Modélisation des flux de gènes par approches Bayésiennes. Application à l'aide à la décision pour la coexistence de cultures OGM et non OGM.

Arnaud Bensadoun

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Ecole Doctorale ABIES
Agriculture Agriculture Alimentation Food Biologie Biology Environnement Environment Santé Health

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement

(AgroParisTech)

Spécialité : Sciences du Vivant

présentée et soutenue publiquement par

Arnaud BENSADOUN

le 17 février 2015

Modélisation des flux de gènes par approche bayésienne.

Application à l'aide à la décision pour la coexistence entre cultures OGM et non OGM

Directeur de thèse : **Hervé Monod**
Co-encadrement de la thèse : **Antoine Messéan**

Jury

M. Etienne KLEIN,	Directeur de recherche, Unité BioSP, INRA, Avignon, France	Rapporteur
M. Joe N. PERRY,	Professeur, Rothamsted research, Royaume-Uni	Rapporteur
Mme Chantal GUIHENNEUC,	Professeur, Université Paris Descartes, France	Examinateur
Mme Joaquima MESSEGUER,	Chercheur, IRTA, Girona, Espagne	Examinateur
M. François PIRAUX,	Statisticien, Arvalis-*Institut du végétal*, France	Examinateur
M. Antoine MESSEAN,	Directeur de recherche, Unité Eco-Innov, INRA, Grignon, France	Co-Directeur
M. Hervé MONOD,	Directeur de recherche, Unité MaIAGE, INRA, Jouy-en-Josas, France	Directeur

INRA
Mathématiques et Informatique Appliquées
Domaine de Vilvert, 78350 Jouy-en-Josas

Institut National de la Recherche Agronomique

Thèse

présentée pour obtenir le grade de Docteur délivré par

L'Institut des Sciences et Industries du Vivant et de l'Environnement (AgroParisTech)

Spécialité : Sciences du Vivant

par

Arnaud Bensadoun

Sous la direction de Hervé Monod et Antoine Messéan

Modélisation des flux de gènes par approches Bayésiennes.

Application à l'aide à la décision pour la coexistence de cultures OGM et non OGM.

Soutenue publiquement le 17 *Février* 2015
devant le jury composé de :

M. Etienne Klein, Directeur de recherche, unité BioSP, INRA, Avignon, France	Rapporteur
M. Joe N. Perry, Professeur, Rothamsted Research, Rothamsted, Royaume-Uni	Rapporteur
Mme Chantal Guihenneuc, Professeur, Université Paris Descartes, Paris, France	Examinateur
Mme Joaquima Messeguer, Chercheur, IRTA, Girona, Espagne	Examinateur
M. François Piraux, Statisticien, Arvalis - *Institut du végétal*, Paris, France	Examinateur
M. Antoine Messéan, Directeur de recherche, unité Eco-Innov, INRA, Grignon, France	Co-directeur
M. Hervé Monod, Directeur de recherche, unité MaIAGE, INRA, Jouy-en-Josas, France	Directeur

---

Thèse préparée à l'INRA, dans les unités
Mathématiques et Informatiques Appliquées, UR 341, 78350 Jouy-en-Josas, France, et
Eco-Innov, UAR 1240, 78850 Thiverval-Grignon, France

Table des matières

Table des matières i

Table des figures v

Liste des tableaux vii

Introduction 1

1 Problématique 5
1 Contexte et enjeux de la coexistence . . . 7
1.1 État des lieux des variétés de maïs OGM et des surfaces cultivées . 7
1.2 Contexe législatif . . . 8
2 Biologie et modélisation des flux de gènes sur le maïs . . . 14
2.1 Résultats expérimentaux sur la pollinisation croisée chez le maïs . . 15
2.2 Approche prédictive pour l'estimation du taux de pollinisation croisée 17
3 Coexistence, modélisation et incertitudes: le projet PRICE . . . 20
3.1 La modélisation au cœur du processus de décision . . . 21
3.2 Des prédictions? Oui, mais avec une incertitude associée! . . . 21
4 Objectifs de la thèse . . . 22

2 Conceptualisation 25
1 Situations de prédiction . . . 27
1.1 Problématique . . . 27
1.2 Exemples et définitions . . . 28
1.3 Conséquences opérationnelles . . . 31
2 Recherche d'un cadre statistique approprié . . . 34
2.1 L'école classique ou Fréquentiste . . . 34
2.2 L'école Bayésienne . . . 34
3 Approches Bayésiennes pour l'estimation des paramètres . . . 35
3.1 Vraisemblance . . . 36
3.2 Distribution *a priori* . . . 37
3.3 Distribution *a posteriori*, Metrolopolis-Hastings et MCMC . . . 37
4 Avantages de l'approche bayésienne pour l'aide à la décision . . . 39
4.1 Quantification des incertitudes pour toutes fonctions des paramètres 39
4.2 Mise à jour perpétuelle de l'information . . . 40
4.3 Modélisation hiérarchique . . . 40
5 Plusieurs types d'expérience/de données . . . 41
5.1 Types d'expériences: Mono-source vs. Multi-sources . . . 41

5.2 Types de données: discrètes vs. continues . . . 42
5.3 Types de configurations: trop de diversité? . . . 47

3 Modélisation 49
1 Représentations des échanges de pollen entre parcelles . . . 51
1.1 Mode global . . . 51
1.2 Mode individuel . . . 53
2 Modèle de l'espérance - Noyaux de dispersion . . . 54
2.1 Noyau de Student bivarié (2Dt) . . . 54
2.2 Noyau Normal Inverse Gaussian (NIG) . . . 55
2.3 Noyau Compound Exponential (CEx) . . . 55
3 Intégration de covariables au modèle . . . 56
3.1 Direction et force du vent . . . 56
3.2 Décalage de floraison . . . 57
4 Modèles d'observation . . . 59
4.1 Cas discret . . . 59
4.2 Cas continu . . . 61
5 Modèle d'ensemble et DAG . . . 62
5.1 Modèle d'ensemble . . . 62
5.2 DAG - Définitions et formalisme . . . 63
5.3 Représentations graphiques . . . 63
6 Maillage adaptatif pour l'approche individuelle . . . 68
6.1 Problématique . . . 68
6.2 Grille de référence et recherche d'approximations . . . 68
6.3 Plan d'expérience . . . 70
6.4 Protocole de simulation . . . 70
6.5 Mise en œuvre et exemples de grilles dégradées . . . 72
6.6 Résultats . . . 74
6.7 Application . . . 75

4 Estimation et Mise en œuvre 79
1 Contexte mono-source . . . 81
1.1 Données . . . 81
1.2 Estimation . . . 82
1.3 Prédiction . . . 92
2 Contexte multi-sources . . . 98
2.1 Données . . . 99
2.2 Intégration du décalage de floraison . . . 102
2.3 Stratégies pour l'estimation . . . 103
2.4 Prédiction avec modèle ajusté en *mono-source* . . . 106
3 Évaluation et Sélection de modèles . . . 112
3.1 Critères statistiques . . . 112
3.2 Analyse en composantes principales . . . 114

5 Discussion - Perspectives 127
1 Apports et difficultés des méthodes statistiques . . . 128
2 Influence des choix de modélisation sur la qualité de prédiction . . . 129
3 Adaptation des modèles aux situations de prédiction . . . 131

4 Mise à disposition et applications 132
5 Perspectives 134

Références **137**

Annexes **151**
Annexes 1 153
Annexes 2 201
Annexes 3 223
Annexes 4 255
Annexes 5 273

Table des figures

1.1 Étapes à prendre en compte pour l'étude de la coexistence entre différents types de filières de production végétale (*Source*: F. Angevin, INRA) 13

2.1 Illustration des différences entre données discrètes et données continues . . . 43
2.2 Schéma de dispositifs expérimentaux *mono-source* type *marqueur coloré* (*Source*: E. Klein (non publié)) 44
2.3 Schéma de dispositifs expérimentaux *mono-source* type *PCR* (*Source*:Van De Wiel et al. (2009)) 45
2.4 Schéma de dispositif de suivi *multi-sources* type *PCR* (*Source*: Messeguer et al. (2006)). 46

3.1 Représentation schématique des modes global et individuel pour la représentation des échanges de pollen entre parcelles. 52
3.2 Réprésentation schématique du métamodèle de prédiction. 62
3.3 Graphe acyclique orienté du modèle de Poisson à espérance fixe. 65
3.4 Graphe acyclique orienté du modèle de Poisson à espérance normale. 66
3.5 Graphe acyclique orienté du modèle ZIP à espérance fixe. 66
3.6 Graphe acyclique orienté du modèle ZIP à espérance normale. 67
3.7 Grille de référence. 72
3.8 Exemples de grilles dégradées pour un récepteur avec $\mathcal{C}_1 = (0, 4)$, $\mathcal{C}_2 = (4, 20)$, $\mathcal{C}_3 = (20, 24m)$, $F_1 = 4$, et avec de gauche à droite $F_3 = 200,\ 500$ et de haut en bas $F_2 = 50,\ 100$. 73
3.9 Front de Pareto - Qualité de prédiction en fonctions du coût de la simulation (nombre de points $\propto$ temps de calculs) 75
3.10 Grille minimisant le maximum de la valeur absolue des résidus $\mathcal{C}_1 = (0, 2)$, $\mathcal{C}_2 = (2, 7)$, $\mathcal{C}_3 = (7, 37)$, $F_1 = 2$, $F_2 = 100$, $F_3 = 1000$. 76
3.11 Illustration de l'adaptation de la grille à un récepteur dont les zones B1 et B2 recouvrent toute la surface cultivée en OGM. 77
3.12 Illustration de l'adaptation de la grille à un récepteur dont les zones B2 et C recouvrent toute la surface cultivée en OGM. 77

4.1 Plans d'échantillonnage utilisés pour les 2 expérimentations *mono-source* du jeu d'entrainement. Les zones OGM sont représentées en rouge. 83
4.2 Plans d'échantillonnage utilisés pour l'expérimentation *mono-source* du jeu de validation. (Source: Palaudelmàs et al. (2012)) 83
4.3 Traces des chaînes de Markov et distributions a posteriori des paramètres du modèle $GPExpoA$, estimés sur les données *mono-source*. 86

4.4 Traces des chaînes de Markov et distributions a posteriori des paramètres du modèle $GZExpoA$, estimés sur les données *mono-source*. 87
4.5 Traces des chaînes de Markov et distributions a posteriori des paramètres du modèle $IPExpoA$, estimés sur les données *mono-source*. 88
4.6 Traces des chaînes de Markov et distributions a posteriori des paramètres du modèle $IZExpoA$, estimés sur les données *mono-source*. 89
4.7 Boxplots du nombre de grains bleus par épi en fonction de la distance à la source OGM la plus proche. Les croix rouges correspondent à la moyenne du nombre de grains bleus pour chaque classe de distance. 94
4.8 Boxplots des taux de pollinisation croisée observés et prédits (par le modèle GZExpoA) en fonction de la distance à la source OGM la plus proche. 95
4.9 Boxplots des taux de pollinisation croisée observés et prédits (par le modèle GZExpoB) en fonction de la distance à la source OGM la plus proche 96
4.10 Boxplot des moyennes (A, B) et des écarts-types (C, D) du taux de pollinisation croisée prédit en fonction de la distance à la source OGM la plus proche. Graphes A et C (respectivement B et C): prédictions issues du modèle $GZExpoA$ (respectivement $GZExpoB$). Les croix rouges correspondent à la moyenne (A, B) et à l'écart type (C, D) du taux observé par intervalle de distance. 97
4.11 Moyennes a posteriori des prédictions moyennes par micro-parcelle, en fonction des observations correspondantes sur les données d'ajustement. À Gauche (resp. à droite), prédiction du modèle $GZExpoA$ (resp. $IZExpoA$). 98
4.12 Moyennes a posteriori des prédictions moyennes par micro-parcelle, en fonction des observations correspondantes sur les données de validation. À Gauche (resp. à droite), prédiction du modèle $GZExpoA$ (resp. $IZExpoA$). 98
4.13 Plans d'échantillonnage utilisés pour les prélèvements dans la zone de Foixà (Messeguer et al., 2006). 100
4.14 Plan d'échantillonnage réalisé et nombre de prélèments pour chaque parcelle et chaque année dans la zone de Foixà (Messeguer et al., 2006). 101
4.15 Traces des chaînes de Markov et distribution a posteriori des paramètres du modèle $GNExpoA$ estimés sur les données *multi-sources*. 104
4.16 Traces des chaînes de Markov et distribution a posteriori des paramètres du modèle $INExpoA$ estimés sur les données *multi-sources* 105
4.17 Médianes a posteriori des prédictions en fonction des observations correspondantes. 109
4.18 Courbe ROC pour un seuil de décision de 0.9%. 111
4.19 Cercle des corrélations entre critéres 118
4.20 Contributions des critères aux deux premières composantes principales et indices de sensibilités des facteurs. En abscisse des deux graphiques du haut: 1: CRPS, 2: r, 3: RMSE, 4: R^2, 5: AUC. Dans les deux graphiques du bas: indices de sensibilité des facteurs sur les deux premières composantes. La partie noire indique l'indice de sensibilité du premier ordre, l'ensemble de la barre indique l'indice de sensibilité totale. DF:Fonction de dispersion, ObsMod: Modèle d'observations, MuDist:Modèle pour μ, E: Ajustement ou Validation 119

Liste des tableaux

1.1 Distances d'isolement (en m) entre OGM et non-OGM proposées dans différents Etats membres (hors production de semences) . . . 11

2.1 Tableau récapitulatif des situations de prédictions . . . 33
2.2 Table des caractéristiques des jeux de données rassemblés pour le projet SIGMEA. (Extrait traduit du livrable D4.25 du projet PRICE, disponible en intégralité à l'annexes 5) . . . 48

4.1 Plan d'expérience utilisé pour la comparaison des modèles. . . . 90
4.2 Distribution *a priori* des paramètres des modèles d'observations. $\mathcal{U}$: loi uniforme; $InvGamma$: loi Gamma inverse. . . . 90
4.3 Caractéristiques des distributions *a priori* des paramètres des fonctions de dispersion. FD: fonction de dispersion, ET: écart-type, CV: coefficient de variation (= ET/Moyenne). . . . 91
4.4 Valeurs des critères calculées entre les observations et les prédictions correspondantes du modèle $IPExpoA$ ajusté sur les données *mono-source*. . . . 108
4.5 Comparaison entre observations et prédictions pour les 27 parcelles échantillonnées à Foixà entre 2004 et 2008. . . . 110
4.6 Valeurs des critères pour toutes les combinaisons de facteur en ajustement sur Montargis 98 . . . 120
4.7 Valeurs des critères pour toutes les combinaisons de facteur en validation (Montargis 99 et Mas Cebria) . . . 121
4.8 Valeurs des critères pour toutes les combinaisons de facteur en ajustement sur Montargis 99 . . . 122
4.9 Valeurs des critères pour toutes les combinaisons de facteur en validation sur Montargis 98 et Mas Cebria . . . 123
4.10 Valeurs des critères pour toutes les combinaisons de facteur en ajustement sur Montargis 98 et Montargis 99) . . . 124
4.11 Valeurs des critères pour toutes les combinaisons de facteur en validation sur Mas Cebria . . . 125

Introduction

Dans le contexte actuel d'innovations technologiques et d'évolution des problématiques agro-environnementales, les acteurs confrontés à la prise de décisions publiques font de plus en plus appel aux scientifiques pour réaliser l'évaluation des risques associés à ces décisions. En effet, différentes questions, au cœur des problématiques actuelles, telles que la diminution des intrants, la durabilité des résistances variétales, l'introduction de nouvelles technologies (OGM, agriculture de précision, clonage animal, etc.) ou le changement climatique (risque d'invasions, déplacement de niches écologiques, ...) amènent à considérer la notion d'évaluation des risques associés aux décisions pour répondre aux attentes de la société.

Ces questions présentent plusieurs caractéristiques communes et impliquent notamment des processus de dispersion, souvent à de larges échelles spatiales, soumis à l'influence du climat mais aussi aux interventions éventuelles d'acteurs sur le terrain pouvant être très nombreux et ayant des intérêts potentiellement contradictoires. L'évaluation des risques associés à ces processus nécessite donc la définition de nouvelles méthodes pour permettre la prise en compte de ces composantes multiples. De plus la grande incertitude qui existe, aussi bien sur les réactions du processus aux variations (volontaires ou non) des paramètres environnementaux, que sur les variations possibles de tels paramètres, plaide pour son intégration systématique dans l'aide à la décision et donc dans toute évaluation des risques. Les problématiques agro-environnementales actuelles requièrent donc des méthodes spécifiques pour aider les acteurs concernés à prendre des décisions en évaluant les risques associés et en intégrant l'incertitude à l'évaluation de ces risques.

