.041	0.108	0.048
$n\Delta = 200$	$c_0 = \sqrt{2}$	$\hat{c}$	1.416	0.315	1.413	0.387
$n = 500$	$\beta_0 = 1.5$	$\hat{\beta}$	1.488	0.126	1.488	0.126
$\Delta = 1$	$b_0 = 1$	$\hat{b}$	1.158	0.497	1.143	0.525
$n\Delta = 500$	$c_0 = \sqrt{2}$	$\hat{c}$	1.317	0.340	1.361	0.447
$n = 1000$	$\beta_0 = 1.5$	$\hat{\beta}$	1.488	0.092	1.488	0.092
$\Delta = 1$	$b_0 = 1$	$\hat{b}$	1.179	0.528	1.126	0.473
$n\Delta = 1000$	$c_0 = \sqrt{2}$	$\hat{c}$	1.462	0.357	1.412	0.414

TAB. 4.6 – Mixed estimators for Model 2, $\beta_0 = 1.5$, $\alpha_0 = 1$ and $c_0 = \sqrt{2}$ and 150 simulations

4.6 Conclusion

In this chapter, we were interested for mean-reverting hidden diffusions in computing and comparing different methods :
- Whittle and Tapered Whittle methods,
- empirical moment method
- and a two step statistical method based on Whittle contrast where the expectation of the process is estimated using an empirical moment method, and the other parameters are estimated using the Whittle or Tapered Whittle method.

For all these estimators, simulations have been made to compare their performances. In a whole the two step Tapered Whittle method is the more satisfying. Even if the obtained estimates are not always better than the other methods, it allows a time saving for computation of estimators. The parameter β is well estimated. However for the α and c parameters, the results are not always satisfying. This work shows the necessity to have large sample sizes for the observations (more than 2000, an order of 10 000) which is often possible with financial data.

4.7 Appendix

We recall the needed lemmas from Dahlhaus (1988), necessary to proofs of Theorems 4.1 in section 4.4.

Let $(X_i)_{0\leq i\leq n-1}$ be an ARMA process satisfying assumptions (H1)-(H5) and assume, in the taper case, that the tapered function h is the Tukey-Hanning taper described in paragraph 4.3.3.

Theorem A1.2 :
Let $\phi^1, ...\phi^l$ be bounded functions. Then $\{\sqrt{n}(J_n(\phi^i) - J(\phi^i)\}_{i=1,...,l}$ tends weakly to a Gaussian random vector ξ with mean 0 and covariance

$$\mathrm{Cov}(\xi_i, \xi_j) = 2\pi \left\{ \int_{\Pi} \phi^i(\alpha)\overline{(\phi^j(\alpha)} + \overline{\phi^j(-\alpha))}\, d\alpha \right.$$
$$\left. + \int_{\Pi^2} \phi^i(\alpha_1)\overline{\phi^j(-\alpha_2)} \frac{f_4(\alpha_1, -\alpha_1, \alpha_2)}{f(\alpha_1)f(\alpha_2)}\, d\alpha \right\}$$

Lemma A1.3 :
Suppose that ϕ is a bounded function. Then
(a) $\lim\limits_{n\to+\infty} EJ_n(\phi) - J(\phi) = 0.$
(b) $\lim\limits_{n\to+\infty} \sqrt{n}E[J_n(\phi) - J(\phi)] = 0.$

Lemma A1.4 :
Suppose that ϕ, ϕ^1 and ϕ^2 are bounded functions. Then
(a) $\lim\limits_{n\to+\infty} \mathrm{Var}\{J_n(\phi)\} = 0.$
(b) $\lim\limits_{n\to+\infty} n\,\mathrm{Cov}\{J_n(\phi^1), J_n(\phi^2)\} = \mathrm{Cov}(\xi_1, \xi_2)$

Chapitre 5

Estimation for mean-reverting stochastic volatility models with a leverage effect

Sommaire

5.1 Introduction

Stochastic volatility models are often used to study financial time series. Indeed they permit to assume that the conditional variance of the observed process is itself a stochastic process. Hence they have an important role in statistics (see e.g. Ghysels *et al* (1996) for a survey).
However these models can be too restrictive to model financial data. Then it is possible to incorporate the so-called leverage effect and to obtain an asymmetric stochastic volatility model. This model was first studied by Black (1976).

In this document, we are interested in the following form of the "leverage model" (LSV), as in Barndorff-Nielsen et Shepard (2001) :

$$\begin{cases} dY_t = \sigma_t dB_t + \rho dW_t, \quad Y_0 = 0 \\ dV_t = b(\theta, V_t)dt + a(\theta, V_t)dW_t \\ V_t = \sigma_t^2, \quad V_0 = \eta \end{cases} \tag{5.1}$$

where (B_t, W_t) is a standard Brownian motion of $\mathbb{R}^2$, η a variable independent of $(B_t, W_t)_{t\geq 0}$, θ an unknown parameter and ρ the leverage effect. In finance, it is assumed negative. (When ρ is equal to 0 we have the classical stochastic volatility SV model)
This effect means, in finance, that a positive increment in the volatility process will have a negative effect on the stock price.

A discrete sampling of the integrated process (Y_t) at regularly spaced times
$Z_i = \frac{1}{\sqrt{\Delta}} \int_{(i-1)\Delta}^{i\Delta} \sigma_s \, dB_s$ is observed (with Δ fixed).

As for classical (SV) models a difficulty in LSV is that the exact likelihood function is not easily computed. Moreover the model structure doesn't permit to easily extend statistical results obtained in SV models to estimate unknown parameters. For example : Generalized methods of moments (Genon-Catalot *et al* 2000) ; contrast based on conditional likelihood (Genon-Catalot *et al*, 2003), prediction-based estimating functions (Sørensen, 2000). In fact, these methods are based on computations of moments of order one and more. And few are explicit, except for some special cases : for example integrated Ornstein-Uhlenbeck Levy processes (Barndorff-Nielsen and Shepard (2001)).

To facilitate estimation, we assume that the non-observed process is a discretized diffusion (V_i) instead of a continuous diffusion (V_t). Hence the observations can be written $Z_i = F(V_i, r_i)$ where (r_i) is a two-dimensional sequence of i.i.d. standard Gaussian vectors.
To our knowledge, few works are done on the leverage models. In a particular case, for log-normal SV models, i.e. $\ln \sigma_i^2 = a + b \ln \sigma_{i-1}^2 + c\varepsilon_i$, Jacquier *et al* (2004) use Bayesian Markov chain Monte-Carlo methods. It is to note that they don't consider the same formulation to model the leverage effect : they consider that the two Brownian motions are correlated.

This document is organized as following : first we describe the studied model, a discretized version of (5.1) for the mean-reverting model, i.e. the function b having the following shape $b(\theta, V_t) = \alpha(\beta - V_t)$.
We study the probabilistic properties of the solution process , in particular it is a strictly stationary process and α-mixing with exponentially decreasing coefficients.
Second, the model being stationary and easy to simulate, the statistical chosen approach is to use an indirect inference method, introduced by Gourieroux *et al* (1993) and Gallant and Tauchen (1996), to estimate parameters of the model.
We describe this method and according to the particular structure of a mean-reverting model, we propose an auxiliary model for which the likelihood function is computable contrary to the initial model. Moreover we show that assumptions for consistency and asymptotic convergence in distribution of estimators are satisfied. Finally simulations have been made to evaluate the performance of these estimations.

5.2 Probabilistic properties for the studied model

5.2.1 Model and assumptions

We consider the simplified discrete-time model associated to (5.1), defined by, for $i \geq 1$,

$$\begin{cases} Z_i = \sigma_i \eta_i + \rho \varepsilon_i \\ V_i = \alpha\beta\Delta + (1 - \alpha\Delta)V_{i-1} + \sqrt{\Delta} a(V_{i-1})\varepsilon_i, \quad V_0 = \eta_0 \\ V_i = \sigma_i^2 \end{cases} \tag{5.2}$$

with the assumption (**A0**) : (η_i) and (ε_i) are independent and identically distributed standard Gaussian random variables, V_0 is independent of those.
α, β and Δ are positive ; ρ is real. And we note $\gamma = \alpha\beta\Delta$.

The model (5.2) is based on an Euler approximation associated to the mean-reverting model (5.1).

Only the process (Z_i) is observed. The Markov Chain (V_i) is called the volatility. The parameter ρ, assumed negative in finance, is called the leverage effect. This corresponds to a negative correlation between volatility and asset price in economic models and implies that a positive increment in the volatility process (V_i) has a negative effect on the stock price for example.

We make now assumptions necessary to ensure the existence of a strictly stationary Markov chain, solution of (5.2) with some specific probabilistic properties.

(**A1**) : The Markov Chain (V_i) has $I =]0, +\infty[$ as state space.

(**A2**) : a is a real strictly positive and continuous function on I.
It may depend on unknown parameters.

(**A3**) : there exists $k > 0$ such that $a^2(x) \leq kx^2 + 1$

(**A4**) : For k defined in (A3), parameters satisfy : $(1 - \alpha\Delta)^2 + k\Delta < 1$

5.2.2 Probabilistic properties

Lemma 5.1 *Assuming (A0)-(A4), then there exists an unique Markov chain , (V_n), solution of (5.2). Moreover (V_n) admits a stationary measure with moment of order two ; (V_n) is ergodic and α-mixing with exponentially decreasing coefficients.*

Proof :
To begin, there exists an unique strictly stationary solution (V_n) with moment of order two. To show this, we follow the reasoning of Mean and Tweedie (1993).
For $x \in I$, let $P(x, dy)$ the transition probability of the (V_n) Markov chain. In our case, we have

$$P(x, dy) = \frac{1}{\sqrt{2\pi\Delta a^2(x)}} \exp\left(-\frac{[y - (\gamma + (1 - \alpha\Delta)x)]^2}{2\Delta a^2(x)}\right) dy$$

First, it is easy to see that (V_n) is an irreducible Feller Chain using assumption (A2) and the Dominated Convergence Theorem. (If h is a bounded and continuous function on I then $Ph(x) = \int_{\mathbb{R}} h(y)P(x, dy)$ is continuous on I.)

Second, we define the function g by $g(x) = x^2 + 1$. We have

$$E[V_i^2 | V_{i-1} = x] = \Delta a^2(x) + [\gamma + (1 - \alpha\Delta)x)]^2$$

since the distribution of $(V_i | V_{i-1} = x)$ is a $\mathcal{N}(\gamma + (1 - \alpha\Delta)x), \Delta a^2(x))$.
Using Assumption (A3) we deduce that

$$E[V_i^2 | V_{i-1} = x] \leq (k\Delta + (1 - \alpha\Delta)^2)x^2 + 2\gamma(1 - \alpha\Delta)x + L$$

where L is a finite constant depending of parameters but not depending of x.

Let $\lambda = \frac{1}{2}(1 - (k\Delta + (1 - \alpha\Delta)^2)$. Under Assumption (A4), we have $\lambda > 0$. Hence there exists a measurable set C and a constant b finite (Meyn and Tweedie, 1993) such that $Pg(x) - g(x) \leq -\lambda g(x) + b\mathbf{1}_C(x)$.

Hence (V_n) is g-geometric ergodic and the equation (5.2) admits an unique strictly stationary solution with a moment of order two.

Finally, as (V_n) is irreducible, aperiodic and g-geometric ergodic (with $g(x) = x^2+1$), (V_n) is α-mixing with exponentially decreasing coefficients (Meyn and Tweedie, 1993, p386 and p388).

Proposition 5.1 *Under the assumptions (A0)-(A4), the process (Z_i) is strictly stationary with moment of order two, ergodic and α-mixing with exponentially decreasing coefficients.*

Proof :
Let $U_i = (V_{i-1}, V_i)$ for $i \geq 1$.

According to Lemma (5.1), (U_n) is a strictly stationary ergodic Markov chain with state space I^2. Moreover (U_n) is α-mixing with exponentially decreasing coefficients. (The same reasoning, as Lemma (5.1), hold with the function $j(x_1, x_2) = x_2^2 + 1$.)

Under assumption (A2), we have $a(x) > 0$ on I. Hence we can write

$$Z_i = G(U_i, \eta_i) = \sqrt{V_i}\eta_i + \rho f(U_i)$$

where (η_n) is a sequence of i.i.d. random variables, independent of the Markov chain (U_n) and f a real function on I^2. Hence (Z_n) is a hidden Markov chain with (U_n) as hidden Markov chain.
Indeed, conditionally on $\mathcal{G}_n = \sigma(V_0, ..., V_n) = \sigma(U_1, ..., U_n)$, the random variables (Z_j) are independent and have a Gaussian distribution $\mathcal{N}(\rho f(U_i), V_i)$. So we have for $(\lambda_1, ..., \lambda_n) \in \mathbb{R}^n$

$$E\left[\exp(\sum_{j=1}^{n} i\lambda_j Z_j)|\mathcal{G}_n\right] = \exp\left(i\rho\sum_{j=1}^{n}\lambda_j f(U_j)\right)\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\lambda_j^2 V_j\right)$$

We deduce from this equality that (Z_n) is a hidden Markov chain with (U_n) as hidden Markov chain.

Therefore the process (Z_n) has the same properties that the process (U_n) (Proposition 3.1 of Genon-Catalot *et al*, 2000). ◇

A first possibility would be to use these stochastic properties as Genon-Catalot *et al*, 2000) and to apply an empirical moments method. However the computations of expectations terms, such as $E(Z_i^4)$ or $E(Z_i^2 Z_{i+1}^2)$, are not explicit.
Hence we choose to exploit the GARCH structure of mean-reverting stochastic volatility models without a leverage effect. (cf previous Chapter 4)

5.3 The indirect inference method

The aim is to estimate the parameter vector $\theta = (\alpha, \beta, c, \rho)$ with $\theta \in \Theta \subset \mathbb{R}^4$ where we assume that c is an unknown one-dimensional parameter in the function a.

A classical method is to maximize the conditional likelihood function. However computation of this function is difficult in this case. Indeed, it requires computation of the integral of the conditional density function of $(Z_1, ..., Z_n)$ given v_0 which is a n-dimensional integral not directly computable. So we will not study the maximum likelihood estimator and its asymptotic properties in this document.

However we first remark that the mean-reverting model is strictly stationary and has a particularly structure, to be more precise the beginning of (V_i) is the beginning of the variance term of a GARCH process.
Second it is possible to simulate values of $\mathbf{Z_n} = (Z_1, ..., Z_n)$ for a given initial condition (z_0, v_0) and a given value of the parameter vector θ.
This is why we will apply an indirect inference method to estimate parameters, described in this paragraph (Gourieroux *et al*, 1993 ; Gallant and Tauchen, 1996). The

obtained indirect estimators are consistent and converge asymptotically in distribution.

The first step : choice of an auxiliary criterion.

The first step is the choice of an auxiliary criterion $Q_n(\mathbf{Z_n}, \omega)$ with an auxiliary parameter vector $\omega \in \Omega \subset \mathbb{R}^p$. An estimator of ω is obtained by maximizing the previous criterion :

$$\hat{\omega}_n = \arg\max_{\omega \in \Omega} Q_n(\mathbf{Z}_n, \omega) \tag{5.3}$$

It is assumed that

(H1) Q_n converges almost surely as $n \to +\infty$ to a deterministic limit which depends on the unknown parameter of the data θ and called $Q_\infty(\theta, \omega)$.

(H2) Q_∞ is continuous in ω and admits an unique maximum in ω_0.

Hence under assumptions (H1)-(H2), $\hat{\omega}_n$ is a consistent estimator of ω_0 and $(\hat{\omega}_n)$ convergences in distribution to rate $\sqrt{n}$ (Gourieroux *et al*, 1993).