Le débat autour des OGM en Europe et la question d'une possible coexistence entre cultures OGM et non-OGM (Beckie and Hall, 2008; Messéan et al., 2009) représente un bon exemple de ces problématiques. Ici, la question est de savoir s'il est possible de cultiver des OGM sans pénaliser les cultures conventionnelles voisines en raison des risques de flux de gènes à l'échelle des paysages. L'enjeu est donc d'évaluer ces risques afin de permettre la coexistence, tout en intégrant les incertitudes associées aux facteurs agissant sur le processus de flux de gènes. Les règles actuellement envisagées par les décideurs publics reposent uniquement sur des distances de séparation fixées à l'échelle nationale et ne permettent pas d'adaptation à la diversité des situations possibles. De plus elles ne prennent pas en compte les incertitudes sur les variations potentielles des paramètres environnementaux. Or, d'une part les connaissances acquises dans le domaine des flux de gènes sont nombreuses et doivent permettre de proposer des solutions pour l'adaptation des mesures de coexistence à des cas spécifiques, et d'autre part l'incertitude doit être intégrée à l'évaluation des risques qui permet d'établir ces règles.

Le travail de thèse présenté dans ce mémoire a notamment pour objectif de développer une approche par modélisation, alternative aux règles trop rigides actuellement envisagées, permettant d'adapter les mesures à prendre afin de garantir la coexistence, à la diversité des situations rencontrées en pratique. La démarche repose sur la conception d'un modèle prédictif, permettant la simulation des flux de pollen entre champs cultivés et la prédiction du taux de pollinisation croisée dans les champs non OGM, en valorisant au maximum l'information locale lorsqu'elle est disponible. L'inférence des paramètres du modèle est réalisée dans un cadre bayésien de manière à intégrer de façon explicite aussi bien la variabilité intrinsèque du système étudié que les approximations et incertitudes sur les processus modélisés. Ce cadre probabiliste offre l'avantage non seulement de mieux

prendre en compte l'incertitude dans l'évaluation des risques mais aussi de permettre une mise à jour aisée des valeurs des paramètres du modèle à mesure que de nouvelles données sont disponibles. Outre la prédiction du taux de pollinisation croisée, le modèle doit permettre de remplir des objectifs tels que *i)* classer des lots de récoltes conventionnelles vis-à-vis d'un seuil afin d'en déterminer le type (OGM ou non OGM) ; *ii)* établir des stratégies d'échantillonnage dans les dispositifs de surveillance des OGM afin d'en optimiser les coûts. De plus ce travail de thèse, mené en concertation avec le projet européen PRICE, a eu pour objectif de contribuer au développement d'un volet opérationnel de la modélisation, sous la forme d'un outil d'aide à la décision informatique.

Ce mémoire est divisé en cinq chapitres et contient quatre annexes. Le premier chapitre est dédié à la description de la problématique de la coexistence entre cultures OGM et non OGM et des principaux résultats acquis dans ce domaine par de précédents travaux, ainsi qu'à la présentation détaillée des objectifs de la thèse. Le deuxième chapitre est consacré à la présentation des éléments de conceptualisation du problème ; nous y présentons tous d'abord les situations ou scénarios de prédiction, puis le cadre statistique adopté et ses particularités, enfin les différents jeux de données existants, leurs caractéristiques et leur diversité. Le troisième chapitre spécifie les éléments constitutifs d'un modèle de prédiction des taux de pollinisation croisée tel que défini dans cette thèse. Nous y abordons également les approximations développées pour améliorer les performances des algorithmes de prédiction en termes de temps de calcul et rendre possible l'estimation des modèles. Le quatrième chapitre présente les résultats relatifs aux estimations des modèles décrits au chapitre précédent. On y expose dans un premier temps les différents cas d'estimation et les résultats qui en découlent. Puis, les méthodes employées pour l'évaluation et la sélection de modèles sont présentées. Enfin, le cinquième et dernier chapitre est consacré à la discussion ; nous y présentons les conclusions qui peuvent être tirées de ce travail de thèse et nous intéressons aux perspectives qu'un tel travail permet de dessiner.

Les quatre premières annexes de ce mémoire comportent les publications relatives à cette thèse. La première annexe est un article qui décrit le cadre méthodologique qui y développé et les premiers résultats obtenus. Il a été soumis à *Environmental Modelling & Software* en décembre 2014. La deuxième annexe est un article paru dans *AgBioForum* en 2014 qui présente une application spécifique et originale des méthodes décrites dans le premier article. La troisième annexe est un article soumis à *Risk Analysis* en décembre 2014. Cet article présente une extension méthodologique ainsi qu'une application spécifique et originale relative au cadre et aux résultats du premier article. La quatrième annexe correspond aux actes d'une communication à la 6ème conférence internationale sur la coexistence entre culture OGM et non OGM (GMCC 2013). Ces actes décrivent la définition de situations de prédiction pour les modèles ainsi que l'implémentation et la mise à disposition de l'outil d'aide à la décision élaboré, en partie, dans le cadre de cette thèse. La cinquième et dernière annexe est un livrable rédigé dans le cadre du projet Européen PRICE, il fait l'inventaire des jeux de données disponibles dans la littérature et consolidés dans le cadre du projet européen SIGMEA (Messéan et al., 2009).

Chapitre 1

Problématique

Table des matières

1 Contexte et enjeux de la coexistence 7

1.1 État des lieux des variétés de maïs OGM et des surfaces cultivées 7

1.2 Contexe législatif 8

1.2.1 En Europe 8

1.2.2 En France 12

2 Biologie et modélisation des flux de gènes sur le maïs 14

2.1 Résultats expérimentaux sur la pollinisation croisée chez le maïs 15

2.2 Approche prédictive pour l'estimation du taux de pollinisation croisée 17

3 Coexistence, modélisation et incertitudes: le projet PRICE . 20

3.1 La modélisation au cœur du processus de décision 21

3.2 Des prédictions? Oui, mais avec une incertitude associée! . . . 21

4 Objectifs de la thèse 22

L'objectif général des travaux de recherche sur la coexistence est d'être capable de proposer des configurations techniques permettant de respecter, à la récolte, différents seuils préalablement définis (Meynard and Le Bail, 2001). Les travaux qui sont présentés dans cette thèse s'inscrivent pleinement dans cet objectif général avec une attention particulière portée à la prise en compte des incertitudes pour aider la prise de décision concernant ces configurations techniques.

Dans un premier temps, un état des lieux des mises en cultures de variétés transgéniques et les contextes législatif et technique de recherche seront présentés. Le processus de flux de gènes ainsi que certains résultats expérimentaux seront ensuite décrits. Enfin, le projet européen PRICE, dans lequel s'insère la thèse, sera présenté, ainsi que les hypothèses de départ et les principaux résultats attendus.

1 Contexte et enjeux de la coexistence entre filières OGM et non-OGM

1.1 État des lieux des variétés de maïs OGM et des surfaces cultivées

Le maïs (*Zea mays* L. ssp. *mays*) est la deuxième plante génétiquement modifiée (PGM) en surface cultivée dans le monde après le soja, avec 55,2 millions d'hectares représentant 30% des PGM en 2014. Les variétés de maïs transgénique cultivées dans le monde possède deux types de traits, soit isolés soit en combinaison : la tolérance aux herbicides (e.g., Soja Roundup Ready ou le T25) et la résistance aux insectes (e.g. MON810 de Monsanto, Bt11 de Syngenta et 1507 de Pioneer). Les variétés MON810, Bt11 ont été modifiées par l'insertion du gène cry de la bactérie du sol *Bacillus thuringiensis (Bt)* qui exprime la protéine CryA1b. Elle produit une δ-endotoxine ciblant certains lépidoptères et particulièrement la pyrale (*Ostrinia nubilalis* Hbn.).

La France a été le premier pays d'Europe à cultiver des OGM, avec l'inscription en février 1998 de trois variétés de maïs transgéniques contenant l'événement Bt176 au catalogue national. Le T25 et le MON810 ont été autorisés en août 1998 mais le T25 n'a jamais été cultivé. Le Bt176 a été retiré de la commercialisation par l'entreprise Novartis (devenu depuis Syngenta) (CE, 2007a), Il ne reste donc plus que le maïs MON810.

En 1998, les cultures commerciales de maïs transgénique représentaient en France 1 500 ha mais ne se sont pas développées du fait de l'instauration d'un moratoire de fait en 1999. En 2005, les cultures commerciales de maïs officiellement déclarées au ministère de l'agriculture représentaient 492,8 ha. En 2007, déclarer les cultures OGM devient obligatoire (CE, 2007a), et les surfaces de maïs cultivé déclarées atteignent 22 135 ha Par ailleurs, jusqu'en 2007, de nombreux essais d'OGM à des fins de recherche en plein champ ont été autorisés (plusieurs centaines au total). Depuis, leur nombre a fortement diminué, en raison de l'absence de perspectives de développement et à la suite de destructions. Il n'y a désormais plus aucun essai. Par ailleurs, en janvier 2008 le gouvernement français interdit la culture de MON 810 en invoquant la clause de sauvegarde, confirmée par une loi adoptée en 2014.

1.2 Contexe législatif

1.2.1 En Europe

Le concept de coexistence est défini comme *"la capacité des agriculteurs à faire un choix effectif entre cultures génétiquement modifiées, biologiques et conventionnelles dans le respect des obligations légales en matière d'étiquetage et/ou de normes de pureté"* (CE, 2003a).

Pour garantir cette capacité, il est nécessaire de déterminer les mesures appropriées afin de permettre aux producteurs d'exercer ce choix dans les zones où des OGM sont cultivés. Les mesures de coexistence visent à éviter la présence accidentelle d'OGM dans d'autres produits, afin de prévenir l'incidence des mélanges entre cultures génétiquement modifiées et autres cultures et le préjudice économique potentiel.

Dans l'Union européenne, la mise sur le marché de variétés OGM est encadrée par la Directive 2001/18 relative à la dissémination volontaire d'organismes génétiquement modifiés dans l'environnement (CE, 2001a) et par le règlement 1829/2003 (CE, 2003b) concernant les denrées alimentaires et les aliments pour animaux. Ces législations ont été mises en place pour protéger la santé humaine et l'environnement, en proposant un cadre unifié pour l'évaluation des risques d'une potentielle mise sur le marché. Elles assurent également la libre circulation des produits au sein de l'Union et doivent aussi garantir la liberté de choix des consommateurs entre différents types de produits.

Le règlement 1830/2003 (CE, 2003c) concernant la traçabilité et l'étiquetage indique que les OGM et les produits pour l'alimentation humaine et animale qui en sont dérivés doivent être clairement étiquetés de façon à laisser le choix libre pour le consommateur. Cependant, des exceptions relatives aux traces d'OGM dont la présence serait fortuite ou techniquement inévitable sont prévues. Un seuil de 0.9% est fixé en dessous duquel, et dans des conditions spécifiques, l'étiquetage n'est pas obligatoire (CE, 2003c). Ce seuil est aussi appliqué aux produits destinés à l'alimentation humaine et animale (CE, 2003b), qu'ils soient issus de l'agriculture conventionnelle ou biologique (CE, 2007b). Il reste valable dans les cinq ans que suivent un potentiel retrait d'autorisation(CE, 2007a).

Contrairement à la législation précédente (Directive 90/220/CE révisée en 2000), ce seuil reste applicable que le transgène ou les protéines produites soient détectables ou non, ce qui implique l'étiquetage de produits raffinés comme les huiles, et la mise en place de procédures de traçabilité systématiques et tout au long des différentes filières de production.

Bien que des recommandations quant à l'échantillonnage et la détection des OGM aient été formulées (CE, 2004; CE, 2011a), il reste une ambiguïté sur l'unité dans laquelle ce seuil peut-être exprimé : pourcentage de grains OGM, résultats issus de PCR exprimés en pourcentage de masse ou de génome, ce qui génère, en plus des différences de fiabilité entre ces méthodes de mesures (Anklam et al., 2002; Heinemann et al., 2004; Weber et al., 2007; Macarthur et al., 2010; Trapmann et al., 2010; Njontie et al., 2011), des problèmes potentiels dans la définition de mesures de coexistence (Haut Conseil des Biotechnologies, 2011; Paul et al., 2012).

Pour l'agriculture biologique et pour certaines filières de l'alimentation humaine (notamment semoule et amidon), des seuils plus stricts peuvent être exigés par les opérateurs. En l'occurence, c'est souvent le seuil de quantification de la PCR qui est demandé dans les récoltes (0.1%) voire, dans certains cas, le seuil de détection soit 0.01% (Meynard and Le Bail, 2001; Raveneau, 2005). Ces cas spécifiques sont considérés comme des standards privés et ne relèvent donc pas de la législation en vigueur. Enfin, il n'existe pas de seuils de tolérance officiels pour les évènements non autorisés dans l'Union européenne. D'après la législation européenne, de tels OGM ne peuvent ni être cultivés ni vendus dans l'UE.

Les mesures de coexistence ne concernent donc que des OGM autorisés et ne relèvent en aucune façon de la gestion des risques dans le domaine de la santé (humaine ou animale) ou de l'environnement. Leur objectif principal est de limiter au maximum les sources potentielles de mélange entre des variétés OGM et non-OGM et, par conséquent, l'impact économique pour les agriculteurs cultivant des variétés non-OGM (*i.e.* déclassement de la récolte).

La Commission européenne a prescrit des recommandations définissant les lignes directrices pour l'établissement de mesures de coexistence "de la semence au silo" (CE, 2003a). Voici les principes généraux édictés pour la définition de stratégies nationales :

- fonder ses décisions sur des résultats scientifiques ;
- s'appuyer sur des méthodes et pratiques de ségrégation existantes ;
- respecter le principe de proportionnalité ; autrement dit les mesures doivent :
  - être efficaces,
  - avoir un bon rapport coût/efficacité,
  - ne pas dépasser ce qui est techniquement nécessaire pour respecter le seuil légal,
  - être adaptées aussi bien au contexte local ou régional qu'à l'espèce cultivée ;
- privilégier les mesures à l'échelle de l'exploitation agricole ou la coordination avec des exploitations attenantes et limiter les mesures supplémentaires à l'échelle régionale à des cas particuliers (espèce, type de production) ;
- avoir des mesures spécifiques en fonction des espèces et variétés mais aussi du contexte régional (climat, topographie, rotations, part d'OGM dans la sole).

Compte tenu de la diversité des systèmes de production, des structures d'exploitation et des conditions économiques et physiques dans l'agriculture de l'UE, il a été décidé que les mesures de coexistence à mettre en place relèveraient de la compétence de chaque État membre. En 2009, 15 des 27 États membres avaient mis en place des lois concernant la coexistence (CE, 2009), qui consistaient principalement en mesures individuelles à mettre en œuvre, à l'échelle de la parcelle, par l'agriculteur cultivant des variétés transgéniques. La plus courante est la définition des distances d'isolement permettant de respecter le seuil légal d'étiquetage fixé à 0.9% (CE, 2003c) en sortie de champ (Tableau 1.1). Ces distances peuvent être complétées ou remplacées par des zones tampons où des variétés non-OGM sont semées autour du champ OGM.

Par ailleurs, le Conseil de l'Union a confié en 2006 un mandat spécifique à la Commission européenne pour engager des travaux supplémentaires concernant la coexistence. Il consiste notamment à sélectionner, en coopération étroite avec les États membres et

les différentes parties prenantes (opérateurs, organisations non-gouvernementales), un ensemble de bonnes pratiques pour la ségrégation OGM/non-OGM et à définir des lignes directrices par espèce. Celles-ci doivent être suffisamment flexibles pour pouvoir être adaptées au contexte de chaque État membre. Un groupe de travail a été mis en place par la Commission sous l'égide du "European Coexistence Bureau (EcoB)". Il a édicté de bonnes pratiques de coexistence chez le maïs (Riznov and Rodríguez-Cerezo, 2014) et aborde désormais le cas du soja. Cet appui méthodologique s'est révélé nécessaire notamment par le fait que la gamme des distances proposées par les différents pays de l'Union est extrêmement large (voir tableau 1.1) et considérée comme très supérieure à la variabilité qui pourrait être expliquée par les différences de contextes agro-climatiques.

Etat membre	Maïs conventionnel	Maïs biologique
Allemagne	150	300
Autriche	∅	∅
Belgique	∅	∅
Danemark	150	150
Espagne	∅	∅
Finlande	∅	∅
France	∅	∅
Hongrie	400	400
Irlande	50	75
Italie	∅	∅
Lettonie	200	200
Lituanie	200	200
Luxembourg	600	600
Pays-Bas	25	250
Portugal	200	300
Rép. Tchèque	70	200
Roumanie	200	200
Slovaquie	200	300
Suède	50	50

TABLE 1.1: Distances d'isolement (en m) entre OGM et non-OGM proposées dans différents Etats membres (hors production de semences)

1.2.2 En France

La loi du 25 juin 2008 retranscrit la Directive européenne 2001/18 en Droit français (République Française, 2008). Elle reprend les recommandations sur l'autorisation des OGM et met en place des instances spécifiques :

- Le Haut Conseil des Biotechnologies (HCB) qui se compose d'un comité scientifique et d'un comité économique, éthique et social. Il "*a pour missions d'éclairer le Gouvernement sur toutes questions intéressant les organismes génétiquement modifiés ou toute autre biotechnologie et de formuler des avis en matière d'évaluation des risques pour l'environnement et la santé publique que peuvent présenter l'utilisation confinée ou la dissémination volontaire d'OGM ainsi qu'en matière de surveillance biologique du territoire [...]*".
- Un Comité de surveillance biologique du territoire qui a un rôle consultatif sur les protocoles et méthodologies d'observations nécessaires à la mise en œuvre de la surveillance biologique du territoire ainsi que sur les résultats de cette surveillance. Dans cette optique de transparence, un registre national indiquant la nature et la localisation des parcelles d'OGM est également mis en place.

Cette loi permet la définition d'un cadre réglementaire pour la coexistence entre différents types de cultures et la reconnaissance de la liberté de produire et de consommer avec ou sans OGM. Elle stipule que la mise en culture, la récolte et le transport des OGM autorisés sont soumis au respect des conditions techniques relatives notamment aux distances entre les différents types de cultures ou à leur isolement, afin d'éviter la présence dite fortuite d'OGM dans d'autres productions.

Ces distances doivent permettre que la dissémination entre cultures n'autorise pas le dépassement du seuil règlementaire fixé à l'échelle de l'Union européenne.

La loi institue de surcroît la responsabilité individuelle du producteur d'OGM en regard du préjudice économique résultant de la présence accidentelle d'OGM dans la récolte d'un autre agriculteur ainsi que l'obligation de souscrire à une garantie financière couvrant cette responsabilité. Elle stipule également les sanctions encourues en cas de non respect du cadre technique : destruction à la charge du producteur d'OGM et sanctions pénales et financières.

Enfin, dans la loi, la notion de "sans organismes génétiquement modifiés" se réfère à la définition communautaire. Cependant, dans l'attente d'une définition effective au niveau européen, le HCB a été saisi pour émettre un avis qui servirait à fixer le seuil correspondant au niveau réglementaire. Celui-ci est composé d'une recommandation du comité économique, éthique et social du HCB, et part du constat que, compte tenu de la coexistence de filières et du recours à des importations de matières premières agricoles, la notion de "sans OGM", intuitivement liée à une absence totale d'ADN transgénique, doit être remplacée par une définition fondée sur le respect d'un seuil maximal de présence d'ADN transgénique. L'avis précise que :

- Pour les produits végétaux, une mention "sans OGM" devrait être réservée aux produits contenant moins de 0.1% d'ADN transgénique ;
- Pour les produits animaux, une mention "nourri sans aliments OGM" ou "issu d'animaux nourris sans aliments OGM" devrait être réservée aux produits issus

d'animaux nourris avec des aliments dans lesquels la présence d'ADN transgénique est inférieure à 0.1%.

Ce seuil a été officialisé en janvier 2012 dans un décret portant sur les règles facultatives d'étiquetage pour denrées alimentaires issues de filières qualifiées "sans OGM" (République Française, 2012). Depuis juillet 2012, il existe donc un seuil d'étiquetage "OGM" pour les produits contenant plus de 0.9% d'ingrédients issus d'OGM, un seuil maximal de 0.1% pour les produits "sans OGM" et, de fait, une *zone grise* entre ces deux limites. Cette structuration de l'étiquetage au niveau du produit fini a des conséquences en début de filière de production, c'est-à-dire au champ, compte tenu des risques cumulatifs de mélanges existants tout au long d'une filière (voir figure 1.1) et des difficultés techniques afférentes pour respecter des seuils bas. Autrement dit, pour assurer un seuil de 0.9% dans le produit fini on ne pourra pas se contenter d'assurer ce seuil à la récolte, il faudra, à ce stade, un seuil plus faible tenant compte de facteurs de risques en aval (post récolte) de la filière.

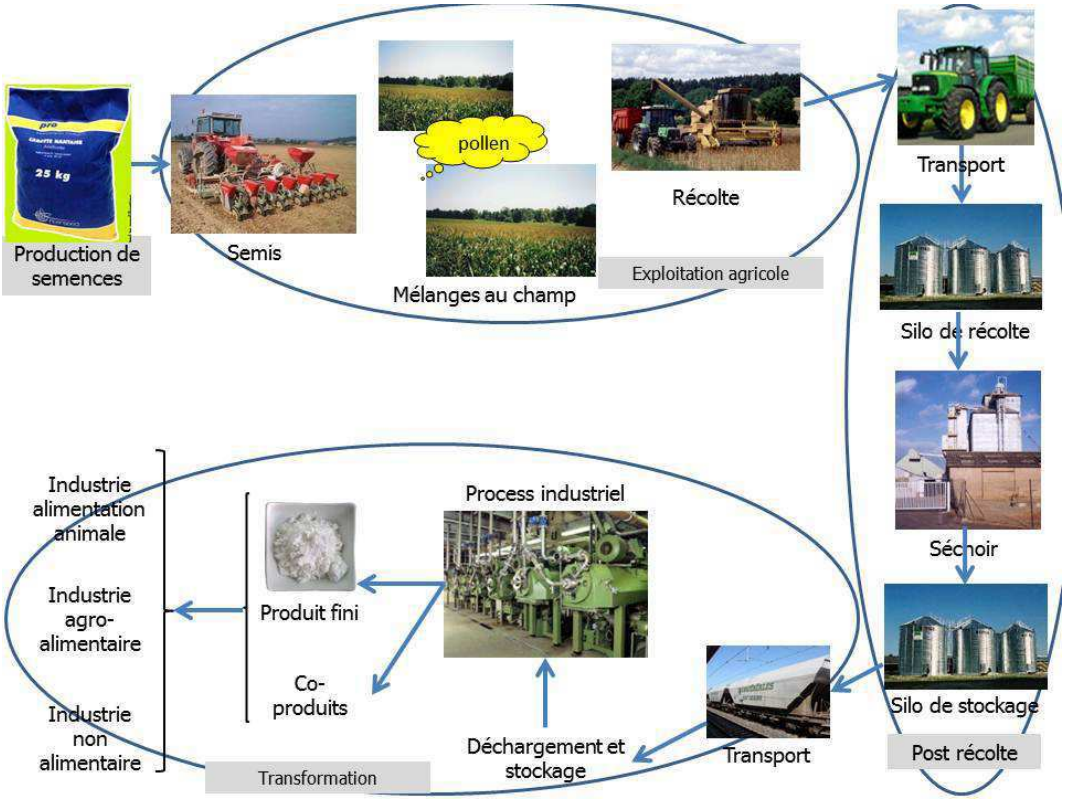


FIGURE 1.1: Étapes à prendre en compte pour l'étude de la coexistence entre différents types de filières de production végétale (*Source* : F. Angevin, INRA)

2 Biologie et modélisation des flux de gènes sur le maïs

Dans l'optique de quantifier le taux de pollinisation croisée dans les récoltes des champs de maïs non-OGM, il convient d'identifier tout d'abord les sources potentielles de croisement entre différents types de culture. Celles-ci sont liées à la dispersion de graines ou de pollen, dans le temps et dans l'espace (Colbach and Clermont-Dauphin, 2001).

Les impuretés liées aux graines peuvent avoir différentes origines :

- plantes déjà présentes dans les champs (repousses de cultures précédentes de la même espèce) ;
- semences OGM déjà présentes dans les lots de semences conventionnelles ;
- graines déposées par les équipements agricoles (semoir, moissonneuse) ;
- pertes de graines de variétés sauvages de la même espèce que la culture et présentes dans les zones non cultivées (bordure, jachère, friches, ...).

Les impuretés liées au pollen proviennent de la fécondation croisée de plantes non-OGM avec

- des champs voisins contenant des cultures de la même espèce ;
- des repousses de la même espèce ;
- des variétés sauvages.

Il faut noter que la pollinisation croisée ne peut avoir lieu que pendant une période limitée durant laquelle émetteurs et récepteurs sont simultanément en floraison.

L'importance relative des différentes sources dépend avant tout de la biologie de l'espèce considérée (ploïdie, durée de floraison, ...), de celle de ses graines (dormance, durée de vie, ...) mais également des conditions environnementales (climat, configuration spatiale, topographie, ...) et agronomiques (système de culture : rotations, choix variétal, date de semis).

Le maïs est une plante essentiellement allogame (Purseglove, 1972) dont le pollen est disséminé en très grande majorité par le vent (Bateman, 1947a; Treu and Emberlin, 2000). Il existe divers travaux décrivant l'activité de butinage des abeilles dans les champs de maïs (Ibrahim and Selim, 1972; Mason and Tracewski, 1982; Aly and Hassan, 1999), cependant il apparaît que les fleurs de maïs femelles sont très peu attractives et il semble avéré que la contribution potentielle des insectes auxiliaires à la pollinisation provient d'un tranfert indirect lié à la mise en suspension du pollen dans l'air, comme cela a pu être montré sur d'autres espèces (Pierre et al., 2002).

Les panicules (fleurs mâles) et les soies (fleurs femelles) sont séparées sur la plante et la plupart des variétés cultivées expriment de la protandrie, c'est-à-dire que la floraison mâle précède la floraison femelle, ce qui favorise l'allogamie (Bateman, 1947b; Struik and Makonnen, 1992; Emberlin et al., 1999). Le pollen de maïs est relativement lourd et sa dispersion décroît rapidement à longue distance (Bateman, 1947b; Raynor et al., 1972). Toutefois la dispersion à longue distance peut également intervenir (Jones and Brooks, 1950; Byrne and Fromherz, 2003; Bannert and Stamp, 2007). Une partie du pollen de maïs peut effectivement être emportée par des mouvements convectifs en altitude, se déposer

à de très longues distances et donner lieu à une pollinisation efficace, uniquement si le pollen est encore vivant (Brunet et al., 2009).

La plupart des variétés de maïs cultivées sont des hybrides et leurs semences sont produites dans des champs où la quantité de pollen émise est sensiblement plus faible que dans des parcelles destinées à la production de graines pour la consommation ou la transformation. Sans rentrer dans les détails techniques de la production de semences, il est à noter que cette production est particulièrement sensible à la pollinisation croisée (Ireland et al., 2006; Messéan et al., 2006).

En ce qui concerne le risque de présence d'OGM au travers de la dispersion des graines, des repousses de maïs peuvent être observées l'année suivant une culture de maïs lorsque les hivers sont doux, principalement dans le Sud de l'Europe (Palaudelmàs et al., 2009). Dans des suivis effectués en Espagne, leur densité est apparue variable, dépendante du type d'irrigation pratiqué, et pouvait atteindre 10% des plantes présentes dans un champ. Ces repousses avaient une faible vigueur, produisaient rarement des épis, mais émettaient du pollen pouvant féconder les plantes alentour. Les taux de pollinisation croisée mesurés se sont avérés faibles avec un maximun de 0.16% dans une situation considérée comme à risque (Palaudelmàs et al., 2009). Dans les systèmes de culture français, le taux d'humidité à la récolte réduit les risques de pertes de graines de l'épi. Par ailleurs, les pratiques agricoles majoritairement utilisées, notamment le labour ou le froid hivernal limitent l'occurence des repousses dont l'impact sur le taux de pollinisation croisée peut dès lors être négligé.

Concernant les sources de mélanges liées au matériel agricole, elles ont été estimées à l'occasion de différentes études (Angevin et al., 2002; Messéan et al., 2006) et sont considérées comme faibles par rapport aux risques liés à la pollinisation entre champs ou au sein d'un même champ, suite à la présence d'impuretés dans les lots de semences (Dietiker et al., 2011; Njontie et al., 2011; Paul et al., 2012).

2.1 Résultats expérimentaux sur la pollinisation croisée chez le maïs

Les flux de gènes entre champs de maïs et plus particulièrement les flux de pollen ont été largement étudiés au cours des dernières décennies (Bateman, 1947a,b; Jones and Brooks, 1950; Raynor et al., 1972) afin d'améliorer la pureté génétique dans les productions de semences (Bateman, 1947a,b). Les résultats obtenus dans ces études ainsi que les connaissances empiriques sur le sujet ont abouti à l'établissement de cahiers des charges pour ces filières (Luna et al., 2001; Ireland et al., 2006). De plus, ils ont permis de déterminer les principaux facteurs influençant ce processus :

- la distance entre champs émetteurs et récepteurs (Bateman, 1947a; Raynor et al., 1972; Langhof et al., 2010; Rühl and Langhof, 2011) ;
- la synchronisation des floraisons (Bateman, 1947b; Langhof et al., 2010) ;
- les quantités relatives de pollen émises par les variétés (Goggi et al., 2007; Dietiker et al., 2011) ;
- les conditions climatiques, notamment la vitesse et la direction du vent (Lonnquist

and Jugenheimer, 1943; Bateman, 1947b; Raynor et al., 1972), et dans une moindre mesure la pluviométrie (Angevin et al., 2008).

Selon Ireland et al. (2006), ces facteurs sont classables en deux composantes interagissant entre elles dans l'isolement reproductif du maïs :

- l'isolement biologique (période et durée de floraison, densité de pollen émise, quantité totale de pollen émise et éventuellement capacité du pollen à féconder) ;
- l'isolement physique, qui relève de facteurs limitant la capacité du pollen à entrer dans le champ récepteur, c'est-à-dire la distance entre champs, l'orientation des parcelles entre elles et leurs tailles relatives.

Lors de la mise en culture des premières variétés génétiquement modifiées, une synthèse des connaissances disponibles sur le sujet a été publiée (Ingram, 2000). Elle fournissait un cadre méthodologique d'interprétation des données antérieures selon les facteurs affectant la pollinisation croisée et soulignait la difficulté d'extrapoler à des situations de production les résultats d'essais mis en place pour d'autres objectifs (établissement de schémas de production de semences ou estimation des probabilités de croisements interspécifiques sur des plantes isolées), dans des conditions favorables au phénomène étudié (e.g. synchronisation des floraisons, champ récepteur sous le vent) pouvant amener à des surestimations des distances d'isolement dans le cas de la production de maïs pour la consommation ou la transformation. Les distances d'isolement champ à champ établies par l'auteur pour le respect des seuils 0.1, 0.5 et 1% sont basées sur les résultats d'un essai dans une situation de risque maximal : vent fort et conditions sèches pendant l'expérimentation (Jones and Brooks, 1950). De plus, l'unique essai sur lequel sont basées les distances d'isolement a été mené avec des variétés de maïs populations et non des hybrides ce qui diminue plus encore son réalisme. Bien qu'Ingram les qualifie de robustes, il ne garantit pas que ces recommandations (200m et 300m pour 1 et 0.5% respectivement) soient suffisantes dans tous les cas de figure et conseille de nouveaux essais dans différents contextes pédoclimatiques de manière à s'assurer de leur pertinence.