Let introduce the binding function $b : \mathbb{R}^4 \to \mathbb{R}^p$ defined by $b(\theta) = \arg\max_{\omega \in \Omega} Q_\infty(\theta, \omega)$.
We have $\omega_0 = b(\theta_0)$.
We can notice that if the function b was known and one to one, a consistent estimator of θ_0 could be the solution $\tilde{\theta}_n$ of $\hat{\omega}_n = b(\tilde{\theta}_n)$. The idea of the second step is then to replace the unknown function b by an estimate based on simulations of the observed process (Z_i).

It is assumed that
(H3) The function b is one to one. Moreover b is derivable and the matrix $\dfrac{\partial b}{\partial \theta'}(\theta_0)$ is of full column rank.

The second step :

First, for a given value of θ and an initial condition v_0, we can simulate a path $\tilde{\mathbf{z}}_{nH}(\theta) = \{z_i(\theta), i = 1, ..., nH\}$ of the model (5.2), with $H \geq 1$.
Then the parameter ω is estimated by replacing the observed process with this simulation in equation (5.3) :

$$\tilde{\omega}_{nH}(\theta) = \arg\max_{\omega \in \Omega} Q_n(\tilde{\mathbf{z}}_{nH}(\theta), \omega) \tag{5.4}$$

Hence $\tilde{\omega}_{nH}(\theta)$ is a consistent functional estimator of $b(\theta)$ since we have
$\lim_{n \to +\infty} \tilde{\omega}_{nH}(\theta) = b(\theta) \quad a.s.$

The idea is then to obtain a value of θ such that $\tilde{\omega}_{nH}(\theta)$ is close to $\hat{\omega}_n$.
The indirect inference estimator of θ is then defined by :

$$\tilde{\theta}_{nH} = \arg\min_{\theta \in \Theta} [(\hat{\omega}_n - \tilde{\omega}_{nH}(\theta))' M_n (\hat{\omega}_n - \tilde{\omega}_{nH}(\theta))]$$

where M_n is a random positive definite matrix converging to a deterministic positive definite matrix M.

In the following, we assume that

(H4) The function Q_n is derivable compared to ω.

As the gradient of Q_n can be explicitly computable (by iteration), we work with the indirect estimator proposed by Gallant and Tauchen (1996) which is asymptotically equivalent to this proposed by Gourieroux *et al* (1993). It permits to have only one step of optimization in θ contrary to the initial indirect estimator (with two optimizations). It is based on the score function of the auxiliary model, and the value of θ is chosen such that this score function is minimal (as close as possible to 0).
Hence the second indirect estimator is defined by

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{\partial Q_{nH}}{\partial\omega'}(\tilde{\mathbf{z}}_{nH}(\theta),\hat{\omega}_n)\Sigma_n\frac{\partial Q_{nH}}{\partial\omega}(\tilde{\mathbf{z}}_{nH}(\theta),\hat{\omega}_n)$$

where Σ_n is a random positive definite matrix converging to a deterministic positive definite matrix Σ.
Under the assumptions (H1)-(H4), $(\tilde{\theta}_{nH})$ and $(\hat{\theta}_n)$ are consistent estimator of θ_0 as n tends to $+\infty$.

Asymptotic normally distribution

To begin, we recall the necessary assumptions done by Gourieroux *et al* (1993) to ensure the convergence in distribution of the estimator $(\hat{\theta}_n)$ to a normal distribution. Moreover we precise some additional assumptions not done by Gourieroux *et al* (1993) but used for the proof of results.

We assume that the auxiliary criterion has the shape $Q_n(\mathbf{Z_n},\omega) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} q_i(Z_i,\omega)$.

And we define $Q_n^h(\tilde{\mathbf{z}}_n^h(\theta),\omega) = \dfrac{1}{n}\displaystyle\sum_{i=nh}^{n(h+1)-1} q_i(\tilde{z}_i(\theta),\omega)$.

(H5)

$$\xi_n = \sqrt{n}\frac{\partial Q_n}{\partial\omega}(\mathbf{z_n},\omega_0) - \frac{\sqrt{n}}{H}\sum_{h=1}^{H}\frac{\partial Q_n^h}{\partial\omega}(\tilde{\mathbf{z}}_{\mathbf{n}}^{\mathbf{h}}(\theta_{\mathbf{0}}),\omega_0)$$

is asymptotically normal with zero mean and finite asymptotic covariance matrix $W = \lim_{n\to+\infty}\mathrm{Var}_{\theta_0}(\xi_n)$.

(H6) Q_n is two times derivable compared to ω and the following limits exist :

$$I_0 = \lim_{n\to+\infty}\operatorname*{Var}_{\theta_0}\left[\sqrt{n}\frac{\partial Q_n}{\partial\omega}(\tilde{z}_n(\theta_0),\beta_0)\right]$$

$$J_0 = \lim_{n\to+\infty}\left[-\frac{\partial^2 Q_n}{\partial\omega\partial\omega'}(\tilde{z}_n(\theta_0),\beta_0)\right] = -\frac{\partial^2 Q_\infty}{\partial\omega\partial\omega'}(\theta_0,\beta_0) \quad a.s.$$

Moreover I_0 and J_0 are assumed non-singular.

Finally it is assumed that

(H7) The limits of the following matrix (of dimension $4 \times p$ and $p \times 4$) exist

$$\lim_{n\to+\infty}\left[\frac{\partial^2 Q_n}{\partial\theta\partial\omega'}(\theta,\omega)\right] = \frac{\partial^2 Q_\infty}{\partial\theta\partial\omega'}(\theta,\omega)$$

$$\lim_{n\to+\infty}\left[\frac{\partial^2 Q_n}{\partial\omega\partial\theta'}(\theta,\omega)\right] = \frac{\partial^2 Q_\infty}{\partial\omega\partial\theta'}(\theta,\omega)$$

Under these regularity conditions (H1)-(H7), the indirect estimator is asymptotically normal , when H is fixed and n tends to infinity :

$$\sqrt{n}(\tilde{\theta}_{nH} - \theta_0) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}(0, W(H,M))$$

where $W(H,M) = \left(1+\frac{1}{H}\right) l(\theta_0,M)^{-1}\frac{\partial b'}{\partial\theta}(\theta_0) M J_0^{-1} I_0 J_0^{-1} M \frac{\partial b}{\partial\theta'}(\theta_0) l(\theta_0,M)^{-1}$ and $l(\theta_0,M) = \frac{\partial b'}{\partial\theta}(\theta_0) M \frac{\partial b}{\partial\theta'}(\theta_0)$

Then the optimal choice of M is obtained for $M^* = J_0 I_0^{-1} J_0$. And in this case

$$W^*(H,M) = \left(1+\frac{1}{H}\right)\left(\frac{\partial^2 Q_\infty}{\partial\theta\partial\omega'}(\theta_0,\omega_0) I_0^{-1} \frac{\partial^2 Q_\infty}{\partial\omega\partial\theta'}(\theta_0,\omega_0)\right)^{-1}$$

The second indirect estimator of θ, $\hat{\theta}_n$ is optimal and asymptotically equivalent to the first when $\Sigma^* = J_0^{-1} M^* J_0^{-1} = I_0^{-1}$.

The indirect inference method described above is a general method. In the following section, we propose an auxiliary criterion for mean-reverting discrete-time models and we check that assumptions are satisfied to have asymptotic properties.

5.4 The proposed auxiliary criterion

Intuition leads to choose an auxiliary model having some similarities with the true model so that the indirect estimator is good. Moreover, the estimation method has two steps of numerical minimization. So it is preferable to have a quite simple shape for the auxiliary model.

5.4.1 The criterion

For the mean-reverting model defined in (5.2),when ρ is equal to 0, Z_i and the beginning of the (V_i) process are the beginning of a GARCH process.
In particular, when $a(x) = cx$, the process (Z_t) solution of the stochastic differential equation :

$$dY_t = \sigma_t dB_t + \rho dW_t, \quad Y_0 = 0$$

$$dV_t = \alpha(\beta - V_t)dt + a(\theta, V_t)dW_t, \quad V_t = \sigma_t^2, \quad V_0 = \eta$$

can be seen as the diffusion approximation of a GARCH(1,1) model (Nelson, 1990).

As θ is four-dimensional, the idea is then to use a GARCH(2, 1) process as auxiliary model, i.e.

$$\begin{cases} y_i = \sqrt{h_i}\nu_i \\ h_i = \omega_0 + \omega_2 h_{i-1} + \omega_3 h_{i-2} + \omega_1 y_{i-1}^2 \end{cases} \tag{5.5}$$

with (ν_i) i.i.d. $\mathcal{N}(0,1)$; $\omega_0 > 0$ and $0 \leq \omega_1 + \omega_2 + \omega_3 < 1$ (to ensure the existence of a solution with moment of order two).

Remark : We also tested the choice of a GARCH(1, 2) as auxiliary model. However, the obtained results on simulations are not as good as the results with the GARCH(2, 1) case.

Hence for the auxiliary criterion, we take the conditional log-likelihood function associated to a GARCH(2,1) process

$$Q_n(\mathbf{y_n}, \omega) = \frac{1}{n-1} \sum_{i=2}^{n} \left(\ln(h_i) + \frac{y_i^2}{h_i} \right) \tag{5.6}$$

5.4.2 Properties

Since the process (Z_n) is strictly stationary, we can consider its extension to a process indexed by $\mathbb{Z}$, with the same finite-dimensional distributions (as Elie and Jeantheau, 1995). For all $i \in \mathbb{Z}$, we define the infinite past from i by $\underline{Z}_i = (Z_i, Z_{i-1}, ...)$, which is a vector of $\mathbb{R}^{\mathbb{N}}$.
Moreover, in the following, we note the process $(Z_i) = (Z_i(\theta))$.

Proposition 5.2 *Assuming (A0)-(A4), when n tends to infinity, we have that* $Q_n(\mathbf{Z_n}(\theta), \omega)$ *converges almost surely, under* $\mathbb{P}_{\theta_0}$*, to*

$$Q_\infty(\theta_0, \omega) = E_{\theta_0} \left[\ln(h_2(\omega, \underline{Z}_1(\theta)) + \frac{Z_2^2(\theta)}{h_2(\omega, \underline{Z}_1(\theta))} \right]$$

Proof :
First it is known that we can write (see Fan and Yao 2003, for example)

$$h_i = h_i(\omega, \underline{y}_{i-1}) = \frac{\omega_0}{1 - (\omega_2 + \omega_3)} + \sum_{j=1}^{\infty} d_j y_{i-j}^2$$

with $d_j \geq 0$ for all j.

In particular, $h_2 = h_2(\omega, \underline{y}_1) = \dfrac{\omega_0}{1 - (\omega_2 + \omega_3)} + \displaystyle\sum_{j=1}^{\infty} d_j y_{2-j}^2$

Then the auxiliary criterion can be written

$$Q_n(\mathbf{Z_n}(\theta), \omega) = \frac{1}{n-1} \sum_{i=2}^{n} \left(\ln(h_i(\omega, \underline{Z}_{i-1}(\theta)) + \frac{Z_i^2(\theta)}{h_i(\omega, \underline{Z}_{i-1}(\theta))} \right)$$

According to Proposition 5.1, the process $(Z_i(\theta))$ is strictly stationary and ergodic. Hence, the ergodic theorem gives that $Q_n(\mathbf{Z_n}(\theta),\omega)$ converges almost surely, under $\mathbb{P}_{\theta_0}$, to $Q_\infty(\theta_0,\omega) = E_{\theta_0}\left[\ln(h_2(\omega,\underline{Z}_1(\theta)) + \frac{Z_2^2(\theta)}{h_2(\omega,\underline{Z}_1(\theta))}\right]$, as n tends to infinity. $\diamond$

We note $q_i(Z_i,\underline{Z}_{i-1}) = \ln(h_i(\omega,\underline{Z}_{i-1}(\theta)) + \frac{Z_i^2(\theta)}{h_i(\omega,\underline{Z}_{i-1}(\theta))}$.

And we assume

(A5) There exists some positive δ such that $E(|q_2(Z_2,\underline{Z}_{1-1})|^{2+\delta}) < +\infty$.

Lemma 5.2 *Under assumptions (A0)-(A5), the assumptions (H5), (H6) and (H7) are satisfied.*

Proof :

We can write $Q_n(\mathbf{Z_n}(\theta),\omega) = \frac{1}{n-1}\sum_{i=2}^{n} q_i(Z_i,\underline{Z}_{i-1})$.

Under (A0)-(A4) the process (Z_i) is strictly stationary and α-mixing with exponentially decreasing coefficients. Then under (A0)-(A5), we can apply a central limit theorem (Hall and Heyde 1980, p.132) and we obtain assumption (H5).

For (H6) and (H7), it is the same. Under (A0)-(A5) we can apply the ergodic theorem for the existence of the matrix J_0 and those of (H7). And a central limit assure the existence of the matrix I_0.

Proposition 5.3 *Under assumptions (A0)-(A4) and (H1)-(H3), $(\tilde{\theta}_{nH})$ is a consistent estimator of θ_0 as n tends to $+\infty$.*

Proof :

We recall that $\tilde{\theta}_{nH} = \arg\min_{\theta\in\Theta}[(\hat{\omega}_n - \tilde{\omega}_{nH}(\theta))' M_n(\hat{\omega}_n - \tilde{\omega}_{nH}(\theta))]$ where M_n is a random positive definite matrix tending to a deterministic positive definite matrix M.

Under assumptions (A0)-(A4) and (H1)-(H3) and according to Proposition 5.2 and Gourieroux *et al* (1993), we have $\hat{\omega}_n$ converging under $\mathbb{P}_{\theta_0}$ to $\beta_0 = b(\theta_0)$ almost surely. Also, $\tilde{\omega}_{nH}(\theta)$ converges to $b(\theta)$ almost surely.

We define the function

$$U_n(\theta) = (\hat{\omega}_n - \tilde{\omega}_{nH}(\theta))' M_n(\hat{\omega}_n - \tilde{\omega}_{nH}(\theta))$$

As M_n converges to a deterministic matrix M, then $U_n(\theta)$ converges under $\mathbb{P}_{\theta_0}$ to the function $K(\theta_0,\theta) = (b(\theta_0) - b(\theta))' M (b(\theta_0) - b(\theta))$. As M is a positive definite matrix and b is a continuous function, the function K is continuous and positive. Moreover $K(\theta_0,\theta) = 0$ if and only if $b(\theta_0) = b(\theta)$. Under the assumption that b is one-to-one, it is the case if and only if $\theta = \theta_0$.
Then the consistency of the estimator $(\tilde{\theta}_{nH})$ follows (for example Dacunha-Castelle and Duflo 1993). $\diamond$

Remark : The estimator $(\hat{\theta}_n)$ is also a consistent estimator of θ_0 as n tends to $+\infty$ (see Gallant and Tauchen 1996).

Proposition 5.4 *Under assumptions (A0)-(A4) and (H1)-(H7),* $\sqrt{n}(\hat{\theta}_n-\theta_0)$ *converges in distribution to a* $\mathcal{N}(0,(1+\frac{1}{H})B^{-1}AI_0A'(B^{-1})')$ *as* n *tends to* $+\infty$*, with*

$$A=\frac{\partial^2 Q_\infty}{\partial\theta\partial\omega'}(\theta_0,\omega_0)\Sigma \quad and \quad B=\frac{\partial^2 Q_\infty}{\partial\theta\partial\omega'}(\theta_0,\omega_0)\Sigma\frac{\partial^2 Q_\infty}{\partial\omega\partial\theta'}(\theta_0,\omega_0)$$

with B *being assumed non singular.*

Proof : It is based on Gourieroux *et al* (1993).
We recall that $\hat{\theta}_n=\arg\max_{\theta\in\Theta}\frac{\partial Q_{nH}}{\partial\omega'}(\tilde{\mathbf{z}}_{nH}(\theta),\hat{\omega}_n)\Sigma_n\frac{\partial Q_{nH}}{\partial\omega}(\tilde{\mathbf{z}}_{nH}(\theta),\hat{\omega}_n)$
where Σ_n is a random positive definite matrix converging to a deterministic positive definite matrix Σ.

Moreover let

$$A_n=\frac{\partial^2 Q_{nH}}{\partial\theta\partial\omega'}(\tilde{\mathbf{z}}_{nH}(\theta_0),\omega_0)\Sigma_n$$

$$B_n=\frac{\partial^2 Q_{nH}}{\partial\theta\partial\omega'}(\tilde{\mathbf{z}}_{nH}(\theta_0),\omega_0)\Sigma_n\frac{\partial^2 Q_{nH}}{\partial\omega\partial\theta'}(\tilde{\mathbf{z}}_{nH}(\theta_0),\omega_0)$$

and

$$C_n=\frac{\partial^2 Q_{nH}}{\partial\omega\partial\omega'}(\tilde{\mathbf{z}}_{nH}(\theta_0),\omega_0)$$

Under the assumption (H7), the three matrix converge almost surely respectively to

$$A=\frac{\partial^2 Q_\infty}{\partial\theta\partial\omega'}(\theta_0,\omega_0)\Sigma,\quad B=\frac{\partial^2 Q_\infty}{\partial\theta\partial\omega'}(\theta_0,\omega_0)\Sigma\frac{\partial^2 Q_\infty}{\partial\omega\partial\theta'}(\theta_0,\omega_0) \text{ and } C=\frac{\partial^2 Q_\infty}{\partial\omega\partial\omega'}(\theta_0,\omega_0)$$

$\hat{\theta}_n$ being a minimum, we have

$$\frac{\partial^2 Q_{nH}}{\partial\theta\partial\omega'}(\tilde{\mathbf{z}}_{nH}(\hat{\theta}_n),\hat{\omega}_n)\Sigma_n\frac{\partial Q_{nH}}{\partial\omega}(\tilde{\mathbf{z}}_{nH}(\hat{\theta}_n),\hat{\omega}_n)=0$$

The Taylor formula at first order gives

$$\frac{\partial Q_{nH}}{\partial\omega}(\tilde{\mathbf{z}}_{nH}(\hat{\theta}_n),\hat{\omega}_n)=\frac{\partial Q_{nH}}{\partial\omega}(\tilde{\mathbf{z}}_{nH}(\theta_0),\omega_0)+\frac{\partial^2 Q_{nH}}{\partial\omega\partial\theta'}(\tilde{\mathbf{z}}_{nH}(\theta_0),\omega_0)(\hat{\theta}_n-\theta_0)+$$

$$\frac{\partial^2 Q_{nH}}{\partial\omega\partial\omega'}(\tilde{\mathbf{z}}_{nH}(\theta_0),\omega_0)(\hat{\omega}_n-\omega_0)+\varepsilon_n(x_n-x_0)$$

with $x=(\hat{\theta}_n,\hat{\omega}_n$ and $x_0=(\theta_0,\omega_0)$.
Moreover ε_n tends to 0 in probability when n tends to infinity.

To simplify notations we define $\frac{\partial Q_{nH}}{\partial\omega}(\theta_0,\omega_0)$ instead of $\frac{\partial Q_{nH}}{\partial\omega}(\tilde{\mathbf{z}}_{nH}(\theta_0),\omega_0)$.
Hence we obtain

$$\sqrt{n}(\hat{\theta}_n-\theta_0)=B_n^{-1}A_n\left[\sqrt{n}\frac{\partial Q_{nH}}{\partial\omega}(\theta_0,\omega_0)+\frac{\partial^2 Q_{nH}}{\partial\omega\partial\omega'}(\theta_0,\beta_0)\sqrt{n}(\hat{\omega}_n-\omega_0)+\varepsilon_n(x_n-x_0)\right]$$

On the other hand, we have

$$0 = \frac{\partial Q_{nH}}{\partial \omega}(\theta_0, \tilde{\omega}_{nH}) = \frac{\partial Q_{nH}}{\partial \omega}(\theta_0, \omega_0) + \frac{\partial^2 Q_{nH}}{\partial \omega \partial \omega'}(\theta_0, \omega_0)(\tilde{\omega}_{nH} - \omega_0) + \varepsilon_{n,2}(\tilde{\omega}_{nH} - \omega_0)$$

with $\varepsilon_{n,2}$ tending to 0 in probability when n tends to infinity.

Hence we can write

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = B_n^{-1} A_n C_n [\sqrt{n}(\hat{\omega}_n - \omega_0) - \sqrt{n}(\tilde{\omega}_{nH} - \omega_0) - \varepsilon_{n,3}\sqrt{n}(\tilde{\omega}_{nH} - \omega_0)]$$

with $\varepsilon_{n,3}$ tending to 0 in probability when n tends to infinity.

By definition $\hat{\omega}_n$ is a maximum of $Q_n(\mathbf{Z}_n, \omega)$ and $\tilde{\omega}_{nH}(\theta)$ is a maximum of $Q_{nH}(\tilde{\mathbf{z}}_{nH}(\theta), \omega)$. Then we have

$$0 = \frac{\partial Q_n}{\partial \omega}(\mathbf{Z}_n, \hat{\omega}_n) = \frac{\partial Q_n}{\partial \omega}(\mathbf{Z}_n, \omega_0) + \frac{\partial^2 Q_n}{\partial \omega \partial \omega'}(\mathbf{Z}_n, \omega_0)(\hat{\omega}_n - \omega_0) + \alpha_n(\hat{\omega}_n - \omega_0) \quad (5.7)$$

with α_n tends to 0 in probability when n tends to infinity. And

$$0 = \frac{\partial Q_{nH}}{\partial \omega}(\tilde{\mathbf{z}}_{nH}(\theta_0), \omega_0) + \frac{\partial^2 Q_{nH}}{\partial \omega \partial \omega'}(\theta_0, \omega_0)(\tilde{\omega}_{nH} - \omega_0) + \alpha_{n,2}(\tilde{\omega}_{nH} - \omega_0) \quad (5.8)$$

with $\alpha_{n,2}$ tending to 0 in probability when n tends to infinity.

We assumed that $Q_{nH}(\tilde{\mathbf{z}}_{nH}(\theta), \omega)$ could be written under the shape

$$Q_{nH}(\tilde{\mathbf{z}}_{nH}(\theta), \omega) = \frac{1}{H}\sum_{h=1}^{H} Q_n^h(\tilde{\mathbf{z}}_n^h(\theta), \omega) \text{ where } Q_n^h(\tilde{\mathbf{z}}_n^h(\theta), \omega) = \frac{1}{n}\sum_{i=nh}^{n(h+1)-1} q_i(\tilde{z}_i(\theta), \omega)$$

Then we obtain by subtracting the two terms (5.7) and (5.8) :

$$\sqrt{n}(\hat{\omega}_n - \omega_0) - \sqrt{n}(\tilde{\omega}_{nH} - \omega_0) = C_n^{-1}\left[\sqrt{n}\frac{\partial Q_n}{\partial \omega}(\mathbf{z_n}, \omega_0) - \frac{\sqrt{n}}{H}\sum_{h=1}^{H}\frac{\partial Q_n^h}{\partial \omega}(\tilde{\mathbf{z}}_n^h(\theta_0), \omega_0)\right] +$$

$$C_n^{-1}(\alpha_n(\hat{\omega}_n - \omega_0) - \alpha_{n,2}(\tilde{\omega}_{nH} - \omega_0))$$

Hence we have with ξ_n introduced in (H5) :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = B_n^{-1} A_n \xi_n + r_n$$

with $r_n = -B_n^{-1} A_n C_n \varepsilon_{n,3}\sqrt{n}(\tilde{\omega}_{nH} - \omega_0) + B_n^{-1} A_n(\alpha_n\sqrt{n}(\hat{\omega}_n - \omega_0) - \alpha_{n,2}\sqrt{n}(\tilde{\omega}_{nH} - \omega_0))$.

Under (A0)-(A5) the process (Z_i) is strictly stationary, α-mixing with exponentially decreasing coefficients. So we can apply a central limit theorem (Hall and Heyde 1980, p.132 for example).

The different terms in ξ_n are independent. Hence, as I_0 is non singular, we obtain that ξ_n converges in distribution to a Gaussian vector $\mathcal{N}(0, (1 + \frac{1}{H})I_0)$.

Moreover $B_n^{-1} A_n$ converges almost surely to $B^{-1}A$ under assumption (H7).

As $\sqrt{n}(\tilde{\omega}_{nH} - \omega_0)$ and $\sqrt{n}(\hat{\omega}_n - \omega_0)$ are supposed to converge in distribution, they are bounded in probability. Therefore the rest r_n tends to 0 in probability because $(\varepsilon_{n,j})_{j=1,2,3}$ and $(\alpha_{n,j})j = 1, 2$ tend in probability to zero.
Then we obtain that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a
$\mathcal{N}(0, (1 + \frac{1}{H})B^{-1}AI_0A'(B^{-1})')$.

Remarks :
- The assumptions (H2), (H3) and (A5) are checked numerically. Indeed, if they were not satisfied, then the indirect inference method would not give results for the study of estimators on simulations.

-If the a function depends from an unknown vector parameter of dimension s, a GARCH(p, q) can be again used as auxiliary model. However it is necessary that $p + q + 1 \geq s + 2$ to ensure the injectivity of the binding function.

5.5 Simulations to study the method performances

In this section we study the performances of the indirect estimator for three models and different values of Δ. The considered models are :

Model 1 : $a(x) = cx$ with $c > 0$. The parameters must satisfy $(1 - \alpha\Delta)^2 + c\Delta < 1$.
The observations (Z_i) are simulated with $\theta_0 = (\alpha_0, \beta_0, c_0, \rho_0) = (2, 5, 0.7, -1)$ and for $\Delta = 0.1$ and 0.05.

Model 2 : $a(x) = c\sqrt{x}$ with $c > 0$. The process (V_i) is then the square-root process used for interest rates by Cox *et al* (1985). The parameters must also satisfy $(1 - \alpha\Delta)^2 + c\Delta < 1$.
The observations (Z_i) are simulated with $\theta_0 = (\alpha_0, \beta_0, c_0, \rho_0) = (2.5, 4, 1, -1)$ and for $\Delta = 0.1$ and 0.05.

Model 3 : $a(x) = cx^{0.7}$ with $c > 0$. The parameters must satisfy $(1 - \alpha\Delta)^2 + c\Delta < 1$.
The observations (Z_i) are simulated with $\theta_0 = (\alpha_0, \beta_0, c_0, \rho_0) = (2.5, 4, 1, -1.5)$ and for $\Delta = 0.1$ and 0.05.

The indirect estimators are computed for 200 replications with $n = 1000, 2000$ and 3000. The number H which gives the length of the "simulated" observations $\tilde{\mathbf{z}}_{nH}$, is chosen equal to 20 when $n = 1000$ and 10 when $n = 2000, 3000$ (tests show that greatest values for H don't change much estimated values of parameters).
The results are in the Tables (5.1), (5.2) and (5.3). They contain mean and empirical standard deviation of the estimated parameters. Moreover the results are given for $b = \alpha\Delta$ instead of α.

First more n is large, better are the estimations and empirical variances.

We notice that in a whole the results are better for Model 2 than Model 1 as Whittle estimation in the previous chapter. For Model 1, a reason suggested by Genon-Catalot *et al* (1999) could be that the parameter involving the number of degrees of freedom of a Student distribution is badly estimated even for independent

observations (Blattberg and Gonedes (1974)). For Model 2, if the process (V_i) is observed from a discrete manner, the estimation of α and β are good using the method proposed by Kessler (2000) for example.

The results show that β is always well estimated and same for ρ in a whole. For Model 1, the parameter c is quite well estimated however the empirical standard deviation is high contrary to the model 2. The estimation of α is rather bad except when $n = 3000$. This corroborates the need of a lot of information for these models. For Model 3, β, c and ρ are well estimated. However the parameter α is difficult to estimate as Model 1 and the standard deviation is high.

Δ	Parameters	$n = 1000$		$n = 2000$		$n = 3000$	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
$\Delta = 0.1$	$\hat{b}$	0.296	0.201	0.256	0.141	0.235	0.085
	$\hat{\beta}$	4.978	0.075	4.982	0.068	4.994	0.038
	$\hat{c}$	0.669	0.165	0.694	0.125	0.707	0.066
	$\hat{\rho}$	-0.962	0.139	-0.990	0.111	-0.996	0.075
$\Delta = 0.05$	$\hat{b}$	0.221	0.197	0.167	0.119	0.127	0.048
	$\hat{\beta}$	5.008	0.081	5.007	0.070	4.996	0.047
	$\hat{c}$	0.803	0.286	0.732	0.212	0.701	0.054
	$\hat{\rho}$	-0.956	0.142	-0.986	0.107	-0.992	0.094

TAB. 5.1 – Indirect estimator for Model 1 with 200 paths. For $\Delta = 0.1$, $b_0 = 0.2$ and for $\Delta = 0.05$, $b_0 = 0.1$.

Δ	Parameters	$n = 1000$		$n = 2000$		$n = 3000$	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
$\Delta = 0.1$	$\hat{b}$	0.280	0.170	0.261	0.113	0.258	0.084
	$\hat{\beta}$	3.997	0.069	4.002	0.053	4.003	0.059
	$\hat{c}$	0.985	0.092	0.988	0.081	0.991	0.075
	$\hat{\rho}$	-0.995	0.094	-0.998	0.074	-0.997	0.069
$\Delta = 0.05$	$\hat{b}$	0.170	0.132	0.154	0.111	0.135	0.065
	$\hat{\beta}$	3.992	0.051	3.997	0.041	3.998	0.035
	$\hat{c}$	0.991	0.042	0.994	0.027	0.996	0.015
	$\hat{\rho}$	-0.987	0.100	-0.994	0.083	-0.9982	0.070

TAB. 5.2 – Indirect estimator for Model 2 with 200 paths. For $\Delta = 0.1$, $b_0 = 0.25$ and for $\Delta = 0.05$, $b_0 = 0.125$.

Δ	Parameters	$n = 1000$		$n = 2000$		$n = 3000$	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
$\Delta = 0.1$	$\hat{b}$	0.355	0.295	0.323	0.223	0.279	0.170
	$\hat{\beta}$	3.995	0.168	4.010	0.106	3.999	0.021
	$\hat{c}$	0.980	0.418	0.980	0.382	0.983	0.070
	$\hat{\rho}$	-1.503	0.106	-1.493	0.104	-1.501	0.051
$\Delta = 0.05$	$\hat{b}$	0.235	0.199	0.220	0.152	0.169	0.117
	$\hat{\beta}$	3.978	0.081	4.001	0.070	3.996	0.021
	$\hat{c}$	0.967	0.336	0.977	0.143	0.991	0.473
	$\hat{\rho}$	-1.495	0.151	-1.494	0.078	-1.493	0.062

TAB. 5.3 – Indirect estimator for Model 3 with 200 paths. For $\Delta = 0.1$, $b_0 = 0.25$ and for $\Delta = 0.05$, $b_0 = 0.125$.