Le développement des variétés transgéniques et de la problématique de coexistence a abouti à la mise en place de plusieurs expérimentations ayant pour but de déterminer des règles techniques permettant de limiter le taux de pollinisation croisée dans les champs non-OGM. Dans le cadre du projet européen SIGMEA, une vingtaine de jeux de données de flux de gènes entre champs de maïs ont été rassemblés et analysés. La plupart de ces données étaient issues d'expérimentation sur la pollinisation croisée champ à champ, à courte distance, avec comme émetteur des maïs OGM ou des variétés de maïs spécifiques exprimant un marqueur de couleur dominant (blanc ou bleu). Voici les principaux enseignements tirés de ces études :

- une forte décroissance de la dispersion de pollen avec la distance ;
- une très grande variabilité des taux de pollinisation observés en fonction des conditions expérimentales et environnementales ;
- la persistence d'une pollinisation croisée à longue distance à des taux faibles mais non nuls ;
- un effet très marqué de la direction du vent.

Les résultats des essais champ à champ ont par ailleurs fait l'objet de méta-analyses dans le but de définir des règles de coexistence (Gustafson et al., 2006; Sanvido et al.,

2008; Riesgo et al., 2010). Ces études établissent des recommandations prenant la forme de distances d'isolement estimées pour des situations dites à risques. En effet, ce type d'analyse statistique ne permet pas d'analyser en profondeur les facteurs influençant le flux de gènes. Il est donc très difficile voire impossible d'adapter ces mesures de coexistence à des contextes agro-climatiques spécifiques.

Les essais au champ sont utiles afin de déterminer et comprendre les phénomènes en jeu dans les flux de gènes. Cependant, l'échelle spatiale, l'effet du climat, la dépendance à certaines pratiques agricoles et la nécessité de tester des combinaisons de techniques d'isolement rendent difficiles la généralisation de l'approche expérimentale pour définir des règles de coexistence adaptées aux différentes régions de production du maïs.

2.2 Approche prédictive pour l'estimation du taux de pollinisation croisée

Dans le cadre global d'étude des flux de gènes, on trouve dans la littérature quatre grands types d'objectifs (Lavigne et al., 2004) :

- fournir des prédictions précises (e.g. moyenne du taux de pollinisation croisée à la parcelle) pour décider de l'implantation ou non d'une parcelle de maïs GM dans un paysage donné ;
- prédire la variabilité à l'intérieur d'une surface d'intérêt (e.g. une parcelle ou un ensemble de parcelles) pour aider à définir des plans d'échantillonnage optimaux ;
- classer des scénarios de manière à déterminer quels systèmes ou solutions techniques minimisent le risque de dépasser le seuil légal ;
- prédire des extrêmes, c'est-à-dire fournir des prédictions, même peu précises, d'événements difficilement observables (e.g. hybridation entre variétés cultivés et variétés sauvages apparentées).

Le type de modélisation approprié est alors déterminé par l'objectif d'utilisation. Pour répondre à ces questions relativement diverses, plusieurs types de modèles ont été développés.

Des modèles mécanistes ou physiques. Ces modèles décrivent la trajectoire des grains de pollen en trois dimensions et reposent sur la modélisation des processus de libération, de transport et de dépôt des grains de pollen. Ils permettent de prédire la proportion de pollen issue d'une ou plusieurs sources atteignant une parcelle cible (Aylor et al., 2003; Loos et al., 2003; Jarosz et al., 2004; Dupont et al., 2006; Lipsius et al., 2006; Arritt et al., 2007b) et ils prennent explicitement en compte l'influence des turbulences atmosphériques sur le transport du pollen. Ainsi, le modèle lagrangien SMOP développé par Jarosz et al. (2004) calcule les trajectoires des grains de pollen pour un environnement turbulent donné. Les modèles Aquilon et ARPS (Foudhil et al., 2005; Dupont et al., 2006; Dupont and Brunet, 2006) permettent quant à eux de calculer les champs de vent et de turbulence, tels qu'ils sont affectés par les caractéristiques structurelles du paysage, et ils utilisent ces champs pour simuler la dispersion.

Ce type d'approches est nécessaire pour acquérir une compréhension très détaillée des processus physiques impliqués dans la dispersion. Elles permettent, par exemple, la prise

en compte de mouvements convectifs très complexes qui amènent certains grains de pollen en altitude, avec des retombées à très longue distance. Cependant la nature des données récoltées pour calibrer ces modèles ne permet que très rarement d'évaluer la capacité de fécondation des grains de pollen à leur point d'arrivée. Il s'agit le plus souvent en effet de dispositifs de capture et comptage de grains de pollen à différentes distances d'une source avec peu de moyens pour mesurer la viabilité des grains de pollen capturés ou la réceptivité des soies environnantes. De ce fait, l'intérêt de l'utilisation de ces modèles pour quantifier l'efficacité de mesures de coexistence est relativement limité, cette quantification nécessitant de connaître des taux de pollinisation efficace (Beckie and Hall, 2008). D'autre part, ces modèles ayant une très haute résolution, ils requièrent une connaissance très fine d'une multitude de données de terrain, relevant notamment de la micrométéorologie, difficilement accessible en pratique (Lipsius et al., 2006). Pour toutes ces raisons, ces modèles, même s'ils permettent d'obtenir des informations précieuses sur les principaux processus à prendre en compte dans la dispersion ainsi que sur la hiérarchie de leurs effet respectifs, ne semblent pas adaptés au déploiement sur des paysages complexes pour la prédiction de taux de pollinisation croisée.

Des modèles empiriques. À l'inverse, compte du grand nombre de facteurs influençant les flux de gènes et de leurs interactions, certains auteurs se sont concentré, notamment pour leur simplicité, sur des approches qualifiées d'empirique. Ici, la modélisation consiste à réaliser un ajustement statistique de fonctions mathématiques sur des jeux de données plus ou moins représentatifs. Ces modèles donnent une probabilité de pollinisation croisée à une distance donnée d'une source émettrice en tenant compte éventuellement de la direction ou du positionnement vis-à-vis du vent dominant (Bateman, 1947c; Goggi et al., 2006; Robson et al., 2011). À partir des courbes ajustées, il est possible de simuler des taux de pollinisation croisée à l'échelle d'une parcelle (Damgaard and Kjellson, 2005; Weekes et al., 2005; Gustafson et al., 2006; Allnutt et al., 2008; Šuštar Vozlič et al., 2010) afin de tester de manière plus réaliste l'effet sur ces taux de facteurs tels que les tailles relatives des parcelles émettrices et réceptrices, la distance d'isolement ou la présence de rangs de bordure.

Les paramètres de ces modèles n'ont, la plupart du temps, pas de signification biologique ou physique, ce qui rend leur utilisation difficile dans des contextes autres que ceux des expérimentations ou suivis dont les résultats ont été utilisés pour les calibrer. Leur qualité prédictive dépend donc très fortement de la variabilité des contextes agro-climatiques et des situations incluses dans les données pour la calibration (Lavigne et al., 2004; Beckie and Hall, 2008). Dans tous les cas cités ici, les simulations ne concernent que des situations de dispersion champ à champ. Il faut noter que certaines de ces approches empiriques autorisent la prise en compte de plusieurs champs émetteurs notamment celles de Messeguer et al. (2006), de Ivanovska et al. (2009) et de Debeljak et al. (2012). Néanmoins tous ces modèles restent très dépendants des données utilisées pour leur calibration. De plus, aucune de ces approches empiriques ne prend en compte le phénomène de dilution de la concentration du pollen OGM par l'émission locale de pollen non OGM.

Des approches intermédiaires. Une approche, intermédiaire aux deux précédentes (mécaniste et empirique), a été développée et peut être qualifiée de *quasi* mécaniste. Elle repose sur la définition d'une fonction de dispersion individuelle qui donne la probabilité d'une fécondation d'un ovule par un grain de pollen émis à une distance donnée d'une

source. Cette fonction intègre la libération du pollen, sa dispersion et la fécondation de la fleur femelle sur l'ensemble de la période de floraison (Klein et al., 2003, 2006b,a). Ce type de modèle de dispersion, dit quasi mécaniste, prend en compte des paramètres climatiques associés notamment à la vitesse du vent, la vitesse de sédimentation et le niveau de turbulences mais également des paramètres biologiques liés par exemple à la différence de hauteur entre la source d'émission du pollen et les ovules récepteurs. Ces paramètres ont l'avantage d'avoir une signification physique ou biologique et ont été estimés via des essais au champ (Klein et al., 2003). Ces modèles permettent donc, à la différence des modèles purement empiriques, l'utilisation dans des contextes autres que ceux des expérimentations servant à leur calibration. Il reste cependant la difficulté de relier ces fonctions aux dynamiques de floraison des sources et du récepteur. En effet ces fonctions sont calibrées, la plupart du temps, dans des contextes de synchronisme de floraisons ce qui limite leur capacité à prendre en compte des situations réelles où le décalage de floraison est souvent observé entre différentes variétés de maïs. De plus, ces fonctions ayant été établies sur des dispositifs continus, leur pertinence pour rendre compte de l'effet des obstacles et de divers éléments d'hétérogénéité d'un paysage, notamment en bordure de champ se révèle relativement limitée (Angevin et al., 2002).

Des modèles intégratifs. Enfin, une approche que nous qualifions ici d'intégrative a été développé par Angevin et al. (2008) avec notamment pour objectif d'intégrer l'effet des facteurs agronomiques sur les taux de pollinisation croisée. Le modèle MAPOD (*MAtricial Approach to Pollen Dispersal*), issu de cette approche, intègre également les fonctions de dispersion de l'approche quasi mécaniste et permet donc de pallier en quelque sorte ses limites pour l'évaluation de mesures de coexistence. Ce modèle dynamique simule, au pas de temps journalier, la croissance des plantes et détermine ainsi les dates à partir desquelles les mâles émettent du pollen et les soies des femelles sont réceptives. Une fois ces dates déterminées, le modèle simule la quantité de pollen produite par chaque variété à partir de caractéristiques variétales puis simule leur dispersion via une fonction de dispersion individuelle. Ainsi, MAPOD intègre la plupart des processus clés pour l'établissement des taux de pollinisation croisée et permet une adaptation à des paysages complexes qui peuvent être très différents des situations dans lesquelles il a été calibré.

MAPOD constitue donc un très bon exemple de modèle de prédiction de taux de pollinisation croisée *i)* pour l'évaluation et, le cas échéant, l'établissement de mesure de coexistence appropriées ; *ii)* pour l'aide à la décision. Cependant, malgré les efforts consentis par Angevin et al. (2008) notamment pour rendre ce modèle utilisable dans une gamme suffisament large de situations, MAPOD requiert tout de même une très grande quantité de données pour son initialisation (e.g. température et pluviométrie sur toute la période de floraison, caractéristiques variétales, pratiques culturales, ...), et ne permet pas de fournir des prédictions s'il manque une partie de ces données ce qui complique son usage dans une grande partie des situations réelles. De plus, même si l'intégration de paysages complexes et étendus est possible avec ce modèle, les temps de calcul imposent de fortes limites à son utilisation dans ces cas là. Par ailleurs, le modèle MAPOD est un modèle déterministe et donne donc comme prédictions des valeurs ponctuelles dont il est difficile d'évaluer l'incertitude. Or, dans un contexte de prédiction pour l'aide à la décision et d'évaluation des risques, la quantification et la prise en compte de l'incertitude des prédictions sont nécessaires (Gouache et al., 2013) pour mieux cerner les impacts de chaque décision et les incertitudes qui demeurent sur la réalisation d'un processus résultant

de cette décision (Parent and Bernier, 2007; Boreux et al., 2009).

3 Gérer la coexistence en intégrant les incertitudes : le projet PRICE

Le projet européen PRICE (PRactical Implementation of Coexistence in Europe) s'est donné comme objectif ambitieux d'analyser les conditions permettant de rendre possible d'un point de vue pratique la coexistence entre cultures OGM et non OGM en Europe. Ce projet fait suite à plusieurs autres programmes interdisciplinaires européens (CoExtra, SIGMEA) ou nationaux (ANR GCOM2AP). Ces projets ont permis de réaliser une série d'objectifs allant aussi bien de la collecte et l'analyse des jeux de données européens disponibles sur les flux de gènes et les impacts environnementaux des grandes cultures génétiquement modifiées, à de la conception des modèles de flux de gènes prédictifs à l'échelle du paysage ainsi qu'à l'analyse de la faisabilité technique et la pertinence économique de la coexistence dans les principales régions agricoles de l'Europe (Messéan et al., 2009). Les principaux livrables de ces projets ont mis à disposition :

- une base de données regroupant les résultats des expérimentations sur les flux de gènes, consolidée à l'échelle européenne (Messéan et al., 2009) ;
- des modèles intégratifs de prédiction des taux de pollinisation croisée à l'échelle du paysage (Angevin et al., 2008).

Ces éléments ont déjà permis de réaliser des avancées significatives pour l'analyse et la compréhension des principaux facteurs rendant, ou non, la coexistence possible techniquement. Néanmoins, certaines conditions permettant la faisabilité de la coexistence et garantissant la liberté de choix aux agriculteurs comme aux consommateurs font encore défaut. Or, l'expertise acquise dans les précédents programmes de recherche permet de tirer des conclusions intéressantes afin de respecter à la fois l'efficacité des mesures et leur caractère proportionné. Celles-ci vont toutes dans le même sens et plaident pour l'élaboration d'un outil d'aide à la décision pour la coexistence de cultures OGM et non OGM et font des recommandations sur le cahier des charges d'un tel outil :

1. Si l'outil est basé sur un modèle de prédiction, celui-ci doit être suffisament flexible pour permettre l'adaptation au niveau d'information disponible dans la situation où il est utilisé, quel qu'il soit.

2. L'outil doit fournir une estimation de l'incertitude pour toutes prédictions issues des modèles afin d'informer l'utilisateur sur le niveau de confiance attaché à une valeur prédite et le risque lié à sa décision.

3. L'outil doit être ergonomique et convivial afin d'être utilisé par un public très large, allant des agriculteurs aux gestionnaires publics.

L'élaboration et la mise à disposition de cet outil d'aide à la décision constitue le cœur du projet PRICE.

3.1 La modélisation au cœur du processus de décision

Globalement, les différents programmes de recherche préalables au projet PRICE ont plaidé pour plus de flexibilité dans la mise en place de mesures de coexistence (Sausse et al., 2013; Devos et al., 2014). Cette flexibilité peut être introduite dans le processus de décision par plusieurs moyens. Cependant le recours à la modélisation apparaît spécialement adapté aux objectifs du projet. En effet, comme on l'a vu dans la section précédente, il existe des modèles (e.g. MAPOD) qui permettent de simuler l'effet de différentes pratiques culturales ou choix techniques sur le taux de pollinisation croisée et qui peuvent s'adapter à des paysages complexes et divers. Ce type de modèles permet d'une part d'éviter d'avoir recours à des expérimentations fort coûteuses en temps et en argent et donc de simuler et évaluer l'efficacité de plusieurs mesures de coexistence presque gratuitement et dans un temps qui reste très inférieur à une saison de culture. D'autre part, il offre la possibilité de tester ces mesures sur une gamme de paysages larges et variés et permet donc plus de flexibilité dans l'établissement de ces mesures. Il devrait même permettre d'évaluer la faisabilité de coexistence dans chaque situation individuelle. Cependant, comme cela a été évoqué dans la section précédente, les deux grandes limites de MAPOD sont *i)* la grande quantité de données nécessaire à son utilisation et l'impossibilité de fournir des prédictions s'il en manque ; *ii)* la difficulté d'obtenir une estimation de l'incertitude associée à ses prédictions.

L'un des objectifs du projet PRICE est de profiter de l'expertise acquise dans les différents programmes de recherche et par l'établissement de ces modèles intégratifs pour élaborer et proposer un ou des modèles offrant des possibilités comparables à MAPOD tout en répondant mieux aux besoins et aux cahiers des charges d'un outil d'aide à la décision tel que défini plus haut. On rappelle que les deux principales caractéristiques à considérer pour un tel modèle sont d'une part la flexibilité, c'est-à-dire la capacité à fournir des prédcitions même si le niveau d'information sur les caractéristiques environnementales et agronomiques est faible et, d'autre part, la possibilité d'évaluer l'incertitude sur les prédictions qui en sont issues.

D'autre part, un modèle vérifiant ces propriétés est prévu pour être le cœur du processus de décision. Il devra donc être suffisament flexible pour permettre de fournir une aide à la décision dans toute situation potentielle, mais également ergonomique et facilement accessible de manière à permettre son utilisation par différents types d'acteurs et pour des cas d'utilisation potentiellement variés.