Remark :
For computation of estimator of θ we will need a consistent estimator of the optimal matrix $\Sigma^* = I_0^{-1}$. For this, we can use the Newey and West (1987) formula. If $Q_n(\omega) = \frac{1}{n}\sum_{i=1}^{n}\psi_i(\omega)$ then I_0 is estimated by

$$\hat{I}_n = \hat{\Gamma}_0 + \sum_{k=1}^{K}\left(1 - \frac{k}{K+1}\right)(\hat{\Gamma}_k + \hat{\Gamma}_k')$$

with $\hat{\Gamma}_k = \frac{1}{n}\sum_{j=k+1}^{n}\frac{\partial\psi_{j-k}}{\partial\omega}(\hat{\omega}_n)\frac{\partial\psi_j}{\partial\omega'}(\hat{\omega}_n)$
(K is a function of n which doesn't increase at a high rate, but enough to ensure the consistency of $\hat{I}_n$.)

5.6 Concluding remarks

A possible generalization could be to consider that the function a has the form $a(x) = cx^\gamma$ where $\gamma \in (\frac{1}{2}, 1)$ is an unknown parameter to estimate.
Or to more accurately take into account the structure of the stochastic volatility model with a leverage effect, it could be interesting to take a nonparametric GARCH model as auxiliary model and to choose an auxiliary criterion.

Troisième partie

Mélange de modèles mixtes pour l'analyse des appariements de chromosomes chez le colza

Chapitre 6

Modèle mixte avec mélange : application à l'analyse des appariements de chromosomes chez le colza

Sommaire

[1]Ce travail a été fait en collaboration avec Sylvie Huet, Hervé Monod, Eric Jenczewski et Frédérique Eber. Il fait l'objet de deux articles parus dans le *Journal de la Société Française de Statistique*, **143**, p 147 153 ; et dans *Genetics*, **164**, p645-653.

Résumé

Lors de la méiose de colzas haploïdes (ne contenant qu'une copie du génôme), un certain nombre de chromosomes homologuent s'apparient et les autres, dits univalents, restent non appariés. Le nombre d'univalents est variable et dépend en particulier de la variété de colza. L'objectif est de savoir si le contrôle des appariements de chromosomes, lors de la méiose, est dû à l'action d'un gène unique ou non.

Le modèle adapté aux données observées est un modèle de mélange, dont chacun des deux composants suit un modèle mixte. Les paramètres sont estimés par maximum de vraisemblance, calculé avec un algorithme ECM. Le test du rapport de vraisemblance est présenté pour deux hypothèses sur les paramètres, dont l'une inclut la nullité d'un paramètre de variance.

6.1 Introduction

Des modèles de mélange apparaissent fréquemment en génétique, lorsque l'on analyse des données issues de croisements entre des parents différents génétiquement (Loisel *et al*, 1994). Les données issues de l'expérience décrite ci-dessous présentent une structure de covariance qui est prise en compte à l'aide d'un modèle de mélange de deux modèles mixtes. Nous montrons comment les méthodes d'analyse du modèle mixte s'adaptent au cas d'un modèle de mélange.

6.2 Présentation des données

Lors de la méiose de colzas haploïdes (ne contenant qu'une copie du génôme), un certain nombre de chromosomes s'apparient et les autres, dits univalents, restent non appariés. Le nombre d'univalents est variable et dépend en particulier de la variété de colza. L'objectif est de savoir si le contrôle des appariements de chromosomes est dû à l'action d'un gène unique ou non.
Les haploïdes issus de la variété Darmor ont un fort taux d'appariement (donc peu d'univalents), alors que ceux issus de la variété Yudal ont un faible taux d'appariement. Ces deux variétés de colza ont été croisées et des haploïdes ont été produits à partir des variétés parents et à partir des hybrides de première génération, dits F1. Des comptages de chromosomes univalents ont été réalisés, en quatre lots d'observations, sur des cellules de plantes haploïdes issues des variétés parents (de 15 à 50 cellules prélevées par individu, en moyenne 20) et sur des cellules de plantes haploïdes produites à partir des F1 (de 14 à 149 cellules prélevées par individu, en moyenne 20). La répartition des données entre les différents lots d'observation est résumée dans le Tableau 1.

lot	Nombre d'haploïdes (nombre de cellules)		
	Yudal	Darmor	F1
1	0 (0)	0 (0)	55 (1611)
2	10 (193)	20 (411)	109 (2208)
3	0 (0)	0 (0)	35 (688)
4	3 (124)	7 (182)	45 (1011)

TAB. 6.1 – Résumé des données disponibles.

La distribution bi-modale du nombre d'univalents, observée chez les haploïdes issus de plantes F1 (Fig. 1), met en évidence l'existence d'un gène dit gène majeur, qui a une influence prépondérante sur le taux d'appariement des chromosomes. Ce gène est nécessairement présent sous deux formes alléliques différentes dans les variétés

Darmor et Yudal, l'une plus défavorable aux appariements que l'autre. Chaque plante hybride F1 porte les deux allèles, reçus des parents Darmor et Yudal. Par contre, les haploïdes issus de ces hybrides F1 ne portent que l'un des deux allèles, déterminé aléatoirement pour chaque plante haploïde. La distribution bi-modale correspond donc à un mélange de deux populations, l'une constituée des haploïdes portant l'allèle "Darmor" du gène majeur, l'autre portant l'allèle "Yudal".
Ces observations conduisent à considérer que si plusieurs gènes agissent sur le taux d'appariement de chromosomes, il s'agit d'un gène majeur et d'un ensemble de gènes mineurs appelé fonds polygénique. En l'absence de fonds polygénique, les distributions d'univalents des deux populations d'haploïdes issus de F1 doivent être identiques aux deux distributions d'haploïdes issus des variétés parents. La présence d'un fonds polygénique, par contre, peut, d'une part modifier la moyenne des distributions par rapport aux variétés parents, d'autre part engendrer une variabilité entre plantes haploïdes.

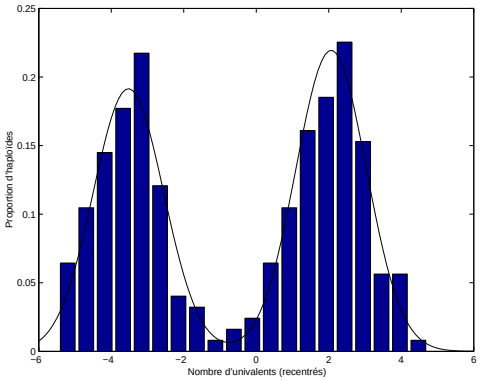


FIG. 6.1 – Histogramme des nombres d'univalents et densité estimée (moyennes par plantes haploïdes issues de F1)

Nous présentons dans la section suivante un modèle qui s'appuie sur cette première interprétation des données observées, et dont l'objectif est, d'une part de préciser l'influence du gène majeur, d'autre part de tester s'il existe de plus un effet lié à un fonds polygénique.

6.3 Modélisation

6.3.1 Modèle

Toutes les cellules haploïdes issues d'une même variété parent sont supposées homogènes génétiquement. On note z_{glij} la réponse pour la cellule j de l'haploïde i de génotype g (Yudal ou Darmor) et provenant du lot l. Pour ces observations, le modèle est

$$z_{glij} = \gamma_l + \mu_g + \varepsilon_{glij}$$

où γ_l ($l \in \{2,4\}$) est la moyenne (fixe) du lot l ; $\mu_g (g \in \{D,Y\})$ est l'effet variété (fixe) ; ε_{glij} est l'erreur résiduelle, supposée gaussienne centrée et de variance σ_E^2 ; $i \in \{1,...,n_{gl}\}$ où n_{gl} est le nombre d'haploïdes issus des parents ; $j \in \{1,...,m_{gli}\}$ où m_{gli} est le nombre de cellules observées. Pour éviter la surparamétrisation, on définit $\mu = \mu_D = -\mu_Y$.
Les haploïdes issus de plantes F1 se répartissent comme décrit précédemment en deux populations, que nous noterons Pd et Py, associées chacune à l'un des deux allèles du gène majeur. Pour les cellules issues de ces haploïdes, on note y_{lij} la réponse pour la cellule j de la plante i provenant du lot l. Pour ces observations, le modèle est

$$y_{lij} = \gamma_l + a_{li} + b_{li} + \varepsilon_{lij}$$

où γ_l ($l \in \{1,2,3,4\}$) est la moyenne (fixe) du lot l ; a_{li} est l'effet population ; b_{li} est un effet plante lié à l'éventuel "fonds génétique" et supposé aléatoire, gaussien, centré de variance σ_H^2, indépendant entre haploïdes mais commun à toutes les cellules issues d'un même haploïde ; ε_{lij} est l'erreur résiduelle, supposée gaussienne centrée et de variance σ_E^2 ; $i \in \{1,...,n_l\}$ où n_l est le nombre d'haploïdes issus de F1 ; $j \in \{1,...,m_{li}\}$ où m_{li} est le nombre de cellules observées. Les $\varepsilon_{lij}, a_{li}$ et b_{li} sont indépendants entre eux.
L'allèle du gène majeur porté par les haploïdes issus de plantes F1 étant inconnu, l'effet population a_{li} prend deux valeurs possibles μ_{Pd} et μ_{Py} avec probabilités p et $1-p$. Ainsi, pour des modalités fixées de la variable a_{li}, le modèle comprend des facteurs à effets fixes (lot, population, variété) et un facteur à effets aléatoires (plante issue de F1). On a donc un modèle mixte avec mélange, le mélange portant ici sur les effets du facteur population.

6.3.2 Expression de la log-vraisemblance

On note $\theta = (p, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \mu, \mu_{\mathrm{Pd}}, \mu_{\mathrm{Py}}, \sigma_E^2, \sigma_H^2)$ le vecteur des paramètres. Les données sont indépendantes entre haploïdes, dépendantes entre cellules d'un même haploïde. La log-vraisemblance, notée $L_n(\theta)$ s'exprime donc comme la somme des log-vraisemblances associées aux vecteurs d'observations de l'ensemble des cellules d'un même haploïde, notés Z_{gli} et Y_{li}.

$$\begin{aligned} L_n(\theta) &= \sum_{l=1}^{4} \sum_{i=1}^{n_l} \log\left[p f(Y_{li}, \gamma_l + \mu_{\mathrm{Pd}}, V_{li}) + (1-p) f(Y_{li}, \gamma_l + \mu_{\mathrm{Py}}, V_{li})\right] \\ &\quad + \sum_{g \in \{D,Y\}} \sum_{l \in \{2,4\}} \sum_{i=1}^{n_{gl}} \log\left[f(Z_{gli}, \gamma_l + \mu_g, \sigma_E^2 I_{m_{gli}})\right] \end{aligned}$$

où $V_{li} = \sigma_H^2 J_{m_{li}} + \sigma_E^2 I_{m_{li}}$; I_k désigne la matrice identité d'ordre k ; J_k désigne la matrice $k \times k$ dont tous les éléments sont égaux à 1 ; $f(Y, \varphi, V)$ représente la densité d'un vecteur gaussien d'espérance φ et de matrice de variance V.

6.4 Estimation des paramètres par maximum de vraisemblance

Dans notre modèle, on peut considérer que le vecteur des données est constitué de deux composantes : le vecteur des données observées (les nombres d'univalents par cellule) et le vecteur des données manquantes (l'allèle du gène majeur porté par les haploïdes issus de F1).
L'estimateur du maximum de vraisemblance dans le cas d'un modèle de mélange se calcule en général à l'aide d'un algorithme EM (Expectation-Maximisation, Dempster *et al*, 1977), en exploitant la décomposition en données observées et données manquantes que nous venons de décrire. Il s'agit d'un algorithme itératif, chaque itération comportant deux étapes : une étape, dite étape E, où l'on calcule l'espérance de la log-vraisemblance conditionnellement aux observations et aux valeurs courantes des paramètres, notée $Q(\theta)$, et une étape, dite étape M, de maximisation de la fonction Q.
Dans notre cas, la structure de modèle mixte nous a conduit à utiliser un algorithme ECM (Expectation-Conditional Maximisation, Meng et Rubin, 1993). Il s'agit également d'un algorithme itératif avec, à chaque itération, une étape E comme pour l'algorithme EM suivie de plusieurs étapes CM à la place de l'étape M de l'algorithme EM.
Pratiquement, nous considérons trois étapes CM. A l'itération t, la première étape CM consiste à calculer p et η, où η désigne le vecteur des paramètres associés à la partie fixe du modèle, $\eta = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \mu, \mu_{\mathrm{Pd}}, \mu_{\mathrm{Py}})$. L'estimation de p s'obtient par un calcul direct et η comme solution d'un système linéaire, similaire à celui obtenu pour la partie fixe d'un modèle mixte classique en supposant connues les composantes de la variance. On recherche ensuite les valeurs de σ_E^2, puis de ρ, où $\rho = \frac{\sigma_H^2}{\sigma_E^2}$, qui maximisent la vraisemblance obtenue en fixant les autres paramètres à leur valeur courante.
Les conditions de convergence de l'algorithme vers un point stationnaire de la log-vraisemblance données dans Meng et Rubin (1993) se vérifient facilement pour le modèle étudié.

6.5 Résultats

6.5.1 Estimation des paramètres

Les valeurs de θ estimées par maximum de vraisemblance sont présentées dans le Tableau 2.
On peut montrer que l'estimateur du maximum de vraisemblance possède les propriétés de consistance et de convergence en loi, lorsque l'on fait tendre vers l'infini les nombres d'haploïdes n_{gl} et n_l.

6.5.2 Test sur la ségrégation mendélienne

Une première hypothèse à tester est que la ségrégation du gène majeur suit les lois de Mendel, c'est-à-dire H_0 : "$p = \frac{1}{2}$". La statistique du test de rapport de vraisemblance de l'hypothèse H_0 contre l'alternative A_0 : "$p \neq \frac{1}{2}$", converge en loi vers un chi-deux à 1 degré de liberté.
L'hypothèse a été acceptée au niveau asymptotique 5% (cf. Tableau 2).

6.5.3 Test sur l'action d'un gène unique

Dans le modèle tel que nous l'avons présenté, l'hypothèse de l'action d'un gène unique s'exprime par H_0' : "$\sigma_H^2 = 0; \mu_D = \mu_{\mathrm{Pd}}; \mu_Y = \mu_{\mathrm{Py}}$". Sous cette hypothèse, le paramètre de variance σ_H^2 appartient à la frontière de son domaine de définition. Nous ne sommes pas dans un des cas référencés dans l'article de Self et Liang (1987), mais la configuration des paramètres se rapproche du cas 6. On montre, de manière similaire, que la statistique du test de rapport de vraisemblance de l'hypothèse H_0 contre l'alternative A_0' : "$\sigma_H^2 > 0$ ou $\mu_D \neq \mu_{\mathrm{Pd}}$ ou $\mu_Y \neq \mu_{\mathrm{Py}}$" converge en loi vers un mélange 50 : 50 de chi-deux à 2 et 3 degrés de liberté. La loi de ce mélange n'étant pas tabulée, on calcule le quantile empirique en simulant un échantillon de taille 100000.
L'hypothèse est nettement rejetée (cf. Tableau 2).

Bibliographie

DEMPSTER, A., LAIRD, N.M. et RUBIN, D.B. (1977) : *Maximum likelihood estimation from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, B **39**, pp. 1-38.

LOISEL, P., GOFFINET, B., MONOD, H. et MONTES DE OCA, G. (1994) : *Detecting a major gene in an F2 population.* Biometrics, **50**, pp. 512-516.

MENG, X.L., et RUBIN, D.B. (1993) : *Maximum likelihood estimation via the ECM algorithm : A general framework.* Biometrika, **80**, pp. 267-278.

SELF, S.G., et LIANG, K. (1987) : *Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard condition.* Journal of the American Statistical Association, **82**, pp. 605-610.

	Modèle sans contraintes	sous H_0	sous H_0'
$\hat{p}$	0,47	0,50	0,47
$\hat{\gamma}_1$	8,17	8,16	7,53
$\hat{\gamma}_2$	7,72	7,72	7,14
$\hat{\gamma}_3$	7,77	7,77	7,11
$\hat{\gamma}_4$	9,16	9,16	8,59
$\hat{\mu}_D$	-3,72	-3,72	-2,90
$\hat{\mu}_Y$	3,72	3,72	2,90
$\hat{\mu}_{\mathrm{Pd}}$	-3,53	-3,53	-2,90
$\hat{\mu}_{\mathrm{Py}}$	2,07	2,07	2,90
$\hat{\sigma}_E^2$	3,16	3,16	3,92
$\hat{\sigma}_H^2$	0,80	0,80	-
Maximum de la log-vraisemblance	-13036	-13037	-13503
Statistique de test	-	1,12	933,46
quantile	-	3,84	7,07

TAB. 6.2 – Estimations et résultats des tests sous différentes hypothèses

6.6 English article

Copyright © 2003 by the Genetics Society of America

PrBn, a Major Gene Controlling Homeologous Pairing in Oilseed Rape (*Brassica napus*) Haploids

Eric Jenczewski,*,[1] Frédérique Eber,* Agnès Grimaud,[†] Sylvie Huet,[†] Marie Odile Lucas,* Hervé Monod[†] and Anne Marie Chèvre*

**UMR ENSAR-INRA, Station de Génétique et Amélioration des Plantes, F-35653 Le Rheu, France and [†]INRA-Unité de Biométrie, F-78352 Jouy-en-Josas, France*

Manuscript received July 23, 2002
Accepted for publication February 4, 2003

ABSTRACT

Precise control of chromosome pairing is vital for conferring meiotic, and hence reproductive, stability in sexually reproducing polyploids. Apart from the *Ph1* locus of wheat that suppresses homeologous pairing, little is known about the activity of genes that contribute to the cytological diploidization of allopolyploids. In oilseed rape (*Brassica napus*) haploids, the amount of chromosome pairing at metaphase I (MI) of meiosis varies depending on the varieties the haploids originate from. In this study, we combined a segregation analysis with a maximum-likelihood approach to demonstrate that this variation is genetically based and controlled mainly by a gene with a major effect. A total of 244 haploids were produced from F_1 hybrids between a high- and a low-pairing variety (at the haploid stage) and their meiotic behavior at MI was characterized. Likelihood-ratio statistics were used to demonstrate that the distribution of the number of univalents among these haploids was consistent with the segregation of a diallelic major gene, presumably in a background of polygenic variation. Our observations suggest that this gene, named *PrBn*, is different from *Ph1* and could thus provide complementary information on the meiotic stabilization of chromosome pairing in allopolyploid species.

POLYPLOIDY has played a major role in the evolution of higher plants. Recent estimates suggest that up to 70% of all angiosperms have experienced one or more episodes of polyploidization during the course of their evolution (Masterson 1994) or domestication (Hilu 1993; van Raamsdonk 1995). Nonetheless, most if not all polyploids behave as diploids at meiosis, indicating that precise control of chromosome pairing is a prerequisite and confers evolutionary advantages in polyploid species. In allopolyploids containing homeologous chromosomes with sufficient homology to be able to pair at meiosis, cytological diploidization requires homeologous pairing to be suppressed. This process can be achieved by two complementary systems: (i) differentiation of homeologous chromosomes due to either structural changes or gene mutations, which leads to differential affinity and preferential pairing of homologs, and (ii) a genetic control that distinguishes between the differentiated sets of chromosomes and precludes pairing between homeologues. Although the presence of genetic systems regulating pairing has been suspected in a wide range of polyploids [*e.g.*, cotton (Kimber 1961), oat (Gauthier and McGinnis 1968), and fescues (Jauhar 1975)], evidence to date has been circumstantial. It is only in wheat that the presence of pairing regulators has been indisputably demonstrated with the characterization of the *Ph1* locus (Riley and Chapman 1958; Sears and Okamoto 1958) that suppresses homeologous pairing and contributes to the karyotypic stability of wheat (Sánchez-Morán *et al.* 2001). Several other weaker loci have also been shown to either restrict or promote homeologous pairing (Riley and Law 1965). The question remains whether similar pairing regulators are widespread among polyploid species and have therefore a general evolutionary significance. Differences in meiotic pathways among closely related species (Shaw and Moore 1998; Cuñado and Santos 1999) and in strategies for bivalent formation in different allopolyploids (Jenkins and Rees 1991) indicate that the genetic mechanisms characterized in wheat are not the same as in other polyploid species. It is therefore necessary to explore new and complementary models to further understanding of polyploid meiotic diploidization. This issue has practical applications. A large number of successful alien introgressions have been achieved in wheat through homeologous recombination (Friebe *et al.* 1996) and, notably, by suppressing the control exerted by *Ph1* (*e.g.*, Riley *et al.* 1968; Luo *et al.* 1996 and references therein; Benavente *et al.* 2001). It may be anticipated that a better understanding of the genetic systems regulating homeologous pairing in other polyploid species could help in promoting and engineering introgressions.

[1]*Corresponding author:* UMR APBV, INRA-GAP Rennes, Domaine de la Motte, BP 35327, F-35653 Le Rheu Cedex, France.
E-mail: jenczews@rennes.inra.fr

Oilseed rape (*Brassica napus*) is an allopolyploid species (AACC; $2n = 38$) that originated from hybridization between *B. oleracea* (CC; $2n = 18$) and *B. rapa* (AA; $2n = 20$). This species exhibits a clear bivalent-pairing regime and a disomic inheritance, which demonstrate that homologs pair at meiosis at the expense of homeologous pairing. The basis of such a diploid-like meiotic behavior is hypothetical. Different authors have proposed that homeologous pairing is genetically regulated in oilseed rape (Attia and Röbbelen 1986a; Sharpe *et al.* 1995) and its close relatives (Prakash 1974; Hardberg 1976; Eber *et al.* 1994). Renard and Dosba (1980) and Attia and Röbbelen (1986a) observed that the amount of chromosome pairing in haploid plants (AC; $n = 19$) originating from different oilseed rape varieties was variable and identified high- and low-pairing varieties (at the haploid stage). Chiasmata were formed between paired chromosomes and resulted in both rod- and ring-shaped bivalents, but also in multivalents.

The objectives of this study are to determine if a large part of the variation observed for the amount of chromosome pairing in oilseed rape haploids is genetically based and to establish the genetic basis of this variation. To do so, we analyzed the meiotic behavior of haploids produced from a high- and a low-pairing line and developed proper statistical analyses to account for the different sources of variation (genetic and environmental determinants). We studied the segregation of the meiotic behavior in a population of haploids produced from F_1 hybrids between the high- and low-pairing lines and used likelihood-ratio (LR) statistics to test for alternative modes of inheritance and interpret genetic distribution in terms of both major and minor gene effects.

MATERIALS AND METHODS

Plant materials: The genealogy and structure of the data sets are detailed in Figure 1. All haploids were isolated using microspore cultures as described by Polsoni *et al.* (1988). A total of 13 and 27 haploids were isolated from a spring Korean line (*Yudal*) and a French dwarf winter line (*Darmor-bzh*), which are known to vary in their meiotic behavior at the haploid stage (Figure 1). All the diploid lines used to produce these parental haploids (*Darmor-bzh* F_3 and F_4 progenies and *Yudal* F_9 and F_{13} progenies) were obtained by single-seed descent (SSD). These haploids comprise the parental data set. The parental genotypes differed in their response to produce haploids so that twice as many haploids were scored for *Darmor-bzh* as for *Yudal*. A total of 244 haploids were isolated from F_1 hybrids between *Darmor-bzh* and *Yudal* and comprise the offspring data set.

In an initial phase, three and seven haploids were produced from a few diploid plants from *Yudal* F_9 and *Darmor-bzh* F_3 progenies, respectively. A total of 45 haploids were isolated from F_1 hybrids obtained by crossing a single plant of the *Darmor-bzh* F_3 progeny to a single plant of the *Yudal* F_9 progeny (Figure 1). These parental and offspring haploids were grown together in the greenhouse and floral buds were sampled on almost the same date (three dates within 15 days); these haploids comprise the first set of observations (series 1). In a second phase, 10 and 20 haploids were isolated from *Yudal* F_{13} and *Darmor-bzh* F_4 progenies. A total of 199 haploids were isolated from a few F_1 hybrids obtained by crossing a single *Darmor-bzh* F_4 plant by a single *Yudal* F_{13} plant; three microspore cultures were needed to isolate all the haploids. Accordingly three series of haploids were successively grown in the greenhouse and analyzed separately (series 2–4 of observations). Only one set was grown simultaneously with parental haploids (series 3). For series 2 and 4, floral buds were sampled on three to four dates within 15 days. For series 3, floral buds were sampled on three dates within 1 month. For series 1 and 3, some haploids were observed at each date (or at least more than once) and showed the same meiotic behavior (data not shown). Sixteen other haploids were observed in 2 consecutive years to test for the repeatability of the amount of pairing. These haploids were chosen within series 2–4 to encompass the whole range of meiotic behaviors [3.3 < no. of univalents < 10]. These haploids were conserved as cuttings.

Meiotic observations: Floral buds were fixed in Carnoy's solution (ethanol-chloroform-acetic acid, 6:3:1) for 24 hr and stored in 50% ethanol. Observations on the pollen mother cells (PMCs) were performed at the metaphase I (MI) stage from anthers squashed and stained in a drop of 1% acetocarmine solution. On average, 20 PMCs (minimum, 14; maximum, 149) were examined for each haploid, regardless of their origin.

Statistical analysis: Statistical analyses were performed mainly on the number of univalents. This variable was chosen because it can be reliably scored and because it measures the whole extent of pairing in a synthetic way, reflecting by subtraction the number of chromosomes associated as both bivalents and multivalents. Parental data were first analyzed on their own, to determine to what extent variation in the amount of pairing among *Darmor-bzh* and *Yudal* haploids was genotypically determined. On the basis of this preliminary analysis, the offspring and parental data were then analyzed simultaneously, so that parental and offspring distributions could be compared within a single model. In the models below, we denote by $Y_{g,lij}$ the number of univalents in the PMC j of haploid i observed in the series l from population g, where g refers to haploids produced from *Darmor-bzh* ($g = D$), *Yudal* ($g = Y$), or *Darmor-bzh* × *Yudal* F_1 hybrids ($g = H$).

Analysis of parental data: The model employed for each parental genotype was

$$Y_{g,lij} = \mu_g + \gamma_l + b_{g,li} + \varepsilon_{g,lij}, \qquad (1)$$

where μ_g is the mean for population g (g is either genotype D or genotype Y), γ_l is the effect of series l ($l = 1$ or 3), $b_{g,li}$ is a random haploid plant effect, and $\varepsilon_{g,lij}$ is a residual error term. The $b_{g,li}$ and $\varepsilon_{g,lij}$ random effects were assumed to follow independent normal and centered distributions, with variances denoted by τ_g^2 for haploid effects and by σ_g^2 for residual errors. The parameter estimates in model (1) and their asymptotic standard errors were calculated by residual maximum likelihood (REML) with the PROC MIXED procedure of SAS (SAS Institute 1999). Note that for variance parameter estimates, the standard errors are known to be unreliable and should not be used to construct confidence intervals. The hypotheses on the absence of series effects ($\gamma_1 = \gamma_3 = 0$ *vs.* $\gamma_1 \neq \gamma_3$) and of haploid effects ($\tau_g^2 = 0$ *vs.* $\tau_g^2 > 0$) were tested by an analysis of variance performed with the PROC GLM procedure of SAS. The RANDOM statement of this procedure was used because the haploid random factor was nested within the series factor. On two occasions in results, we propose to quantify and compare the contributions to the variability between haploids that are due to the different factors of the models. For factors with random effects, contributions are

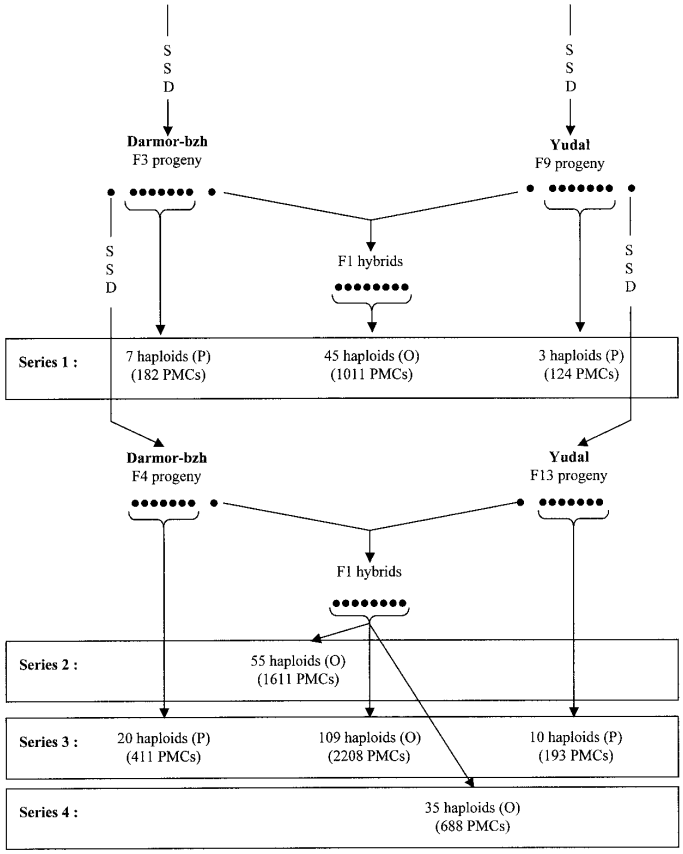


FIGURE 1.—Genealogy and structure of the parental (P) and offspring (O) subpopulations employed for segregation analyses. The total number of pollen mother cells (PMCs) observed is indicated for each series in each subpopulation. All the diploid lines used to produce the parental haploids were obtained by single-seed descent (SSD).

given by the estimated variance parameters. For factors with fixed effects, contributions are calculated as the average squared estimates of the factor effects.

Simultaneous analysis of parental and offspring data: We considered a model with both a segregating diallelic major gene and a completely additive polygenic background. This model allowed the presence of two subpopulations in the offspring data set, one with a behavior similar to *Darmor-bzh* haploids (denoted HD) and the other one with a behavior similar to *Yudal* haploids (denoted HY). Our general model was

$$Y_{D,lij} = \mu_D + \gamma_l + b_{D,li} + \varepsilon_{D,lij}$$

$$Y_{Y,lij} = \mu_Y + \gamma_l + b_{Y,li} + \varepsilon_{Y,lij}$$

$$Y_{H,lij} = \begin{cases} \mu_{HD} + \gamma_l + b_{HD,li} + \varepsilon_{H,lij} & \text{with probability } p \\ \mu_{HY} + \gamma_l + b_{HY,li} + \varepsilon_{H,lij} & \text{with probability } 1-p, \end{cases} \quad (2)$$

where μ_D, μ_Y, γ_l, $b_{D,li}$, $b_{Y,li}$, $\varepsilon_{D,lij}$, and $\varepsilon_{Y,lij}$ are as defined for the parental data; μ_{HD} and μ_{HY} are the means for the two subpopulations HD and HY; $b_{HD,li}$ and $b_{HY,li}$ are the random haploid plant effects for the two subpopulations HD and HY; $\varepsilon_{H,lij}$ is a residual error term; and p and $1 - p$ are the transmission probabilities of the *Darmor-bzh* and *Yudal* major-locus alleles, respectively. The $b_{g,li}$ and $\varepsilon_{g,li}$ random effects were assumed to follow independent normal and centered distributions. Haploid variances τ_g^2 (with $g = D$, Y, HD, or HY) were assumed to depend on the genotype, whereas the residual variance σ^2 was assumed to be the same for all PMCs. Model (2) involves factors with fixed (genotype and series) and random effects (plants) and is therefore a mixed model. In the offspring data, it includes the mixture of two distributions with different means and variances (EVERITT and HAND 1981).