3.2 Des prédictions ? Oui, mais avec une incertitude associée !

Comme on l'a vu dans la définition du cahier des charges de l'outil d'aide à la décision, celui-ci doit fournir, pour toute prédiction, une estimation de l'incertitude qui lui est associée. Cette incertitude peut prendre plusieurs formes (intervalles de confiance ou distribution de probabilité) mais doit toujours être accessible à l'utilisateur du modèle. De plus, cette incertitude se doit d'être facilement calculable, pas uniquement pour les prédictions elles-mêmes, mais aussi pour toute fonction des prédictions de manière à obtenir également l'incertitude associée aux critères calculés à partir des prédictions de l'outil. De ce fait, l'un

des objectifs de PRICE est d'élaborer un modèle pour lequel l'incertitude accompagne *automatiquement* chaque prédiction, et est facilement calculable pour tout critère découlant de ces prédictions. Sans rentrer ici dans des détails techniques, le projet considère les approches bayésiennes comme particulièrement appropriées pour réaliser ces objectifs, notamment pour la facilité qu'elles offrent pour la restitution des incertitudes issues de toute fonction des sorties du modèle. On verra plus loin dans ce mémoire que ces approches statistiques présentent d'autres avantages pour l'aide à la décision, en particulier la possibilité d'intégrer de l'information *a priori*, lorsqu'elle existe, dans les estimations (Clark, 2005) ainsi que la souplesse permise dans la définition de modèles hiérarchiques (Parent and Bernier, 2007; Boreux et al., 2009).

4 Objectifs de la thèse

Les objectifs propres de cette thèse peuvent se décliner en deux grandes catégories : des objectifs scientifiques et des objectifs appliqués ou opérationnels.

Les objectifs opérationnels s'insèrent dans ceux du projet européen PRICE. Ce travail doit donc fournir des solutions techniques, basées sur la modélisation, permettant d'assurer la coexistence entre filères de maïs OGM et non OGM. Sur le plan pratique, l'objectif est d'élaborer et de mettre à disposition des acteurs des filières de production de maïs un ou plusieurs modèles mathématiques de prédiction du taux de pollinisation croisée à l'échelle d'un paysage agricole. Ce modèle ou ensemble de modèles sera destiné à différents types d'acteurs (agriculteurs, conseillers agricoles, coopératives, législateurs) et devra permettre de répondre à différentes questions concrètes.

Ces questions peuvent concerner une échelle spatiale et temporelle très réduite, par exemple s'il s'agit d'implanter ou non une parcelle de maïs OGM en un lieu et pour une saison de culture donnés ; ou bien une échelle beaucoup plus large, par exemple s'il s'agit de l'établissement à long terme de mesures de coexistence adaptées à l'ensemble d'une région agricole. Selon la question posée, les données nécessaires à la prise de décision et leur écahelle spatiale différera. De fait, le système à mettre en place doit être suffisamment flexible pour permettre de répondre à des questions diverses. Il doit de plus être adaptable à une diversité de situations pour lesquelles les niveaux d'information accessibles peuvent être très différents. Au préalable, il est nécessaire de bien définir les situations dans lesquelles ces modèles seront utilisés, les niveaux d'information associés et le type de questions auxquelles ils devront répondre.

Un point supplémentaire essentiel que doit permettre ce travail est la prise en compte et la restitution des incertitudes dans les prédictions des modèles. Pour cela, les modèles définis et calibrés dans cette thèse devront être en mesure de fournir des prédictions probabilistes, c'est-à-dire de donner une distribution de probabilité pour chaque prédiction plutôt qu'une valeur ponctuelle. Les objectifs opérationnels de cette thèse peuvent donc être résumés par ces quatre points :

1. déterminer les cas d'utilisation potentiels ainsi que les niveaux d'information qui y sont associés ;

2. élaborer et fournir des modèles prédictifs du taux de pollinisation croisée dans les champs de maïs conventionnels ;
3. donner aux modèles la capacité de fournir des prédictions probabilistes ;
4. permettre l'adaptation des modèles au niveau d'information disponible pour toute situation ou cas d'utilisation potentiel.

Tous ces éléments amènent à se poser plusieurs questions d'ordre beaucoup plus fondamental qui constituent les objectifs scientifiques de ce travail de thèse. En premier lieu, il apparaît essentiel de rechercher tout d'abord, puis de sélectionner et de comparer, les formalismes mathématiques existants et pertinents pour prendre en compte les facteurs clés de la dispersion du pollen de maïs dans la modélisation des taux de pollinisation croisée dans les champs conventionnels.

Par ailleurs, un des objectifs opérationnels de cette thèse étant de fournir des prédictions probabilistes pour refléter les incertitudes, les approches bayésiennes offrent un cadre particulièrement attractif ici. De ce fait, leur application à notre problème constitue un des objectifs scientifiques de cette thèse. Cette application a pour buts *i)* de définir et calibrer les modèles de manière, d'une part à retranscrire de manière suffisamment fidèle la grande variabilité des taux de pollinisation croisée observée dans les jeux de données disponibles, d'autre part à améliorer la compréhension des éléments qui structurent cette variabilité ; *ii)* de calculer aisément (via la distribution prédictive ; *a posteriori*) l'incertitude existante sur toute prédiction ou fonction des prédictions ; *iii)* de permettre une communication plus facile de cette incertitude.

La prise en compte des caractéristiques spatiales des surfaces émettrices de pollen dans la modélisation de la dispersion entraîne généralement des problèmes en termes de temps de calcul, du moins pour un des formalismes mathématiques étudiés par la suite. Ces temps de calculs peuvent rendre certains modèles inutilisables en pratique dans un contexte *i)* d'estimation par des approches bayésiennes, elles-mêmes fortes consommatrices de temps de calcul même pour des modèles relativement simples ; *ii)* d'aide à la décision nécessitant de fournir des réponses rapides et donc des modèles qui *réagissent* vite. C'est pourquoi un des objectifs de ce travail est de fournir des solutions, basées sur des approximations, pour réduire les temps de calcul inhérents à ce formalisme et rendre possible son utilisation dans un cadre bayésien et à des fins d'aide à la décision.

Enfin, le fait de devoir proposer des modèles adaptables à toute une gamme de situations potentielles d'une part, et la diversité des choix possibles dans la modélisation d'autre part, entrainent la définition d'un très grand nombre de modèles *candidats*. De ce fait, un objectif supplémentaire de ce travail de thèse concerne la sélection de modèle et plus particulièrement *i)* la définition de critères adaptés aux objectifs particuliers ; *ii)* la définition de procédures *ad-hoc* pour sélectionner les modèles lorsque les méthodes usuelles ne suffisent pas et que différents critères de sélection donnent des informations contradictoires.

Les objectifs scientifiques de cette thèse peuvent donc être résumés par ces quatre points :

1. Déterminer les formalismes ou paradigmes de modélisation appropriés pour la modélisation des taux de pollinisation croisée et comparer les options retenues.

2. Appliquer les approches bayésiennes pour permettre
   - une modélisation satisfaisante de l'espérance et de la variance du taux de pollinisation croisée,
   - une meilleure compréhension des éléments qui structurent la variabilité des taux observés,
   - une restitution aisée de l'incertitude afférente à chaque prédiction des modèles.

3. Résoudre les problèmes calculatoires qui se posent lors de l'utilisation d'un type particulier de formalisme, et notamment rendre les temps de calcul compatibles avec l'utilisation opérationnelle envisagée.

4. Proposer des solutions *ad-hoc* pour la sélection de modèles.

Chapitre 2

Conceptualisation

Table des matières

1 Situations de prédiction 27

1.1 Problématique 27

1.2 Exemples et définitions 28

1.3 Conséquences opérationnelles 31

2 Recherche d'un cadre statistique approprié 34

2.1 L'école classique ou Fréquentiste 34

2.2 L'école Bayésienne 34

3 Approches Bayésiennes pour l'estimation des paramètres . . 35

3.1 Vraisemblance 36

3.2 Distribution *a priori* 37

3.3 Distribution *a posteriori*, Metrolopolis-Hastings et MCMC . . . 37

4 Avantages de l'approche bayésienne pour l'aide à la décision 39

4.1 Quantification des incertitudes pour toutes fonctions des paramètres 39

4.2 Mise à jour perpétuelle de l'information 40

4.3 Modélisation hiérarchique 40

5 Plusieurs types d'expérience/de données 41

5.1 Types d'expériences: Mono-source vs. Multi-sources 41

5.2 Types de données: discrètes vs. continues 42

5.3 Types de configurations: trop de diversité? 47

Ce chapitre a pour but de présenter et décrire les travaux et éléments préalables aux étapes de modélisation à proprement parler, indispensables pour répondre aux objectifs de la thèse. Nous décrirons dans un premier temps les situations ou scénarios pour lesquels le ou les modèles proposés doivent fournir des prédictions et les questions auxquelles ils doivent pouvoir répondre. Nous discuterons de la quantité d'information disponible dans chaque situation et des implications en termes *i)* d'hypothèses à effectuer ou de solutions à fournir pour réaliser les prédictions ; *ii)* d'incertitudes dans les décisions résultant de l'utilisation d'un modèle. Dans un second temps, nous nous pencherons sur les différents cadres ou paradigmes statistiques afin d'en dégager les grandes caractéristiques et de justifier notre choix de nous concentrer sur le paradigme bayésien. Nous décrirons les éléments spécifiques ainsi que les avantages de ce paradigme d'une part pour l'estimation des paramètres d'un modèle et d'autre part pour l'aide à la prise de décision en présence d'incertitudes. Enfin, nous présenterons les différents types d'expérimentations ou de dispositifs de suivi qui permettent d'appréhender le processus de flux de gènes ainsi que la nature des données qui en résultent.

1 Situations de prédiction

1.1 Problématique

Avant de proposer un quelconque système ou modèle permettant de réaliser des prédictions, il nous apparaît légitime de définir en premier lieu les situations dans lesquelles on veut être en mesure de fournir ces prédictions (Castellazzi et al., 2010; Sausse et al., 2013). Cette étape de "cadrage" du problème a constitué le premier travail propre à cette thèse. On définit ici une *situation de prédiction* comme un triplet Acteurs/Questions/Niveau d'information.

Les acteurs sont les opérateurs de la filière de production du maïs, ils peuvent être des agriculteurs cultivant des variétés de maïs, OGM ou non, ou des organismes de collecte tels que des coopératives agricoles assurant la collecte, l'allotement et le stockage des produits de la filière. Ces acteurs se posent des questions auxquelles il faut pouvoir fournir des réponses adaptées à chaque situation. Ces questions sont diverses mais reposent toujours sur le respect ou non d'un seuil de présence (seuil légal de 0.9% ou tout autre niveau requis par les filières) d'organismes génétiquement modifiés dans une récolte de variétés de maïs non génétiquement modifiées, ou dans un ensemble de récoltes de ce type dans un bassin de collecte.

Le niveau d'information associé à chaque situation dépend en premier lieu du stade d'avancement de la culture au moment où la question a été posée et doit avoir une réponse. Si on souhaite prendre une décision avant les semis, aucune information certaine n'est encore disponible sur les dates de semis, ni sur des mesures de vitesse et direction du vent pendant la période de floraison (celle-ci n'ayant pas eu lieu). À ce stade, on peut néanmoins faire des hypothèses sur les positions potentielles de parcelles de maïs OGM et non-OGM ou les autres facteurs affectant la pollinisation croisée. Après les semis, les positions de parcelles OGM et non-OGM sont connues, certaines et définitives tout comme

les dates de semis. À ce niveau l'incertitude demeure sur les dates de floraison et sur les caractéristiques climatiques telles que la direction et la force du vent.

D'autres éléments peuvent faire varier le niveau d'information ou plus précisément la confiance que l'on peut attribuer aux informations pour un niveau donné. En effet, certaines informations peuvent être connues de manière imparfaite. C'est le cas de la date de floraison, le plus souvent estimée par des observations visuelles qui ne sont pas sans erreur. Un autre exemple est donné par les mesures climatiques (force et direction du vent principalement) qui, même si les erreurs de mesure sont très faibles grâce à la précision des intruments de mesure, ne sont que très rarement réalisées à l'endroit exact où le processus à lieu, ce qui génère une incertitude sur les valeurs réellement prises par ces grandeurs au moment et lieu exacts du processus modélisé.

1.2 Exemples et définitions

Nous donnons dans un premier temps quelques exemples de situations de prédiction pour chaque grand type d'acteurs tels que décrit ci-dessus. Puis nous généralisons le problème en définissant trois grandes classes de situations de prédiction pertinentes vis-à-vis des objectifs opérationnels spécifiés au Chapitre 1, section 4.

Commençons par les trois exemples :

- Pour l'agriculteur ne souhaitant pas cultiver d'OGM, la question principale peut se formuler ainsi :
  "Sachant que des OGM sont semés dans la région, où et comment (quelle parcelle ou îlot ? Quelle variété ? Quelle date de semis ?) dois-je implanter mon maïs pour minimiser le risque de dépasser le seuil légal du taux de pollinisation croisée ?"
  Il est clair que si cette question se pose et est pertinente, elle ne peut avoir un sens qu'avant les semis de maïs. À ce stade, le niveau d'information est relativement faible, en effet seules les positions potentielles des champs OGM sont connues (car la déclaration des intentions de semis de parcelles OGM est obligatoire et doit intervenir bien avant les semis de maïs), on ne connaît pas encore ni les positions des champs non-OGM, ni les variétés semées, ni les dates de semis. Cette situation est donc qualifiée d'*Ex-ante*, elle correspond à une situation préalable à tout choix (assolement, variété, date de semis) définitif.
- Pour l'agriculteur souhaitant cultiver des OGM, la question peut se formuler ainsi : *"Sachant qu'il y a des agriculteurs ne cultivant pas d'OGM dans la région et que je pourrais être tenu responsable du taux de pollinisation croisée retrouvé chez mes voisins, où et comment dois-je implanter mon maïs pour minimiser le risque de faire dépasser le seuil légal du taux de pollinisation croisée à mes voisins ?"*.
  De la même façon que la situation précédente, cette question n'a de sens qu'avant les semis. Elle correspond également à une situation dite *Ex-ante* car préalable à tout choix définitif. Le niveau d'information est toutefois plus limité car, si les producteurs d'OGM sont tenus de déclarer leurs intentions de semis, cette obligation ne s'applique pas aux parcelles non-OGM.
- Pour un organisme de collecte de maïs une des questions possibles (dépendant de l'avancement de la culture) pourrait être :

"Sachant que, parmi mes clients, il y a des agriculteurs qui cultivent des OGM et d'autres non, peut-on faire des prévisions de lots homogènes en termes de taux de pollinisation croisée ?".
Cette question peut se poser à différents stades d'avancement de la culture :
- juste après les semis : acquisition d'information spatiale (positions des parcelles OGM et non-OGM) et culturale (dates de semis, variétés $\propto$ précocités, durées de floraison) ;
- au début de la floraison : acquisition d'informations sur les dates de floraison ;
- après la floraison : acquisition d'informations climatiques (direction et force du vent) et d'informations sur les durées de floraison des différentes variétés semées.

L'énumération de ces exemples nous permet de dégager trois grandes classes de situations dépendant principalement de l'avancement de la culture de maïs dans le temps et du niveau d'information qui s'y rapporte. Nous définissons à présent ces situations et indiquons pour chacune comment le modèle peut être utilisé pour répondre aux questions posées. Le tableau 2.1 regroupe ces situations :

1. EX ANTE *"Faites vos jeux"* $\Rightarrow$ Cette situation est qualifiée ainsi car le processus que l'on cherche à prédire n'a pas encore eu lieu. L'appellation *"Faites vos jeux"* vient du fait que toutes les stratégies sont encore possibles. Les agriculteurs souhaitant cultiver des OGM peuvent encore y renoncer ou changer leur assolement de manière à réduire *a priori* l'impact sur leurs voisins. Les agriculteurs ne cultivant pas d'OGM peuvent s'entendre avec leurs homologues sur les variétés semées de manière à avoir un décalage de floraison entre les deux types de culture. On rappelle qu'en cas de litiges ou d'impossibilité de s'entendre, c'est l'agriculteur qui introduit l'innovation (ici l'OGM) qui est supposé (Directives européennes toujours en discussion) s'adapter à ses voisins. Pour ce type de situation, le ou les modèles proposés peuvent être utilisés de différentes façons pour répondre aux questions posées :
   - À partir d'une configuration initiale potentielle d'assolement et de répartition OGM/non OGM, le modèle doit permettre d'identifier les parcelles ou zones pour lesquelles le risque de dépasser le seuil légal est suffisamment élevé pour justifier un changement d'assolement et de répartition OGM/non OGM.
   - En l'absence d'une configuration initiale potentielle d'assolement et de répartition OGM/non OGM, le modèle doit permettre de déterminer les éléments (distance minimale entre parcelles, orientation vis-à-vis du vent dominant, nombre de jours de décalage entre les floraisons OGM et non OGM) qui doivent être respectés pour garantir aux agriculteurs ne cultivant pas d'OGM, avec une probabilité donnée, le non dépassement du seuil légal.