The vector of parameters is $\theta = (p, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \mu_D, \mu_Y, \mu_{HD}, \mu_{HY}, \tau_D^2, \tau_Y^2, \tau_{HD}^2, \tau_{HY}^2, \sigma^2)$, with the constraints $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 0$ to avoid overparameterization and $\mu_{HY} - \mu_{HD} > 0$ to ensure that the parameters can be identified. The parameter estimates and their asymptotic standard errors were calculated by Gaussian likelihood maximization. For each g, l, i ($g = D$, Y, or H), $\mathbf{Y}_{g,li}$ denotes the vector of observations on all PMCs from plant i of genotype g in series l. In the model, the $\mathbf{Y}_{g,li}$'s are mutually independent and so the total log-likelihood is equal to the sum of the log-likelihoods for each vector $\mathbf{Y}_{g,li}$. $\phi(\mathbf{Y};\mathbf{m},\Sigma)$ denotes the density of a Gaussian vector with expectation $\mathbf{m}$ and covariance matrix Σ, calculated in $\mathbf{Y}$. For $g = D$ or Y, the likelihood for $\mathbf{Y}_{g,li}$ is $\phi(\mathbf{Y}_{g,li};(\gamma_1 + \mu_g)\mathbf{1}, \tau_g^2\mathbf{J} + \sigma^2\mathbf{I})$, where $\mathbf{1}$ denotes the vector of ones of appropriate length, $\mathbf{I}$ is the identity matrix, and $\mathbf{J}$ is the matrix of ones. For $g = H$, the likelihood for $\mathbf{Y}_{g,li}$ is $p\phi(\mathbf{Y}_{g,li};(\gamma_l + \mu_{HD})\mathbf{1}, \tau_{HD}^2\mathbf{J} + \sigma^2\mathbf{I}) + (1 - p)\phi(\mathbf{Y}_{g,li};(\gamma_l + \mu_{HY})\mathbf{1}, \tau_{HY}^2\mathbf{J} + \sigma^2\mathbf{I})$, which we denote by $f(\mathbf{Y}_{g,li};\theta)$.

The maximization of the likelihood coming from a mixture model is usually carried out using an expectation-maximization (EM) algorithm (DEMPSTER *et al.* 1977). In this study, the numerical procedure was improved by using a generalization of this algorithm, namely an expectation-conditional-maximization (ECM) algorithm (MENG and RUBIN 1993). In-

deed, the ECM algorithm allowed us to maximize the log-likelihood more efficiently, by separating maximization with respect to the variance parameters and maximization with respect to the other model parameters. The algorithm was programmed using the MATLAB software (MathWorks 2000).

The testing procedure was based on the LR test statistic. This procedure tests the null hypothesis that the vector of parameters satisfies a set of q linear constraints against the alternative that at least one of these constraints is not satisfied. LR equals twice the difference between the maximum log-likelihoods under the alternative and null hypotheses. The null hypothesis is rejected at level 5% when the test statistic is greater than the 95% quantile of a χ^2 distribution with q d.f. This test proved to be approximately of level 5% when the numbers of haploids and PMCs per haploid are large (Graybill 1976).

Finally, we used the estimated model parameters to predict the major-locus genotype of all haploids in the offspring data set. According to Bayes' theorem, the probability that a haploid from the offspring data set carries the *Darmor-bzh* allele, conditionally to its vector $\mathbf{Y}_{H,li}$ of observed values, is

$$P(D/\mathbf{Y}_{H,li}) = \frac{p\phi(\mathbf{Y}_{H,li};(\gamma_l + \mu_{HD})\mathbf{1}, \tau^2_{HD}\mathbf{J} + \sigma^2\mathbf{I})}{f(\mathbf{Y}_{g,li};\theta)}$$

and the probability that it carries the *Yudal* allele is $P(Y/\mathbf{Y}_{H,li}) = 1 - P(D/\mathbf{Y}_{H,li})$ (Everitt and Hand 1981).

For estimating the repeatability of our observations, we computed the correlation between repeated measures (mean number of univalents) on the 16 haploids from the offspring data set that have been observed in 2 consecutive years (Falconer and Mackay 1996).

RESULTS

Analysis of the variation among *Darmor-bzh* and *Yudal* haploids: Typical pairing patterns at MI for *Darmor-bzh* and *Yudal* haploids are illustrated in Figure 2. Figure 3A presents the mean numbers of univalents for each plant in the parental data set. Averaged meiotic behaviors for *Darmor-bzh* and *Yudal* haploids estimated by REML and corrected for the series and haploid effects (Table 1) demonstrate that pairing patterns in *Darmor-bzh* and *Yudal* haploids are clear cut. Haploids produced from *Darmor-bzh* showed far more pairing than those originating from *Yudal*; 80% of the PMCs observed in the *Darmor-bzh* haploids had less than six univalents whereas 95% of the PMCs scored in the *Yudal* haploids had more than eight univalents. On average, only 36.8% of the chromosomes paired in the *Yudal* haploids while >75% of the chromosomes were associated in the *Darmor-bzh* haploids. Similar differences were observed with the number of multivalents: 41 trivalents (III) and 60 quadrivalents (IV) were scored in a total of 593 PMCs from 27 *Darmor-bzh* haploids, whereas 19 III and only 2 IV were scored over the 317 PMCs analyzed from the 13 *Yudal* haploids. Regardless of the genotype, bivalents and multivalents were held by chiasmata.

Using the number of univalents as a variable, the estimated values (plus or minus their standard errors) for the parameters of model (1) were $\mu_D = 4.82$ (± 0.08), $\gamma_1 = -\gamma_3 = 0.98$ (± 0.08), $\tau^2_D = 0.02$ (± 0.04), $\sigma^2 = 2.52$ (± 0.15) for *Darmor-bzh* haploids and $\mu_Y = 12.03$ (± 0.44), $\gamma_3 = -\gamma_1 = 0.26$ (± 0.33), $\tau^2_Y = 0.86$ (± 0.44), $\sigma^2 = 3.78$ (± 0.31) for *Yudal* haploids, respectively. The analyses of variance (Table 2), performed separately on each parental line, showed that significant differences existed between the two series of haploids (series 1 and 3) produced from *Darmor-bzh* ($P = 1.8e\text{-}11$), which differed on average by the association of two chromosomes as a bivalent. By contrast, no differences were observed between the two series of haploids produced from *Yudal* whereas haploids within each series were significantly different from one another ($P = 2.2e\text{-}09$). This result is surprising because the diploid *Yudal* plants used for microspore culture were from the same F_9 or F_{13} progenies and were therefore genetically almost homogeneous. By contrast, no differences were detected among the haploids of the same generation produced from *Darmor-bzh*.

According to parameter estimates, 93% of the observed variability for the number of univalents could be

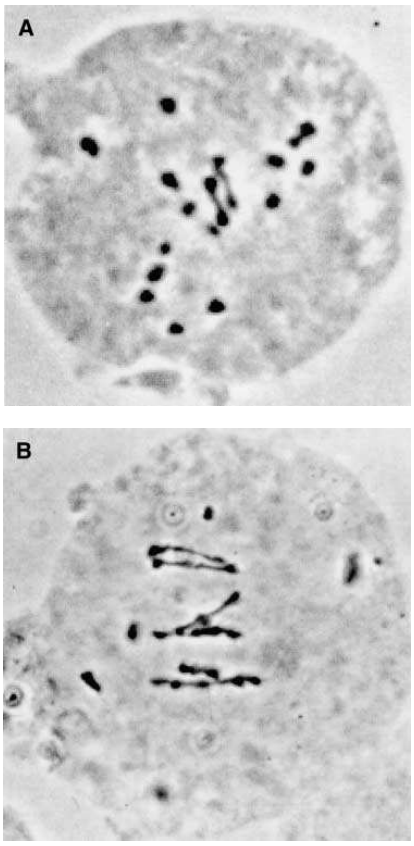


Figure 2.—First metaphase of meiosis in 1% acetocarmine-stained squashes of pollen mother cells of two oilseed rape haploids produced from the parental lines: (A) low pairing in a *Yudal* haploid, with two bivalents and 15 univalents; (B) high pairing in a *Darmor-bzh* haploid, with eight bivalents and only 3 univalents.

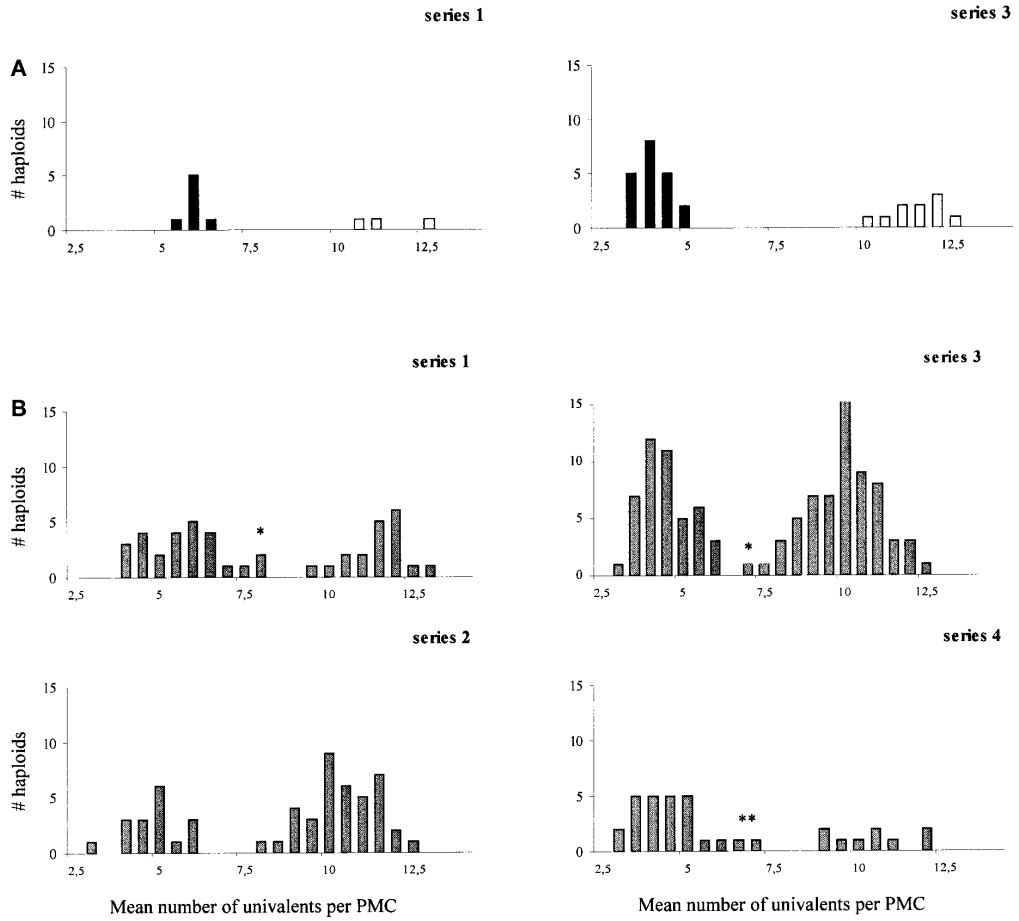


FIGURE 3.—Mean numbers of univalents per PMC for each haploid produced from either (A) the parental lines or (B) the hybrids between *Darmor-bzh* and *Yudal*. Haploids produced from *Darmor-bzh* are represented by solid histograms and those produced from *Yudal* by open histograms. Asterisk (*) points to haploids with intermediate behavior.

attributed to differences between the parental genotypes, 4% to differences between series (calculated by averaging over the two parental lines), and 3% to differences between haploids within a series.

Analysis of the whole data set, including the segregating population of haploids: Figure 3B presents the mean numbers of univalents for each plant in the offspring data set. These values proved to be very repeatable and reliable using Falconer's method; a very high correlation ($r_F = 0.96$) was observed among repeated measures on the 16 haploids that had been chosen in the offspring data set to encompass the whole range of meiotic behaviors. The maximum absolute differences in the mean number of univalents between the 2 years of observation for a plant were 1.7 and then 1.05.

The distribution of the mean number of univalents in the offspring data set was clearly bimodal with a mixture of two distinct distributions (Figure 4). Ac-

TABLE 1

Averaged meiotic behavior in haploids produced from the two parental lines

	No. of cells	I	II	III	IV
Darmor-bzh	597	4.82 ± 0.08	6.77 ± 0.05	0.07 ± 0.01	0.11 ± 0.01
Yudal	317	12.03 ± 0.33	3.37 ± 0.136	0.07 ± 0.02	0.007 ± 0.004

I, II, III, and IV are the mean numbers of univalents, bivalents, trivalents, and quadrivalents per PMC, respectively. These values ±standard errors have been estimated by REML.

TABLE 2

Analyses of variance on the numbers of univalents observed among haploids produced from *Darmor-bzh* and *Yudal* parental lines, respectively

	Darmor-bzh haploids			*Yudal* haploids		
Factor	d.f.	MS	*F* value	d.f.	MS	*F* value
Series	1	462.18	150.34***	1	17.54	0.57 NS
Haploid	25	3.0	1.2 ns	11	23.7	6.3***
Residual	566	2.5		304	3.8	
Error for series	22.63[a]	3.07		10.07[a]	31.03	

NS, not significant; MS, mean square. ***$P < 10^{-3}$.
[a] Approximated degrees of freedom.

cording to model (2), the parameter estimates (plus or minus their standard errors) were $p = 0.46$ (± 0.03), $\mu_D = 4.50$ (± 0.36), $\mu_Y = 12.06$ (± 0.38), $\mu_{HD} = 4.67$ (± 0.35), $\mu_{HY} = 10.27$ (± 0.60), $\gamma_1 = 1.08$ (± 0.36), $\gamma_2 = -0.05$ (± 0.27), $\gamma_3 = -0.56$ (± 0.35), $\gamma_4 = -0.48$ (± 0.62), $\tau^2_D = 0.005$ (± 0.07), $\tau^2_Y = 0.97$ (± 1.05), $\tau^2_{HD} = 0.70$ (± 0.21), $\tau^2_{HY} = 0.91$ (± 1.09), and $\sigma^2 = 3.11$ (± 0.17). Adjusted and observed distributions of the mean number of univalents for each haploid of the offspring data set were in close agreement (Figure 4).

We initially tested whether the distribution of the number of univalents was consistent with the Mendelian segregation of a major gene. The full model that treated transmission probability as an unknown parameter was compared, using a likelihood-ratio test, with the restricted model that fixed $p = 0.5$. As the full model did not provide a better fit than the restricted one ($P = 0.25$), the hypothesis $p = 0.5$ was accepted at the 5% level (Table 3). Therefore the distribution of the mean number of univalents in the offspring data set supports the presence of a major gene with two alleles.

To analyze whether this major gene was the only genetic source of variation, we tested whether the distribution of the number of univalents in the offspring data set was consistent with the mixture of the two parental distributions within a given series: this would be expected if the amount of pairing was completely controlled by the major gene. We compared the full model that treated μ_{HD}, μ_{HY}, τ^2_{HD}, and τ^2_{HY} as free parameters with restricted models that fixed $\mu_D = \mu_{HD}$, $\mu_Y = \mu_{HY}$, $\tau^2_{HD} = \tau^2_D$, or $\tau^2_{HY} = \tau^2_Y$. Note that, as a consequence of using a mixture model, each test was performed on the whole data set and took account of the uncertainty on the subpopulations of offspring haploids. Results, presented in Table 3, clearly showed that the distribution of the number of univalents in the offspring data set was not a mixture of the two parental distributions. First, the mean number of univalents was significantly lower in the HY subpopulation than in the parental *Yudal* haploids. By contrast there were no differences between the *Darmor-bzh* mean and that of the HD subpopulation ($P = 0.17$), although μ_{HD} was slightly higher than μ_D. Second, although τ^2_{HY} was not significantly different from τ^2_Y, the variance for the number of univalents was

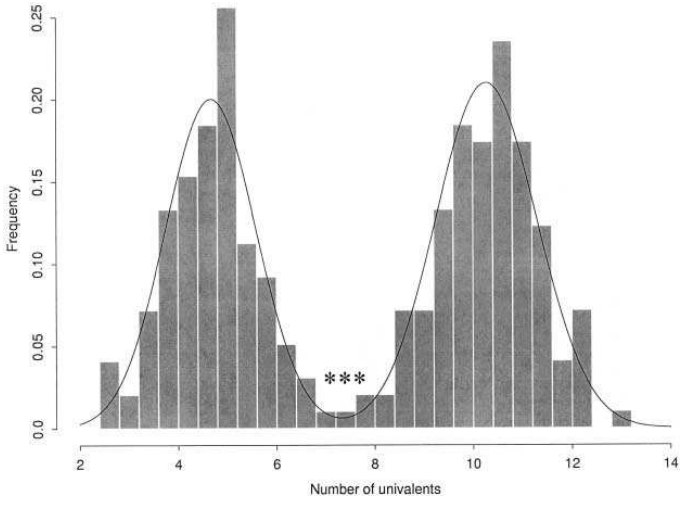


FIGURE 4.—Comparison of the observed (histogram) and estimated (solid curve) frequency distributions of the mean number of univalents in the offspring haploids. The observed distribution was obtained by pooling observations (mean number of univalents, adjusted for the estimated series effects) from all the series of the offspring data set. Asterisks (*) point to haploids with intermediate behavior.

TABLE 3

Likelihood-ratio (LR) tests for different hypotheses assuming a Mendelian single-locus model for the inheritance of the meiotic behavior in oilseed rape haploids

Hypothesis	Parameter constraints	Likelihood-ratio statistic	d.f.	*P* value
Mendelian segregation	$P = 0.5$	1.3	1	0.25
Adjustment of means	$\mu_D = \mu_{HD}$	1.895	1	0.17
	$\mu_Y = \mu_{HY}$	17.23	1	$<10^{-3}$
Adjustment of variance	$\tau_D = \tau_{HD}$	19.13	1	$<10^{-3}$
	$\tau_Y = \tau_{HY}$	0.017	1	0.90
	$\tau_Y = \tau_{HY} = \tau_{HD}$	1.11	2	0.57

significantly higher in the HD subpopulation than in the parental *D* subpopulation; actually, τ^2_D, τ^2_{HY}, and τ^2_{HD} were not different from one another (Table 3).

According to the parameter estimates of model (2), 86% of the observed variability for the number of univalents in the offspring data set was due to differences between the HD and HY subpopulations, 5% to differences between series, and 9% to differences between haploids within a series and a subpopulation. Interestingly, several haploids in the offspring data set exhibited an intermediate pairing behavior (Figures 3B and 4). Prediction of the major-locus genotype of each haploid in the offspring data set showed that three of them had probabilities $P(D/\mathbf{Y}_{H,li})$ and $P(Y/\mathbf{Y}_{H,li})$ <0.9. These three plants had an averaged meiotic behavior of 6.77 univalents (I) + 5.97 bivalents (II) + 0.025 III + 0.05 IV; 90% of their PMCs had 5 I + 7 II, 7 I + 6 II, or 9 I + 5 II. These patterns were quite different from those observed in the other haploids of the offspring data set since 82% of the PMCs had more than nine univalents in the HY subpopulation or less than five univalents in the HD subpopulation. The amount of pairing in two of these intermediate haploids was measured twice and values were similar in both measurements (data not shown).

DISCUSSION

Several authors have observed a variation in the extent of pairing among oilseed rape haploids (Renard and Dosba 1980; Attia and Röbbelen 1986a). In this study, we demonstrated that this variation is genetically based and controlled mainly by the major gene *PrBn* (*Pairing regulator in B. napus*).

Investigations on the meiotic behavior of poly-haploids have been made on a wide range of allopolyploid species, usually at MI or later stages (Kimber and Riley 1963; Magoon and Khanna 1963). These studies have usually demonstrated a low level of bivalent formation and/or various types of secondary associations of univalents (*i.e.*, not held by chiasmata). In this study, we observed a high level of pairing in the oilseed rape haploids isolated from *Darmor-bzh*, with up to 75% of the chromosomes being associated at MI. Even in the low-pairing haploids isolated from *Yudal*, a minimum of two to three bivalents were systematically observed. All these associations were held by chiasmata and probably resulted from both auto- and allosyndesis within and between the A and C genomes of oilseed rape. Autosyndesis, the pairing between two chromosomes originating from the same genome, has been reported within *B. oleracea* (Armstrong and Keller 1982) and *B. rapa* (Armstrong and Keller 1981) as a result of intragenomic duplications (Prakash and Hinata 1980; Schmidt *et al.* 2001). As only one or two autosyndetic pairs are possible within the C and A genomes (Mizushima 1950, 1972; Armstrong and Keller 1981, 1982), additional associations should be considered as allosyndetic (*i.e.*, involving chromosomes from different genomes). This assertion is consistent with the close proximity of A/C homeologous genomes of oilseed rape and with their high affinity for pairing, which is indirectly supported by the high amount of pairing reported in *B. rapa* × *B. oleracea* interspecific hybrids (Attia and Röbbelen 1986b); for example, Inomata (1980) observed that 81% of the PMCs in such a hybrid contained 1 I + 9 II, 8 II + 1 III, 8 II + 3 I, or 2 I + 7 II + 1 III and that the frequency of quadrivalents was ~10%.

Interestingly, high-pairing haploids of oilseed rape exhibit a meiotic behavior that is almost similar to that of raw *B. rapa* × *B. oleracea* interspecific hybrids (direct comparisons are ongoing). This suggests that these haploids express the largest extent of pairing affinities between the A and C genomes. By contrast, low-pairing haploids show a severe restriction in pairing potentialities. Such restriction has been used to infer the presence of pairing regulators (Kimber 1961; Riley and Law 1965) or interpreted as the consequence of an overdifferentiation of homeologous chromosomes since the origin of the polyploid state. This last proposal seems unlikely in oilseed rape haploids for several reasons. First, meiosis in *Darmor-bzh* × *Yudal* F_1 diploid hybrids

is regular (19 II; data not shown), suggesting that these genotypes do not differ by extensive chromosomal rearrangements. In addition, high levels of chromosome pairing in A × AC and AC × C hybrids indicated that the A/C genomes in oilseed rape have remained essentially unaltered with respect to the A/C genomes of their progenitors (Olsson and Hagberg 1955; Attia *et al.* 1987; see also Parkin *et al.* 1995). Finally our study provided direct evidence that the differences between the high- and low-pairing haploids are genetically based.

Our study combined a segregation analysis with a maximum-likelihood approach to test for different modes of inheritance of the pattern of chromosome pairing in oilseed rape haploids. A similar approach has been recently advocated by Wu *et al.* (2001) to combine quantitative genetic and population genetic principles. Our approach assumed normality of the underlying distributions, which appeared consistent with the observed mean numbers of univalents and simplified the form of the likelihood functions. Our statistical treatment provides a powerful and flexible framework to investigate the different sources of variation, test for parameter adjustment in the different parental (*D*, *Y*) and offspring (HD, HY) subpopulations independently, and interpret genetic data in terms of both major and minor gene segregation.

Segregation analysis combined with LR tests clearly demonstrates that pairing patterns in oilseed rape haploids are inherited in a Mendelian fashion and supports the presence of a single major gene. However, the distribution of the number of univalents in the offspring data set was not consistent with the mixture of the two parental distributions; an obvious asymmetry in the evolution of mean and variance parameters in the HD and HY subpopulations (*i.e.*, $\mu_D = \mu_{HD}$ and $\tau^2_{HD} > \tau^2_D$ while $\mu_Y > \mu_{HY}$ and $\tau^2_{HY} = \tau^2_Y$) was detected. This pattern may have resulted from the segregation of additional weaker genes with nonadditive effects that are confounded with the major gene activity or the range of chromosome pairing affinities, environmental variation affecting HD and HY haploids in a different way, or both. These interpretations are tentative although they are supported by additional observations. On the one hand, the meiotic behavior of *Yudal* haploids, which were taken and observed at the same date, was related to their position in the greenhouse (data not shown); this indicates that a large part of the unexpected variation observed between these haploids (Table 2) was due to environmental heterogeneity. By contrast, no variation was detected among *Darmor-bzh* haploids (Table 2), suggesting that pairing in high-pairing haploids was less susceptible to environmental variation than pairing in low-pairing haploids. On the other hand, strong differences between the *Y* and HY subpopulations (Table 3), the presence of intermediate haploids with a repeatable behavior, and the increased variance in the HD subpopulations (while pairing patterns in *Darmor-bzh* haploids did not vary) are consistent with the presence of a polygenic background.

Our results suggest that control of chromosome pairing in oilseed rape haploids is roughly similar to that in wheat in that major genes are involved in both cases. However, it is likely that *PrBn* is different from *Ph1*. First, polymorphism observed among oilseed rape haploids is natural whereas there is hardly any natural polymorphism for *Ph1* (see Ozkan and Feldman 2001); the only known wheat lines defective for *Ph1* have been induced through irradiation (Sears 1977; Giorgi 1978; Roberts *et al.* 1999). Second, *Ph1* prevents homeologous pairing at both the haploid and diploid stage. By contrast, all *B. napus* accessions, regardless of the frequency of chromosome pairing in their dihaploid forms, display regular bivalent associations and disomic inheritance. This indicates that, if the presence of *Ph1* is essential for chromosome stability and fertility in wheat (Sánchez-Morán *et al.* 2001), *PrBn* is not required. Alternatively, *PrBn* could contribute to the regularity of chromosome pairing in all diploid forms of *B. napus*, but the allele present in genotypes with a high-pairing behavior at the haploid stage could be ineffective at the hemizygous stage or at least less efficient than that in the diploid state. Such haplo-ineffective regulating systems have been described in hexaploid fescues (Jauhar 1975) and different Aegilops species (Cuñado and Santos 1999).

This last hypothesis is tentative and clearly deserves further examination. Ongoing genetic mapping and subsequent cloning of *PrBn*, comparative analysis of chromosome pairing at prophase I in high- and low-pairing haploids, and direct studies of the amount of recombination in oilseed rape diploids and haploids should further our understanding of the genetic regulation of chromosome pairing in this species. Then, combined with the extensive and continuous characterization of *Ph1* (Roberts *et al.* 1999), the Brassica model could provide new insights into the nature of the meiotic stabilization of allopolyploid species.

We thank J. C. Letanneur and D. Simmoneaux for technical assistance, Dr. E. Klein for his help in statistical analysis, and Drs. K. Alix, R. Delourme, and D. Gaudet for their fruitful comments on the manuscript.

LITERATURE CITED

Armstrong, K. C., and W. A. Keller, 1981 Chromosome pairing in haploids of *Brassica campestris*. Theor. Appl. Genet. **59:** 49–52.

Armstrong, K. C., and W. A. Keller, 1982 Chromosome pairing in haploids of *Brassica oleracea*. Can. J. Genet. Cytol. **24:** 735–739.

Attia, T., and G. Röbbelen, 1986a Meiotic pairing in haploids and amphihaploids of spontaneous versus synthetic origin in rape, *Brassica napus* L. Can. J. Genet. Cytol. **28:** 330–334.

Attia, T., and G. Röbbelen, 1986b Cytogenetic relationship within cultivated *Brassica* analyzed in amphihaploids from three diploid ancestors. Can. J. Genet. Cytol. **28:** 323–329.

Attia, T., C. Busso and G. Röbbelen, 1987 Digenomic triploids for an assessment of chromosome relationships in the cultivated diploid *Brassica* species. Genome **29:** 326–330.

Benavente, E., K. Alix, J. C. Dusautoir, J. Orellana and J. L.

David, 2001 Early evolution of the chromosomal structure of *Triticum turgidum-Aegilops ovata* amphiploids carrying and lacking the *Ph1* gene. Theor. Appl. Genet. **103:** 1123–1128.

Cuñado, N., and J. L. Santos, 1999 On the diploidization mechanism of the genus *Aegilops*: meiotic behaviour of interspecific hybrids. Theor. Appl. Genet. **99:** 1080–1086.

Dempster, A., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood estimation from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39:** 1–38.

Eber, F., A. M. Chèvre, A. Baranger, P. Vallée, X. Tanguy *et al.*, 1994 Spontaneous hybridization between a male-sterile oilseed rape and two weeds. Theor. Appl. Genet. **88:** 362–368.

Everitt, B. S., and D. J. Hand, 1981 *Finite Mixture Distributions*. Chapman & Hall, London.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longman Science and Technology, Harlow, UK.

Friebe, B., J. Jiang, W. J. Raupp, R. A. McIntosh and B. S. Gill, 1996 Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status. Euphytica **91:** 59–87.

Gauthier, F. M., and R. C. McGinnis, 1968 The meiotic behaviour of a nulli-haploid plant in *Avena sativa* L. Can. J. Genet. Cytol. **10:** 186–189.

Giorgi, B., 1978 A homoeologous pairing mutant isolated in *Triticum durum* cv. Capelli. Mut. Breed. Newsl. **11:** 4–5.

Graybill, F. A., 1976 *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, MA.

Hardberg, D. J., 1976 Cytotaxonomic studies of *Brassica* and related genera, pp. 47–68 in *The Biology and Chemistry of the Cruciferae*, edited by J. G. Vaughan, A. J. MacLeod and B. M. Jones. Academic Press, London.

Hilu, K. W., 1993 Polyploidy and the evolution of domesticated plants. Am. J. Bot. **80:** 1494–1499.

Inomata, N., 1980 Hybrid progenies of the cross *Brassica campestris* × *Brassica oleracea*. I. Cytogenetical studies on F1 hybrids. Jpn. J. Genet. **55:** 189–202.

Jauhar, P. P., 1975 Genetic regulation of diploid-like chromosome pairing in the hexaploid species, *Festuca arundinacea* Schreb. and *F. rubra* L. (Gramineae). Chromosoma **52:** 363–382.

Jenkins, G., and H. Rees, 1991 Strategies of bivalent formation in allopolyploid plants. Proc. R. Soc. Lond. Ser. B **243:** 209–214.

Kimber, G., 1961 Basis of the diploid-like meiotic behaviour of polyploid cotton. Nature **191:** 98–100.

Kimber, G., and R. Riley, 1963 Haploid angiosperm. Bot. Rev. **29** (4): 480–531.

Luo, M. C., J. Dubcovsky, S. Goyal and J. Dvorak, 1996 Engineering of interstitial foreign chromosome segments containing the K+/Na+ selectivity gene *Kna1* by sequential homoeologous recombination in durum wheat. Theor. Appl. Genet. **93:** 1180–1184.

Magoon, M. L., and K. R. Khanna, 1963 Haploids. Caryologia **16:** 191–234.

Masterson, J., 1994 Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science **264:** 421–424.

MathWorks, 2000 *Using MATLAB*. The MathWorks, Natick, MA.

Meng, X. L., and D. B. Rubin, 1993 Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika **80:** 267–278.

Mizushima, U., 1950 Karyogenetic studies of species and genus hybrids in the tribe *Brassicaceae* of Cruciferae. Tohoku J. Agric. Res. **1:** 1–14.