2. EX POST 1 *"Les jeux sont faits"* $\Rightarrow$ Pour cette situation, des choix ont été faits et ils sont irréversibles : choix d'un assolement, choix d'une variété et donc, implicitement, d'une précocité, d'une date de semis et d'une durée de floraison (ces deux dernières caractéristiques étant très dépendantes de la variété). L'appellation *"Les jeux sont faits"* provient du fait que les choix faits au préalable sont définitifs. Autrement dit, on ne pourra plus compter sur des changements de variétés ou de positions des parcelles. À partir de ce stade, l'agriculteur n'a plus la possibilité d'influencer le processus de flux de pollen par des choix techniques, cette situation

concerne donc plutôt les organismes de collecte, de stockage et de vente c'est-à-dire les coopératives agricoles. Ici encore, pour répondre à une même question, plusieurs modes d'utilisation du modèle peuvent être pertinents :

- La première utilisation repose sur la prédiction des taux moyens de pollinisation croisée pour chaque parcelle du paysage considéré sans action autre que la récolte de la part des agriculteurs. En fonction de la valeur prédite et de l'incertitude afférente à cette prédiction (si la valeur est inférieure au seuil et l'incertitude suffisament faible), cette utilisation peut être suffisante à ce stade. Cependant, dans le cas contraire, si la valeur prédite est au dessus du seuil légal ou si l'incertitude de cette prédiction est trop grande, un deuxième mode d'utilisation peut être défini.
- On rappelle qu'à ce stade, on n'a plus la possibilité d'influencer le processus de flux de pollen en lui-même, néanmoins le modèle doit pouvoir servir à simuler différentes techniques ou pratiques culturales permettant a priori de réduire le taux moyen de pollinisation croisée dans une parcelle ; on pense notamment ici au détourage (bordures de parcelle récoltés séparément du reste) et au mélange des récoltes issues de différentes parcelles.

3. EX POST 2 *"Rien ne va plus"* ⇒ Ici on est entre la fin de floraison et la récolte et les agriculteurs ne cultivant pas d'OGM n'ont aucune marge de manœuvre pour diminuer le taux de pollinisation croisée à l'intérieur de leurs parcelles (les jeux ont été faits et donc rien ne va plus !). Cette situation concerne, ici encore, plutôt les coopératives agricoles qui s'intéressent au niveau de pureté des récoltes de leurs clients afin de constituer des lots homogènes en termes de taux de présence fortuite. Encore une fois, pour cette situation, on envisage plusieurs façons d'utiliser le modèle pour aider les acteurs à répondre aux questions qu'ils se posent :
   - La première utilisation repose, comme dans la situation précédente, sur la prédiction des taux moyens de pollinisation croisée pour chaque parcelle du paysage considéré sans action autre que la récolte de la part des agriculteurs. Si à ce stade, toutes les parcelles non GM ont une prédiction qui respecte le seuil et que l'incertitude sur cette prédiction est suffisament faible, ce type d'utilisation peut suffire pour la coopérative agricole à déterminer des lots homogènes en termes de taux de pollinisation croisée.
   - Dans le cas où l'incertitude sur la prédiction demeure très élevée, on suppose que l'acteur aura tendance à préférer se baser sur une estimation *in situ* du taux de pollinisation croisée plutôt que sur les prédictions du modèle, ce qui suppose de faire des prélèvements dans les parcelles et d'analyser les échantillons prélevés). Dans ce cas le modèle doit pouvoir fournir une aide pour la détermination d'un plan d'échantillonnage fournissant une estimation la plus proche possible du taux moyen réel avec un nombre d'échantillons limité. Pour atteindre cet objectif, le modèle ne doit pas seulement être capable de fournir une estimation précise du taux moyen de présence fortuite dans la parcelle mais aussi et surtout de retranscrire suffisamment fidèlement la variabilité de ce taux à l'intérieur de la parcelle réceptrice.

1.3 Conséquences opérationnelles

D'un point de vue opérationnel, la définition de ces trois grands types de situations nous oblige à fournir des solutions de modélisation permettant de réaliser les prédictions dans des situations très différenciées du point de vue de l'information disponible. Je me suis donc concentré sur la définition de deux ensembles de solutions : un premier ensemble basé sur l'utilisation de modèles différents pour des niveaux d'information différents, et un second ensemble basé sur un même modèle dont les hypothèses sur les entrées varient en fonction du niveau d'information disponible.

Le premier ensemble de solutions consiste à proposer un modèle pour chaque niveau d'information. Nous avons fait le choix ici de nous concentrer sur trois variables dont la prise en compte est indispensable pour bien décrire les processus de flux de gènes :

- la distance entre émetteurs et récepteurs ;
- la force et la direction du vent dominant pendant la période de floraison ;
- le décalage temporel pouvant survenir entre la floraison des émetteurs et celle des récepteurs.

À ce stade et en fonction de l'information disponible, nous proposons de mettre à disposition quatre modèles identifiés par leurs variables d'entrée et de laisser à l'utilisateur le choix du modèle qui convient le mieux à sa situation :

1. modèle "Distance" ;
2. modèle "Distance + Vent" ;
3. modèle "Distance + Décalage de floraison" ;
4. modèle "Distance + Vent + Décalage de floraison".

On peut constater que toutes les combinaisons entre les variables d'entrée ne sont pas représentées, en effet il serait incongru de vouloir faire des prédictions en ayant connaissance uniquement du vent dominant ou du décalage de floraison potentiel. Cela nous amène à considérer la variable "Distance" comme une variable *par défaut* indispensable à tous les modèles.

Le deuxième ensemble de solutions proposé consiste à considérer le modèle le plus complet (soit ici "Distance + Vent + Décalage de floraison") et à faire des hypothèses sur les variables qui n'ont pas été mesurées. On rappelle que les distances sont considérées comme toujours connues, les hypothèses à faire reposent donc sur les deux autres variables d'entrée, à savoir les direction et force du vent et les décalages de floraison.

On rappelle qu'on se place dans un contexte d'évaluation du risque (i.e. de dépasser le seuil légal de présence fortuite d'OGM dans une récolte étiquetée "non-OGM"). Cela suggère de considérer, en absence d'information, la situation la plus à risque.
Pour le décalage de floraison, lorsque les dates de floraison ne sont pas disponibles, le problème est relativement simple à résoudre. La situation la plus à risque étant celle où

aucun décalage n'intervient entre les variétés OGM et non OGM, on considère donc un synchronisme de floraison total tant que les informations sur les dates de floraison ne sont pas disponibles.
Pour le vent, la question est plus délicate ; elle nous oblige à distinguer deux situations :

1. existence de données historiques (mesure de vent) dans la zone d'intérêt ;
2. absence de données historiques (mesure de vent) dans la zone d'intérêt.

Pour la situation 1, il y a au moins trois stratégies possibles pour prendre l'information existante en compte :

- déterminer la force et la direction du vent dominant à partir de séries temporelles et considérer uniquement ces quantités ;
- définir un ensemble de forces et directions représentatif des vents observés dans des séries temporelles disponibles et pondérer les prédictions par l'occurrence de chaque vent ;
- considérer la totalité de la distribution de forces et de directions des vents, faire des tirages aléatoires dans cette distribution et répéter les prédictions pour chaque tirage.

Les deux dernières stratégies offrent plus de flexibilité dans les prédictions et donc potentiellement plus de réalisme que la première. Cependant, pour la partie qui concerne mon travail propre, c'est-à-dire la définition, la calibration et l'évaluation de modèle de dispersion, je me suis concentré sur la première possibilité principalement pour le gain de temps de calcul qu'elle permet par rapport aux autres. En effet, sans rentrer ici dans les détails méthodologiques que nous verrons par la suite, les méthodes employées pour la calibration sont très chronophages, il ne paraît donc pas très judicieux, au moins à cette étape, d'augmenter les temps de calcul. D'autre part, même si les deux dernières stratégies peuvent être envisagées, on a, dans la pratique, plus souvent accès à des données agrégées et donc plus proches du type de celles utilisées par la première solution.

La situation 2 est plus délicate. Pour la traiter, deux stratégies ont été envisagées :

- considérer que la dispersion est isotrope ;
- considérer que le récepteur est sous le vent (pour chaque parcelle réceptrice, le vent vient de la parcelle OGM la plus proche).

De la même façon que pour le décalage de floraison, il semble assez évident que la deuxième stratégie est la plus à risque. En absence de mesure de vent dans la zone d'intérêt, on considérera donc que chaque récepteur est sous le vent de la parcelle OGM la plus proche de manière à considérer le cas le plus défavorable (*worst-case scenario*).

Pour conclure sur ces conséquences opérationnelles, il faut préciser que les solutions présentées ici ont toutes été mises à disposition des acteurs concernés par les questions de coexistence entre OGM et non OGM. Les quatre types de modèles proposés pour le premier ensemble de solutions ainsi que le modèle complet du deuxième ensemble ont été calibrés et implémentés dans un outil d'aide à la décision accessible via une application

web sur le site :
http ://www.price.preprod.farmsat.com.

Cette application permet *i)* de dessiner des parcelles ou de rentrer ses coordonnées géographiques via un SIG pour définir une configuration spatiale de départ ; *ii)* de renseigner les données auxquelles on a accès (direction du vent, date de floraison) ; *iii)* de faire tourner le ou les modèles ; *iv)* de visualiser les résultats de simulation. Les choix sur le modèle le plus pertinent vis-à-vis du niveau d'information et les hypothèses à effectuer en fonction du modèle utilisé sont laissés libres à l'utilisateur. La définition de situations de prédiction ainsi que l'implémentation et la mise à disposition de l'outil d'aide à la décision ont fait l'objet d'une communication à la 6ème conférence internationale sur la coexistence entre culture OGM et non OGM (GMCC 2013) (Meillet et al., 2013) dont les actes figurent en annexes.

Situation	Acteurs	Période	Informations	Questions
EX ANTE *"faites vos jeux"*	Agriculteur OGM Agriculteur non-OGM	Déclarations des intentions de semis OGM	Positions OGM potentielles	Assolement OGM ? Assolement nonOGM ? Choix variétal ?
EX POST *"les jeux sont faits"*	Coopératives	Entre fin des semis et floraison	Positions OGM Positions nonOGM Variétés Dates de semis	Allotement ? Observ. floraisons ?
EX POST *"Rien ne va plus"*	Coopératives	Entre la fin de la floraison et la récolte	Positions OGM Positions nonOGM Variétés Dates de semis Climat (vent) Dates de floraison	Allotement ? Echantillonnages post-floraisons (PCR) ?

TABLE 2.1: Tableau récapitulatif des situations de prédictions

2 Recherche d'un cadre statistique approprié

"L'objet principal de la statistique est de mener, grâce à l'observation d'un phénomène aléatoire, une inférence sur la distribution probabiliste à l'origine de ce phénomène c'est-à-dire de fournir une analyse (ou description) d'un phénomène passé, ou une prédiction d'un phénomène à venir de nature similaire." (Robert, 2006). Comme l'explique cette citation, l'outil statistique est particulièrement approprié à notre travail dont un des objectifs est précisément de fournir des prédictions probabilistes en se basant sur des observations d'un phénomène considéré, au moins en partie, comme aléatoire. Pour réaliser cette inférence, il existe deux paradigmes ou écoles qui, dit simplement, reposent sur des hypothèses différentes quant à l'origine de la variabilité observée. De manière à expliciter voire à justifier notre choix Bayésien, nous allons nous attacher à montrer les grandes différences entre ces deux paradigmes.

2.1 L'école classique ou Fréquentiste

Quel que soit le paramètre inconnu θ à estimer, le mode de raisonnement du statisticien fréquentiste est toujours le même. Implicitement, θ a une valeur unique et son estimation requiert une statistique (ou estimateur), c'est-à-dire une fonction des données qui dépend de θ. Sous certaines conditions, les données disponibles permettent de calculer un intervalle de confiance correspondant à un risque de 1ère espèce α fixé. Le paramètre inconnu θ est ou n'est pas dans cet intervalle. La véritable interprétation de l'intervalle de confiance à 95% n'est donc pas *"Il y a 95% de chances pour que le paramètre soit compris dans cet intervalle."* (ce qu'on entend souvent et qui est absurde puisque, dans l'interprétation fréquentiste θ a une *vraie* valeur, certes inconnue mais conceptuellement unique dit autrement, cette valeur est inconnue mais certaine), mais *"Si je répète n fois la même expérience et qu'à chaque fois je calcule l'intervalle de confiance correspondant, ce dernier contiendra θ dans 95% des cas"*. C'est la vision fréquentiste : tout est dans les données. Cette vision implique notamment qu'aucune source d'information autre que les données ne peut intervenir dans le processus d'estimation. Dans les méthodes fréquentistes, les paramètres sont considérés comme des valeurs fixes auxquelles nous n'avons pas accès et qu'on ne peut qu'estimer, sous certaines conditions, via des estimateurs. Il est possible d'obtenir la distribution de ces estimateurs, cependant l'acquisition de cette distribution repose sur un certain nombre d'hypothèses ou sur des considérations asymptotiques difficiles à satisfaire en pratique. Il en résulte le plus souvent une moins bonne prise en compte de la variabilité.

2.2 L'école Bayésienne

Le statisticien Bayésien raisonne différemment puisqu'il considère que le paramètre du modèle statistique θ est incertain. Il s'attache donc à quantifier son incertitude en mobilisant toutes les informations disponibles. C'est là toute la différence avec l'école classique puisque cela revient à conférer au paramètre θ le statut de variable aléatoire. Dès lors, il devient sensé de lui attribuer une distribution *a priori* décrivant le savoir actuel sur ce paramètre. Cette distribution de probabilité, souvent appelé *prior*, est noté $\pi(\theta)$. Le prior

quantifie l'état de connaissance (et donc l'incertitude) d'un expert sur le problème étudié. Cela signifie que l'expert mise plus volontiers sur certaines valeurs de θ que sur d'autres (ne pas confondre incertitude et ignorance). Cette information a d'autant plus de valeur que les données sont rares et que leur acquisition peut s'avérer très coûteuse. L'originalité des méthodes bayésiennes par rapport aux méthodes classiques ou fréquentistes, repose donc sur deux grands principes. Premièrement, les méthodes bayésiennes permettent d'intégrer de l'information a priori sur les paramètres d'un modèle (e.g. via l'étude de processus individuels) et de combiner cette information avec des données expérimentales issues d'observations partielles sur le système complet. Avec les méthodes fréquentistes, il est impossible d'intégrer une autre source d'information, et donc de variabilité, que celle des données. Le second principe réside dans la recherche d'une distribution de probabilité de paramètres, qui sont ici considérés comme variables aléatoires, plutôt que d'une valeur. De ces deux grands principes découle une meilleure prise en compte de la variabilité des paramètres et des observations mais également une description plus réaliste et plus complète de l'incertitude. La connaissance de cette variabilité doit permettre de quantifier plus aisément l'incertitude présente dans les modèles proposés.

Pour toutes les raisons évoquées dans la première partie de cette section nous avons fait le choix de nous placer dans un cadre Bayésien pour toute la suite de ce travail. Nous présentons à présent les éléments spécifiques du paradigme Bayésien et leur pertinence vis-à-vis de nos objectifs, d'une part pour l'estimation des paramètres d'un modèle et d'autre pour l'aide à la prise de décision en environnement incertain.

3 Approches Bayésiennes pour l'estimation des paramètres

Les méthodes bayésiennes tirent leur nom du théorème de Bayes et reposent principalement sur une équation. L'équation de Bayes est une équation de base en probabilités, que l'approche soit fréquentiste ou bayésienne. Elle s'écrit :

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{2.1}$$

où A et B sont deux événements aléatoires. Dans l'approche bayésienne, on applique cette formule avec A associé à θ, le vecteur des paramètres, et B associé à Y, le vecteur des mesures. La forme de l'équation qui nous intéresse est alors

$$P(\theta|Y) = \frac{P(Y|\theta) \times \pi(\theta)}{P(Y)}. \tag{2.2}$$

Dans cette équation, $P(\theta|Y)$ représente la distribution a posteriori des paramètres et $P(Y|\theta)$ est la vraisemblance de Y sachant θ fournie par un modèle statistique. Ces notations sont des raccourcis que nous n'expliciterons pas plus ici. Ce qu'il faut retenir, c'est qu'on traite le vecteur des paramètres θ comme une variable aléatoire, sinon on ne pourrait pas parler de sa distribution. Dans l'approche fréquentiste par contre le vecteur

des paramètres θ a une valeur fixe. C'est pourquoi les méthodes fréquentistes n'appliquent pas cette formule au problème d'estimation des paramètres.

Le fait de raisonner conditionnellement au vecteur Y est différent également de l'approche fréquentiste. Dans celle-ci, la base de la variabilité est la variabilité de l'échantillon. Si on refaisait l'expérience, on mesurerait des valeurs différentes. Ainsi, quand on parle par exemple d'un intervalle de confiance à 90%, pour un fréquentiste cela veut dire que si on refaisait l'expérience suffisamment de fois, et que l'on calculait l'intervalle de confiance chaque fois avec les nouvelles valeurs mesurées, dans 90% des cas la vraie valeur de θ se trouverait dans l'intervalle. C'est l'intervalle qui est aléatoire, parce qu'il est calculé à partir de mesures et ces mesures sont issues d'un échantillon choisi au hasard.

Dans une approche bayesienne, la variabilité que l'on aurait si on refaisait l'expérience est sans intérêt. Tous les calculs se font sur la base des valeurs mesurées dans la seule expérience réellement réalisée. C'est-à-dire que tous les calculs se font conditionnellement à la valeur observée de Y. Dans l'approche bayésienne, on parle d'intervalle crédible, qui est l'équivalent d'un intervalle de confiance mais qui est basé sur l'incertitude des paramètres θ et non pas sur l'incertitude des mesures Y.

Le dénominateur $P(Y)$ dans l'équation (2.2) est la probabilité marginale d'observer la valeur Y. Il peut s'écrire $P(Y) = \int P(Y|\theta) \times \pi(\theta)\ \mathrm{d}\theta$. C'est-à-dire qu'il s'agit de la probabilité d'observer Y pour chaque valeur de θ, intégrée sur les valeurs possibles de θ pondérées par leur densité de probabilité. Par rapport à θ c'est simplement une constante (θ disparaît après intégration). Cela implique que la distribution a posteriori est proportionnelle au produit de la vraisemblance par la distribution a priori :

$$P(\theta|Y) \propto P(Y|\theta) \times \pi(\theta). \tag{2.3}$$

On connaît donc la distribution a posteriori à une constante près, égale à $1/P(Y)$. En revanche, en général on ne connaît pas cette constante, parce que l'intégration $P(Y) = \int P(Y|\theta) \times \pi(\theta)\ \mathrm{d}\theta$ est très difficile à réaliser surtout si θ est multidimensionnel, ce qui est notre cas.

L'équation (2.3) fait apparaître deux termes indispensables au calcul de la distribution a posteriori : la vraisemblance $P(Y|\theta)$ et la distribution a priori $\pi(\theta)$. Nous allons maintenant nous attacher à définir ces termes.

3.1 Vraisemblance

La vraisemblance (ou fonction de vraisemblance) est une quantité fondamentale en statistique quelle que soit l'approche. Formellement, c'est une fonction qui décrit les probabilités conditionnelles des valeurs x d'une loi statistique en fonction de ses paramètres θ. Elle s'exprime à partir de la fonction de densité $f(x|\theta)$ par la relation :

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{2.4}$$

Il faut noter que cette formule n'est valable que si on suppose que les x_i sont indépendants et identiquement distribués entre eux. En pratique, c'est la fonction qui relie les données aux paramètres du modèle. Elle représente en effet la probabilité que les données $(x_1, \ldots, x_n)$ aient été observées pour une valeur particulière des paramètres. Comme dans l'approche fréquentiste, la vraisemblance est fondamentale dans l'approche bayésienne. La définition de cette fonction constitue une des phases importantes de la modélisation. Elle consiste à construire une fonction de vraisemblance permettant de préciser les relations entre variables observées et variables explicatives. Cette relation offre une description statistique de la façon dont les données varient vis-à-vis des paramètres. Il faut noter que le choix d'une fonction de vraisemblance implique un certain nombre d'hypothèses sur la ou les origines de la variabilité des observations et que tous les résultats obtenus par la suite seront conditionnés par un tel choix. Nous verrons au chapitre suivant comment construire cette fonction à partir d'un modèle d'observation ainsi que les hypothèses qui doivent être faites sur la base des caractéristiques des données dont on dispose ainsi que de la façon dont elles ont été récoltées.

3.2 Distribution *a priori*

La loi ou distribution a priori, notée $\pi(\theta)$, représente la connaissance du modélisateur et des experts sur les paramètres d'un modèle avant qu'ils n'aient accès aux données. La notion de distribution a priori n'existe pas dans l'approche fréquentiste, étant donné que θ n'est pas, dans cette approche, une variable aléatoire. Dans l'approche bayésienne, en revanche, on ne peut pas se passer de la distribution a priori pour mener les calculs. Cette distribution est basée sur des connaissances qui peuvent avoir différentes origines ; l'expertise, la littérature mais aussi les expériences passées. Ainsi, la loi a priori peut être vue comme un résumé statistique de toute l'information disponible sur la valeur d'un paramètre au moment où on s'apprête à l'estimer en exploitant les donnéees Y.

Il faut noter ici qu'à mesure que de nouvelles données sont disponibles et que la connaissance des paramètres devient plus précise, on peut utiliser la distribution a posteriori des paramètres en cours, en tant que distribution a priori pour analyser les nouvelles données. Dans notre cas, et comme nous le verrons au chapitre 4, nous ne disposons pas réellement d'information a priori sur les paramètres de nos modèles, principalement parce que, la plupart du temps, ces paramètres n'ont pas de sens biophysique bien définis. Cependant, il est toujours possible de définir des lois a priori dites non informatives en attribuant aux paramètres des ditributions uniformes sur des plages de variations assez larges lorsque cette information fait défaut. L'intérêt pour ce travail de la possibilité d'intégrer de l'information a priori dans l'estimation des paramètres repose donc moins sur l'existence d'une expertise réelle sur les valeurs de paramètres, que sur la facilité de mettre à jour de façon itérative la distribution a posteriori en intégrant les estimations préalables sous forme de distribution a priori.

3.3 Distribution *a posteriori*, Metrolopolis-Hastings et MCMC

D'un point de vue purement théorique et hors de toute considération pratique ou calculatoire, le paradigme bayésien offre un cadre de raisonnement particulièrement adapté à

l'estimation des paramètres. Comme on l'a vu plus haut, la vraisemblance est conditionnelle aux paramètres et la distribution *a priori*, ou prior, décrit l'incertitude de l'expert sur ceux-ci. La règle de Bayes dit simplement comment réactualiser cette expertise lorsque de nouvelles données deviennent disponibles : il faut multiplier le prior par la vraisemblance pour obtenir la distribution *a posteriori*. Cependant, sur le plan pratique, lorsque le produit de la vraisemblance et du prior n'a pas de forme analytique explicite, on doit avoir recours à des méthodes d'approximation empirique de la loi a posteriori. Ces approximations reposent sur la génération d'un très grand nombre de valeurs aléatoires dans le prior, indépendantes et identiquement distribuées, et sur le calcul de leur vraisemblance associée. L'application de ces méthodes requiert donc une puissance de calcul très grande ainsi que la capacité de générer des échantillons aléatoires dans une distribution définie à une constante près et non forcément standard. C'est pourquoi, pendant plus d'une centaine d'années, les méthodes bayésiennes pour l'estimation des paramètres sont restées dans le domaine de la théorie.

L'arrivée des ordinateurs personnels modernes après la seconde guerre mondiale et la découverte des méthodes dites de *Monte Carlo* ont profondément changé la situation et rendu possible l'application du paradigme bayésien. Les méthodes de Monte Carlo permettent de générer des échantillons d'une distribution requise selon différentes modalités. Les méthodes de Monte-Carlo par Chaînes de Markov (MCMC) génèrent ces échantillons en construisant une chaîne de Markov. La principale différence entre une méthode de Monte Carlo classique et une méthode MCMC réside donc dans le caractère indépendant des échantillons générés. En Monte Carlo, les échantillons sont dit *iid* (indépendants et identiquement distribués) alors qu'en MCMC ils forment une chaîne de Markov. La seconde différence est qu'en MCMC on doit évaluer les échantillons générés et choisir ou non de les accepter pour la définition de la distribution a posteriori. Il y a plusieurs façons de construire ces chaînes, mais toutes, y compris l'échantillonneur de Gibbs (Geman and Geman, 1984), sont des cas particuliers du cadre général de Metropolis et al. (1953) et Hastings (1970).

Nous ne rentrerons pas ici dans les détails d'application de l'algorithme de Metropolis-Hastings. Cependant, il s'écrit si simplement qu'on ne résiste pas à l'envie de vous le faire partager. De plus, il est bon de savoir que c'est un algorithme itératif qui fait appel à un très grand nombre de tirages aléatoires (ou plutôt "pseudo-aléatoires", en toute rigueur) dans la distribution a priori et que la façon de générer des valeurs de paramètres candidates ne se fait pas totalement *au hasard* mais passe par la définition, et potentiellement son adaptation *à la volée* (Atchadé and Rosenthal, 2005; Gouache et al., 2013), d'une fonction de transition $\phi(\theta^{(n+1)}|\theta^{(n)})$ permettant de générer des valeurs de paramètres à partir d'une valeur courante. Il faut aussi indiquer que pour utiliser les résultats d'un tel algorithme, il est nécessaire d'évaluer sa convergence, autrement dit, il est impératif de vérifier que la distribution a posteriori donnée par l'algorithme ne dépend pas (ou plus) du nombre d'itérations réalisées (Gelman and Rubin, 1992).

Algorithme de Metropolis-Hastings

1. Fixer arbitrairement $\theta^{(0)}$
2. Tirer une valeur $\theta^{*(n+1)}$ dans $\phi(\theta^{(n+1)}|\theta^{(n)})$
3. Calculer le rapport ρ

$$\rho = \frac{P(Y|\theta^{*(n+1)}) \times \pi(\theta^{*(n+1)})}{P(Y|\theta^{(n)}) \times \pi(\theta^{(n)})} \times \frac{\phi(\theta^{(n)}, \theta^{(n+1)})}{\phi(\theta^{(n+1)}, \theta^{(n)})}$$

4. Si $\rho > 1$: $\theta^{(n+1)} = \theta^{*(n+1)}$
   Sinon :
   - $\theta^{(n+1)} = \theta^{*(n+1)}$ avec probabilité ρ
   - $\theta^{(n+1)} = \theta^{(n)}$ avec probabilité $1 - \rho$
5. Retourner en 1.

4 Avantages de l'approche bayésienne pour l'aide à la décision

Dans un contexte de prise de décision comme celui de la coexistence, lorsque les enjeux sont importants aussi bien en termes économiques qu'environnementaux, la quantification du risque associé à chacune des décisions en compétition est indispensable. Dans cette optique, l'approche bayésienne permet de fournir une information utile aux décideurs en mobilisant d'une part les données disponibles, d'autre part l'expertise reconnue dans le domaine considéré. Nous présentons ici les éléments qui rendent l'approche bayésienne non seulement attractive mais aussi pertinente dans le cadre de l'aide à la prise de décision en présence d'incertitudes.

4.1 Quantification des incertitudes pour toutes fonctions des paramètres

Comme on l'a vu précédemment, le cœur des méthodes utilisées dans l'approche bayésienne réside dans la recherche de la distribution de probabilité a posteriori d'une quantité (ici les paramètres d'un modèle). Une fois que cette distribution est obtenue sous forme exacte ou approchée, toute l'information utile aux décideurs est disponible. La démarche générale pour prédire des données indépendantes de celles utilisées lors de l'estimation consiste à utiliser cette distribution de probabilité en simulation pour obtenir une distribution des prédictions. Cette distribution est qualifiée de *distribution prédictive a posteriori* et représente l'incertitude sur une observation future conditionnellement aux données déjà observées. Un des principaux avantages de l'approche bayésienne pour l'aide à la décision et à l'évaluation des risques repose donc sur sa capacité à fournir une distribution de probabilité pour toutes prédictions du modèle, permettant de refléter leur incertitude.

De plus, hormis les prédictions brutes des modèles (ici le taux de pollinisation croisée, ponctuel ou moyen sur une parcelle), tout critère pouvant être calculé à partir des sorties du modèle et, plus largement toute fonction des paramètres du modèle s'obtient à partir de la distribution prédictive a posteriori ; il est donc possible et relativement aisé de déduire différents critères intéressants pour l'aide à la décision tels que :

- $P(y_s > \delta | X_s)$: la probabilité que le taux de pollinisation croisée y_s dépasse un seuil δ donné dans une configuration X_s connue.
- la distance minimale d à la parcelle OGM la plus proche requise pour que le taux de pollinisation croisée y_s reste sous un seuil δ donné avec une probabilité α donné.

4.2 Mise à jour perpétuelle de l'information

Un autre point fort de l'approche bayésienne pour l'aide à la décision réside dans la cohérence du schéma de mise à jour des connaissances disponibles sur les paramètres à mesure que de nouvelles données sont acquises. Comme on l'a entrevu dans la définition de la distribution a priori, cette mise à jour peut se faire en injectant la distribution a posteriori obtenue avec les données y dans la distribution a priori pour des nouvelles données y^*. Ce schéma permet donc une mise à jour perpétuelle de l'information sur un paramètre tant que de nouvelles données sont récoltées. Dans notre cas, l'intérêt de cette possibilité de mise à jour est basé sur le fait que, à chaque saison de culture, des parcelles conventionnelles sont récoltées et, si les procédures de traçabilité sont respectées, leur taux de pollinisation croisée est estimé par l'analyse d'échantillons prélevés dans la récolte. Cela implique effectivement que chaque année de nouvelles données sont disponibles et qu'on peut refaire l'estimation des paramètres sur ces nouvelles données en utilisant comme distribution a priori la distribution a posteriori obtenue sur les données de l'année passée.

4.3 Modélisation hiérarchique

Un troisième point qui rend les approches bayésiennes attractives dans le cadre de l'aide à la décision et plus spécifiquement dans le cas où il existe plusieurs sources de données exhibant de fortes hétérogénéités, est la grande souplesse qu'elles offrent pour la définition et l'estimation de modèles hiérarchiques (Parent and Bernier, 2007; Boreux et al., 2009). Sans rentrer dans les détails techniques de leur définition et application, on peut néanmoins noter ici que ce type de modèles permet d'imposer une structure commune aux paramètres tout en autorisant des variations de leurs valeurs pour chaque source de données. Cette structure hiérarchique permet d'assurer une certaine cohérence entre les estimations des paramètres. En effet, malgré la forte hétérogénéité des sources de données, il est légitime de considérer que les données issues d'une source particulière devraient apporter aussi de l'information sur les données associées à une autre source. La structure hiérarchique permet de considérer que ces relations entre les données issues d'un même processus (même si ses conditions de réalisation sont différentes) existent et de s'en servir pour l'estimation des paramètres.

5 Plusieurs types d'expérience/de données

Dans la littérature sur les flux de gènes, il existe deux grands types d'approches pour *i)* mettre en place un dispositif permettant d'observer le flux de pollen ; *ii)* permettre l'aquisition de données relatives à ce processus. Cette section a pour objectif de décrire ces différents dispositifs ou expérimentations ainsi que les caractéristiques et différences dans la nature des données qui en découlent.

5.1 Types d'expériences : Mono-source vs. Multi-sources

Dans tous les cas, l'objectif d'une expérience ou d'un dispositif de suivi d'une situation réelle est toujours le même : rendre possible l'observation d'un processus et acquérir des données sur celui-ci. Dans le cadre des flux de gènes, deux grands types de dispositifs existent :

- des expériences en conditions contrôlées ;
- des dispositifs de suivi en conditions réelles.