Mizushima, U., 1972 Evolution of species in *Brassiceae* and their breeding. Kagaku Seibutsu **10:** 78–85.

Olsson, G., and A. Hagberg, 1955 Investigations on haploid rape. Hereditas **41:** 227–237.

Ozkan, H., and M. Feldman, 2001 Genotypic variation in tetraploid wheat affecting homoeologous pairing in hybrids with *Aegilops peregrina*. Genome **44:** 1000–1006.

Parkin, I. A. P., A. G. Sharpe, D. J. Keith and D. J. Lydiate, 1995 Identification of the A and C genomes of amphidiploid *Brassica napus* (oilseed rape). Genome **38:** 1122–1131.

Polsoni, L., S. Kott and W. D. Beversdorf, 1988 Large-scale microspore culture technique for mutation-selection studies in *Brassica napus*. Can. J. Bot. **66:** 1681–1685.

Prakash, S., 1974 Probable basis of diploidization of *Brassica juncea* Coss. Can. J. Genet. Cytol. **16:** 232–234.

Prakash, S., and K. Hinata, 1980 Taxonomy, cytogenetics and origin of crop Brassica, a review. Opera Bot. **55:** 1–57.

Renard, M., and F. Dosba, 1980 Etude de l'haploidie chez le colza (*Brassica napus* L. var *oleifera* Metzger). Ann. Amel. Pl. **30:** 191–209.

Riley, R., and V. Chapman, 1958 Genetic control of the cytologically diploid behaviour of hexaploid wheat. Nature **13:** 713–715.

Riley, R., and C. N. Law, 1965 Genetic variation in chromosome pairing. Adv. Genet. **13:** 57–114.

Riley, R., V. Chapman and R. Johnston, 1968 Introduction of yellow rust resistance of *Aegilops comosa* into wheat by genetically induced homoeologous recombination. Nature **217:** 383–384.

Roberts, M. A., S. M. Reader, C. Dalgliesh, T. E. Miller, T. N. Foote *et al.*, 1999 Induction and characterization of Ph1 wheat mutants. Genetics **153:** 1909–1918.

Sánchez-Morán, E., E. Benavente and J. Orellana, 2001 Analysis of karyotypic stability of homoeologous-pairing (*ph*) mutants in allopolyploid wheat. Chromosoma **110:** 371–377.

SAS Institute, 1999 *SAS/STAT User's Guide*, Version 8. SAS Institute, Cary, NC.

Sears, E. R., 1977 An induced mutant with homoeologous pairing in common wheat. Can. J. Genet. Cytol. **19:** 585–593.

Sears, E. R., and M. Okamoto, 1958 Intergenomic chromosome relationships in hexaploid wheat. 10th Int. Congr. Genet. **2:** 258–259.

Sharpe, A. G., I. A. P. Parkin, D. J. Keith and D. J. Lydiate, 1995 Frequent non-reciprocal translocations in the amphidiploid genome of oilseed rape (*Brassica napus*). Genome **38:** 1112–1121.

Shaw, P., and G. Moore, 1998 Meiosis: vive la difference! Curr. Opin. Plant Biol. **1:** 458–462.

Schmidt, R., A. Acarkan and K. Boivin, 2001 Comparative structural genomics in the Brassicaceae family. Plant Physiol. Biochem. **39:** 253–262.

Van Raamsdonk, L. W. D., 1995 The cytological and genetical mechanisms of plant domestication exemplified by four crop models. Bot. Rev. **61:** 367–399.

Wu, R., B. Li, S. S. Wu and G. Casella, 2001 A maximum likelihood-based method for mining major genes affecting a quantitative character. Biometrics **57:** 764–768.

Communicating editor: P. J. Pukkila

BIBLIOGRAPHIE

Partie I

Abramowitz, M. et Stegun, A.I. editors (1972). *Handbook of Mathematical Functions : with formulas, graphs and mathematical tables.* Dover Books on Advanced Mathematics. Dover Publications.

Akaïke, H. (1973). Information theory and an extension of maximum likelihood principle. *Second International symposium on information theory*, 267-281.

Angevin, F., KleinE., Choimet C., Meynard J.M., de Rouw A., Sohbi Y. (2001). Modélisation des effets des systèmes de culture et du climat sur les pollinisations croisées chez le maïs. Rapport du groupe 3 du programme de recherche "Pertinence économique et faisabilité d'une filière sans utilisation d'OGM".

Barndorff-Nielsen, O.E. (1997). Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling. *Scandinavian Journal of Statistics* **24**, 1-13.

Bateman, A.J. (1947). Contamination of seed crops. II, Wind pollination. *Heredity* **1**, 235-246.

Batschelet, E. (1981). *Circular Statistics in Biology.* Academic Press, London.

Collett, D. (1991). *Modelling binary data.* Chapman and hall, London.

Durbin, J. (1992) The first-passage density of the brownian motion process to a curved boundary. *J. Appl. Prob.* **29**, 291-304.

Huet S., Bouvier A., Gruet M.A. and Jolivet E. (1996). *Statistical tools for nonlinear regression.* Springer-Verlag, New-York, USA.

Hurvich, C.M. and Tsai, C.L. (1995). Model selection for extended quasi-likelihood in small samples. *Biometrics* **51**, 1077-1084.

Karatzas, I. et Shreve, E.S. (1991). *Brownian Motion and Stochastic Calculus.* Seconde édition, Springer-Verlag, New-York, USA.

Klein, E. (2000) *Estimation de la fonction de dispersion du pollen. Application à la dissémination de transgènes dans l'environnement.* Thèse, Université Paris XI, Orsay.

Klein, E.K., Lavigne, C., Foueillassar, X., Gouyon, P.H., Laredo, C. (2003) Corn pollen dispersal : quasi-mechanistic models and field experiments. *Ecological Monographs* **73**, 131-150.

Kloeden, P.E. et Platen, E. (1992). *Numerical solution of stochastic differential equations.* Springer-Verlag, New-York.

Lavigne, C., Klein, E.K., Vallée, P., Pierre, J., Godelle, B. et Renard, M. (1998). A pollen-dispersal experiment with transgenic oilseed rape. Estimation of the average pollen dispersal of an individual plant within a field. *Theoretical and Apllied Genetics* **96**, 886-896.

Loubet, B , Brunet, Y;, Fouieillassar, X., Caltagirone, J.P. *et al.*, (2004). Etude mécaniste du transportet du dépôt de pollen de maïs dans un paysage hétérogène. Rapport de fin de projet.

MacQuarrie A. et Tsai C.L. (1998). *Regression and Time Series Model Selection.* World Scientific.

Mallows, (1973). Some comments on C_p. *Technometrics* **15**,661-675.

McCartney, H.A. et Fitt, B.D.L. (1998). Dispersal of foliar fungal plant pathogens : mechanisms, gradients and spatial patterns. Pages 138-160 dans *The epidemiology of plant diseases.* Kluwer, Dordrecht, The Netherlands.

McCullagh, P. (1983) Quasi-Likelihood functions. *The Annals of Statistics* **11**,N°1,59-67.

McCullagh P. et Nelder J.A. (1989). *Generalized Linear Models.* 2nd Edition Chapman and Hall, London.

Nurminiemi M., Tufto J., Nilsson O., Rognli O.A. (1998). Spatial models of pollen dispersal in the forage grass meadow fescue. *Evolutionary Ecology* **12**, 487-502.

Poilleux-Milhem, H. (2002) *1) Test de validation adaptatif dans un modèle de régression. 2) Modélisation et estimation de l'effet d'une discontinuité du couvert végétal sur la dispersion du pollen de colza.* Thèse, Université Paris XI, Orsay.

Portnoy, S. et Willson, M.F. (1993). Seed dispersal curves : behaviour of the tail of the distribution. *Evolutionary Ecology* **7**, 25-44.

Protter, P. (1992). *Stochastic Integration and Differential Equations. Applications of Mathematics.* New-York : Springer.

Prudnikov A.P., Brychkov Y.A., Marichev O.I. (1986). *Integrals and series.* Gordon and Breach Science Publishers, New-York, USA.

Rogers, L.C.G. et Williams, D. (1994). *Diffusions, Markov processes and martingales, Volume 2 Itô Calculus.* Cambridge University Press, Second edition.

Tufto, J., Engen, S., Hindar, K. (1997). Stochastic Dispersal Processes in Plant Populations. *Theoretical Population Biology* **52**, 16-26.

Van der Vaart, A. W. (1998). *Asymptotic Statistics* Cambridge University Press.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

Yamamura, K. (2004). Dispersal distance of corn pollen under fluctuating diffusion coefficient. *Popul Ecol* **46**, 87-101.

Partie II

Barndorff-Nielsen, O.E. et Shephard, N. (2001) Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion). *J. Roy. Statist. Soc. Ser. B* **63**, 167-241.

Black, F. et Scholes, M. (1973) The valuation of option and corporate liabilities. *Journal of political economy* **81**, 637-284.

Black, F. (1976) Studies of stock price volatility changes. *Proceesinds of the Business and Economic Statistics section, American Statistical Association*, 177-181.

Blatteberg, R. et Gonedes, N. (1974) A comparaison of the stable and Student distributions as statistical models for stock prices. *J. Business* **47**, 244-280.

Brillinger, D. R. (1981) *Time Series : Data Analysis and Theory.* Holden-Day, San Fransisco.

Brockwell, P.J. et Davis, R.A. (1991) *Time Series : Theory and Methods.* Springer-Verlag.

Cox, J. C., Ingersoll, J. E. et Ross, S. A. (1985) A theory of term structure of interest rates. *Econometrica* **53**, 385-407.

Dacunha-Castelle, D. et Duflo, M. (1993) *Probabilités et statistiques, Tome 2.* Seconde édition. Masson

Dahlhaus, R. (1988) Small sample effects in time series analysis : a new asymptotic theory and a new estimate. *Annals of Statistics* **16**, 808-841.

Elie, L ; et Jeantheau, T. (1995). Consistance dans les modèles hétéroscédastiques. *C.R. Acad. Sci. Paris, Séries I* **320**, 1255-1258.

Fan, J. et Yao, Q. (2003). *Nonlinear Time Series.* Springer-Verlag.

Gallant, A.R. et Tauchen, G. (1996) Which moments to match. *Econometric theory* **12**, 657-681.

Gallant, A. R., Hsieh, D., Tauchen, G. (1997) Estimation of stochastic volatility models with diagnostics. *J. Econometrics* **81**, 159-192.

Genon-Catalot, V., Jeantheau, T., Laredo, C. (1998) Limit theorems for discretly observed stochastic volatility models. *Bernouilli* **4**, 283-303.

Genon-Catalot, V., Jeantheau, T., Laredo, C. (1999) Parameter estimation for discretely observed stochastic volatility models. *Bernouilli* **5**, 855-872.

Genon-Catalot, V., Jeantheau, T., Laredo, C. (2000) Stochastic volatility models as hidden Markov models and statistical applications. *Bernouilli* **6**, 1051-1079.

Genon-Catalot, V., Jeantheau, T., Laredo, C. (2003) Conditional Likelihood Estimators for Hidden Markov Models and Stochastic Volatility Models. *Scandinavian Journal of Statistics* **30**, 297-316.

Ghysels, E., Harvey, A., Renault, E. (1996) Stochastic volatility. *Handbook of Statistics* **14** , 119-192.

Gloter A. (2000). Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient *ESAIM : Probability and Statistics* **4**, 207-215.

Gourieroux, C., Monfort, A. et Renault, E. (1993) Indirect inference. *Journal of Applied Econometrics* **8**, 85-118.

Hall, P. et Heyde, C.C (1980) *Martingale Limit Theory and its Application.* New-York : Academic Press.

Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029-1054.

Harvey, A. C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge : Cambridge University Press.

Heston, S. L. (1993) A closed-form solution for options with stochastic volatility, with applications to bond and currency options. *Review of Financial Studies* **6**, 327-343.

Hull, J. et White, A. (1987) The pricing of options on assets with stochastic volatilities. *J. Finance* **42**, 281-300.

Jacquier, E. , Polson, N.G. et Rossi, P.E. (2004) Bayesian analysis of stochastic volatility models with a fat-tails and correlated errors. *Journal of Econometrics* **122**, 185-212.

Kessler, M. (2000). Simple and explicit estimating functions for a discretely observed diffusion process *Scandinavian Journal of Statistics* **27**, 65-82

Meyn, S.P. et Tweedie, R.L. (1993) *Marchov Chains and Stochastic Stability.* Springer-Verlag.

Nelson, D.B. (1990). ARCH models as diffusion approximations. *J. Econometrics* **45**, 7-38.

Newey, W.K. et West, K.D. (1987). A simple positive definite heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703-708.

Rudin, W. (1966). *Real and complex analysis.* Mc. Graw Hill.

Sørensen, M. (2000). Prediction-based estimating functions. *Econom. J.* **3**, 121-147.

Taniguchi, M. et Kakizawa, Y. (2000) *Asymptotic Theory of Statistical Inference for Time Series.* Springer-Verlag.

Whittle, P. (1953). Estimation and information in stationary time series. *Ark. Mat.***2**, 423-434.

Résumé

La première partie de cette thèse est consacrée à l'étude de la dispersion du pollen de maïs. Le grain de pollen est vu comme une particule soumise à un champ de forces et sa trajectoire est modélisée à l'aide de différents processus de diffusion. Lorsque deux champs sont contigüs (milieu homogène), différentes fonctions de dispersion individuelles paramétriques sont alors obtenues, différentes hypothèses étant envisagées pour des temps d'atteinte de processus stochastiques. A partir d'expériences, les paramètres sont alors estimés en considérant un modèle de régression non linéaire. Le choix du modèle le mieux adapté se fait à l'aide d'un critère de type Akaïke et de méthodes graphiques. Par ailleurs ces modèles permettent d'effectuer des prédictions. Les résultats sont alors appliqués lorsque deux champs sont séparés par une autre culture (milieu hétérogène), afin d'étudier l'effet d'une discontinuité sur la dispersion.
Dans la seconde partie, on s'intéresse à des modèles à volatilité stochastique «mean-reverting», souvent utilisés en économie. Le processus observé est fonction d'une diffusion non observable dont on souhaite estimer les paramètres. Une méthode d'estimation à deux pas basée sur la structure ARMA(1,1) du processus est proposée, en utilisant un estimateur de moments et un contraste de Whittle. Des simulations sont réalisées afin de comparer cette méthode avec d'autres méthodes existantes. Ensuite un paramètre dit «leverage» est ajouté et un modèle discrétisé est étudié. Un critère auxiliaire est proposé pour estimer les paramètres à l'aide d'une méthode d'inférence indirecte. Enfin des simulations sont réalisées pour évaluer leurs performances.

Abstract

The first part of this work is devoted to the study of corn pollen dispersion. The pollen grain is seen as a particle subjected to a field of forces and its path is modelled using diffusion processes. In an homogeneous case (two contiguous corn fields), different parametric individual dispersal functions are obtained using different assumptions for stopping time of stochastic processes. From the experiments the parameters are estimated using a non-linear regression model. The choice of the most fitted model is then obtained using a criterion of Akaïke type and graphic methods. Moreover these models allow to make predictions. The results are then applied when two fields are separated by another culture (heterogeneous case), in order to study the effect of a discontinuity on dispersion.
In the second part, we study mean-reverting stochastic volatility models, often used in economy. The observed process is related to a non observable diffusion for which there are unknown parameters to estimate. First a two steps statistical method is proposed based on the ARMA(1,1) structure of the process, by using a moment estimator and a Whittle contrast. Simulations are made in order to compare this method with some known methods. Second a leverage parameter is introduced and a discretized model is studied. An auxiliary criterion is proposed to estimate the parameters using an indirect inference method. Finally simulations are made to evaluate their performances.