Les expériences en conditions contrôlées sont, la plupart du temps, mises en place avec un ou des objectifs donnés. Ces expériences peuvent avoir des objectifs finaux assez différents, cependant leur caractéristique commune repose toujours sur l'aquisition d'un maximum d'informations concernant la situation et le processus qu'on cherche à appréhender. Ces informations doivent ensuite pouvoir être utilisées pour permettre de modéliser le processus observé dans une condition donnée. De ce fait, les expériences en conditions contrôlées sont conçues pour extraire le plus possible d'informations pertinentes et limiter au maximum les sources de variation indépendantes du processus d'intérêt. La quantité de données issues de ces expériences est donc relativement grande par rapport à celle obtenue dans d'autres types de dispositifs. Les figures 2.2 et 2.3 sont des représentations schématiques de ces expérimentations en conditions contrôlées. On peut voir que dans toutes les situations représentées dans ces figures (champs A et B figure 2.2 et sites 1, 2, 3, 4, 5, 6 figure 2.3), il y a toujours une parcelle centrale (émettrice) entourée d'une ou plusieurs parcelles réceptrices. Ces configurations, qualifiées de *mono-source*, résultent d'une approche réductionniste du phénomène, considérant, qu'à partir d'une connaissance suffisament fine des processus impliqués dans ce type de configuration, il suffit d'extrapoler pour permettre la prise en compte d'une configuration plus complexe dite *multi-sources* dans laquelle plusieurs parcelles émettrices entrent en jeu.

Les dispositifs de suivi en conditions réelles sont de natures très différentes. Ils ne présupposent aucune conception préalable et correspondent, comme leur nom l'indique, à des situations réelles de coexistence entre variétés GM et variétés non GM dans un bassin de production. Les principales différences entre ce type de dispositif et les configurations *mono-source* décrites plus haut reposent sur *i)* la multiplicité des parcelles émettrices ; *ii)* la faible quantité de données récoltées. La figure 2.4 est une représentation schématique d'un de ces dispositifs de suivi en conditions réelles.

On peut donc désormais être plus précis sur les deux grands types d'approches décrits plus haut :

- Des expériences en conditions contrôlées avec une seule parcelle émettrice et une ou plusieurs parcelles réceptrices. C'est le dispositif le plus courant, il est qualifié de *mono-source*.
- Des dispositifs de suivi en conditions réelles avec plusieurs parcelles émettrices à l'échelle d'un bassin de production dans des situations réelles de coexistence. Ces dispositifs, qualifiés de *multi-sources*, sont très rares.

5.2 Types de données : discrètes vs. continues

En fonction de l'expérience d'où sont issues les données, on peut avoir affaire à des données de natures très différentes. En effet, selon le type de dispositif, la façon dont les données sont récoltées, les moyens de mesure employés, les données peuvent êtres des comptages (et donc des données discrètes) de grains portant le transgène ou des proportions de génome (et donc des données continues). La figure 2.1 donne un aperçu rapide de ces différences. Elle permet notamment d'entrevoir les problèmes que ces différences vont poser pour la définition d'un modèle d'observations. On verra par la suite que cette dichotomie (discret/continu) oblige à prendre en compte ces deux types de données au travers de modèles différents.

Comme dit plus haut, le type de données peut dépendre du dispositif. En effet, dans les dispositifs de suivi en conditions réelles, des variétés GM sont en coexistence avec des variétés non GM de la même espèce, et il n'y a aucun critère morphologique qui peut permettre de différencier un grain portant le transgène d'un grain n'en portant pas. De ce fait, dans ces dispositifs, on a toujours affaire à des proportions de génome obtenus par la technique de réaction en chaîne par polymérase (PCR, Saiki et al. (1985)), donc des données continues. La figure 2.4 est une illustration de cette situation (i.e. *multi-sources* avec PCR)

La réciproque de cette remarque n'est pas vraie : on n'a pas toujours des données discrètes lorsque les données sont issues d'une expérience *mono-source*. En effet, lorsque l'expérience est réalisée en conditions contrôlées, on peut très bien s'affranchir de l'utilisation de variétés GM en utilisant une variété possédant un allèle dominant pour le gène de la couleur. Dans ce cas où l'OGM est "simulé" par un marqueur coloré, il n'est pas nécessaire d'avoir recours aux techniques de PCR. Un simple (mais coûteux en temps) comptage des grains colorés par le marqueur suffit à déterminer le nombre de grains qui portent le caractère transmis par l'émetteur (ici la couleur) qu'on assimile à la présence du transgène. La figure 2.2 illustre ce type de situation (i.e. *mono-source* avec marqueur coloré). Cependant, l'utilisation du marqueur coloré n'est pas universelle et il existe des situations où les expériences de types *mono-source* sont réalisés avec des variétés GM. Dans ce cas, comme dans le cas *multi-sources*, il n'y a pas de critère permettant de déterminer si un grain porte le transgène ou non. La figure 2.3 est une illustration de cette situation (i.e. *mono-source* avec PCR).

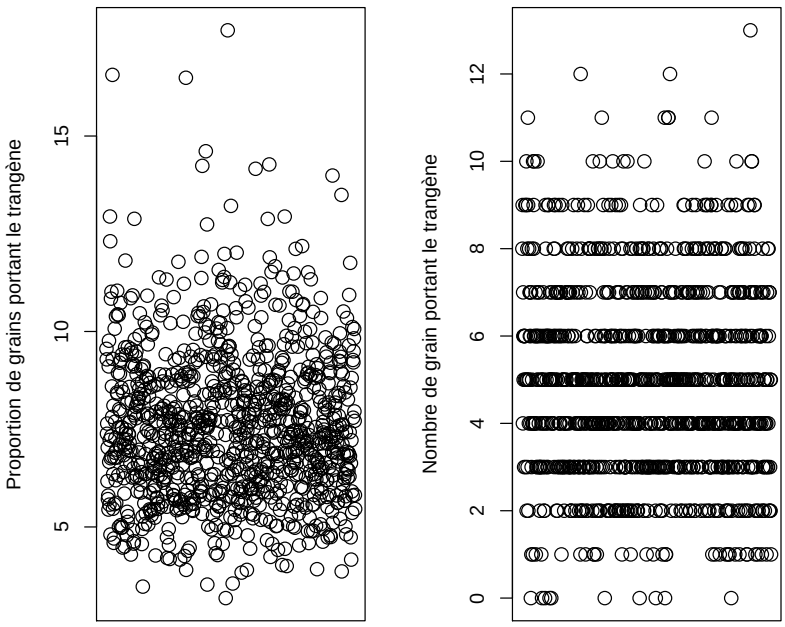


FIGURE 2.1: Illustration des différences entre données discrètes et données continues

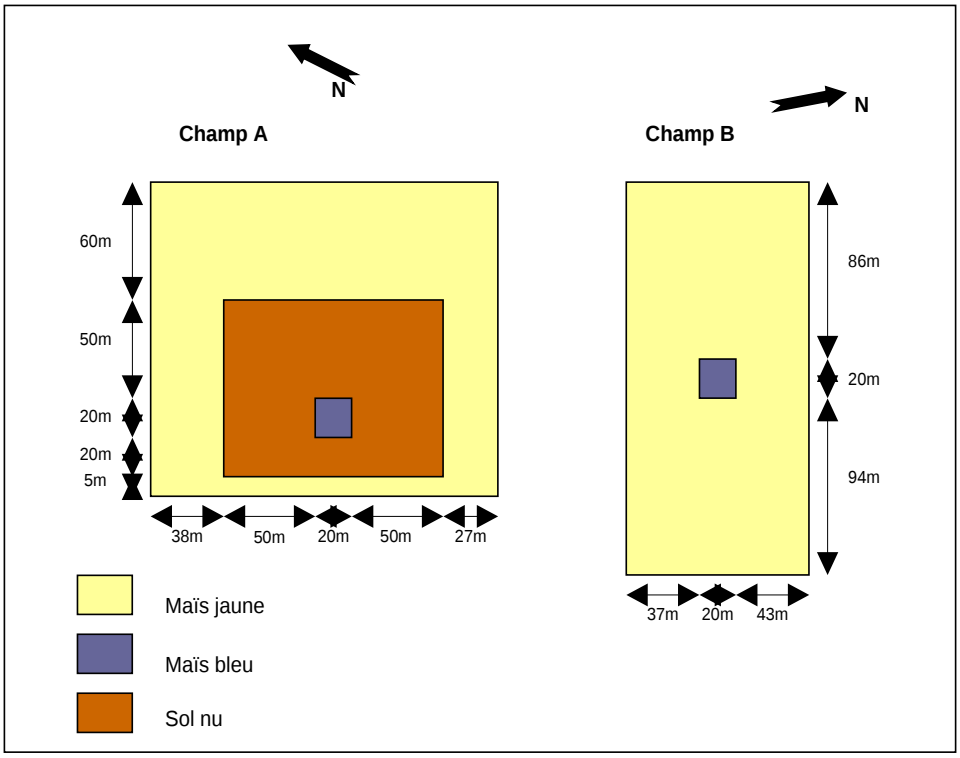


FIGURE 2.2: Schéma de dispositifs expérimentaux *mono-source* type *marqueur coloré* (*Source* : E. Klein (non publié))

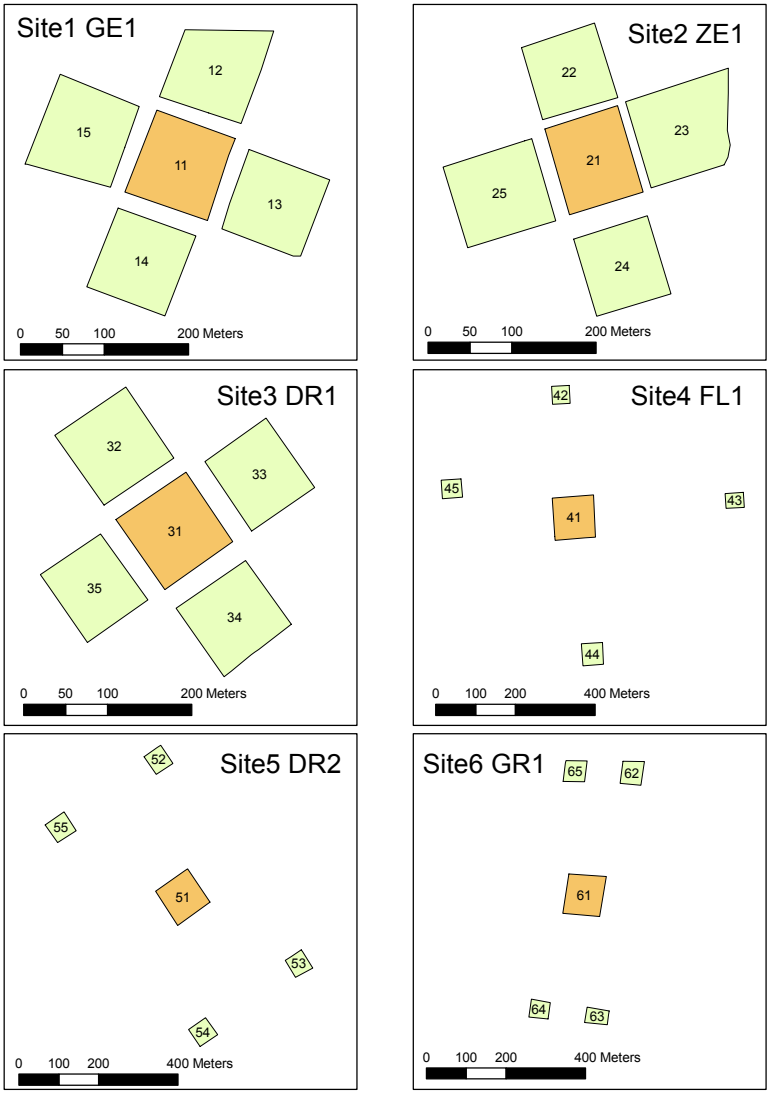


FIGURE 2.3: Schéma de dispositifs expérimentaux *mono-source* type *PCR* (*Source* :Van De Wiel et al. (2009))

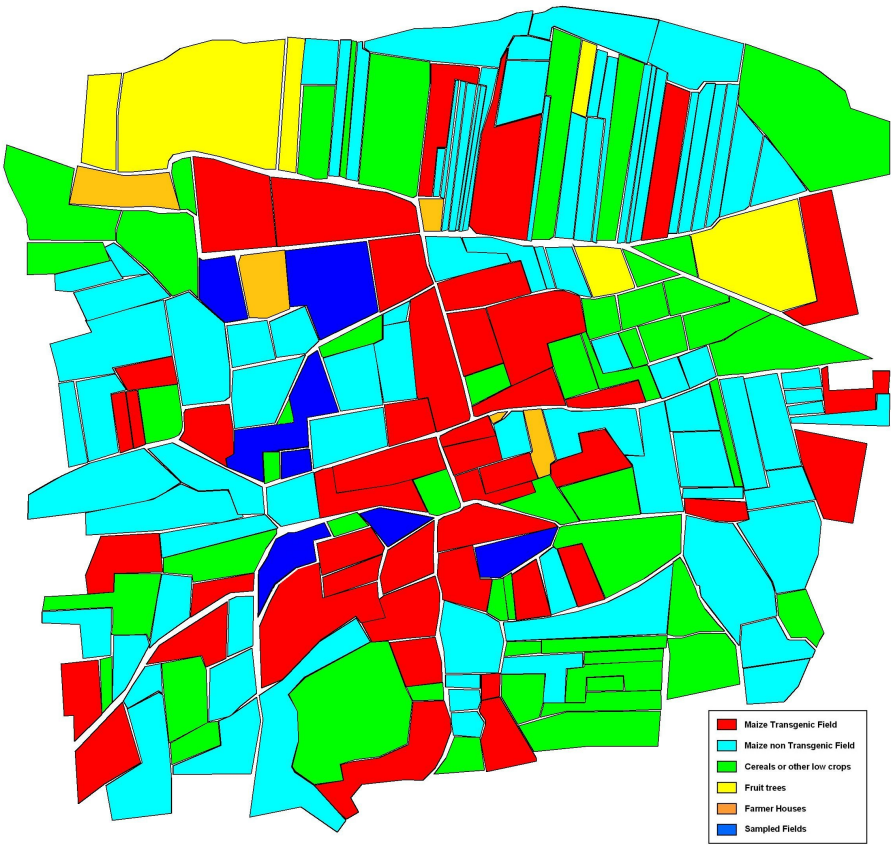


FIGURE 2.4: Schéma de dispositif de suivi *multi-sources* type *PCR* (*Source* : Messeguer et al. (2006)).

5.3 Types de configurations : trop de diversité ?

Outre la dichotomie entre types d'expériences et types de données récoltées, on observe une très grande diversité de configurations au sens large. Il existe en effet des dispositifs dits continus, dans lesquels les parcelles émettrices et réceptrices sont contigües, et des dispositifs discontinus qui, comme leur nom l'indique, présentent des discontinuités entre les émetteurs et les récepteurs. Un autre point concerne le type et l'effort d'échantillonnage. Il n'existe pas de réel consensus sur la meilleure façon de procéder et, de ce fait, on observe à une diversité de type (aléatoire, en transect, stratifié) et d'effort (nombres de points échantillonnés, nombres d'épis par point) d'échantillonnage. De la même façon, les protocoles de détermination des dates de floraison peuvent être très différents d'une expérience à l'autre.

Le tableau 2.2 est un extrait (traduit en français) d'un livrable du projet PRICE (Ce livrable est disponible en intégralité à l'annexes 5). Il fait l'inventaire des jeux de données disponibles dans la littérature et consolidés dans le cadre du projet européen SIGMEA (Messéan et al., 2009), ainsi que de leurs grandes caractéristiques : type de dispositifs, effort d'échantillonnage, données récoltés et protocole pour l'estimation du taux de présence fortuite. Ce tableau illustre la grande diversité de configurations à laquelle on est confronté. Concrètement, sur les onze jeux de données présentés, on n'a jamais deux fois la même configuration ; lorsque les dispositifs se ressemblent, c'est la manière de récolter les données qui diffère, et vice versa.

La diversité de ces jeux de données constitue un des challenges auxquels nous sommes confrontés dans la mise en oeuvre de l'approche bayésienne. En effet, il semble naturel de prendre en compte l'ensemble des jeux de données disponibles pour estimer les paramètres des modèles. Néanmoins, les cas où une méta-analyse bayésienne a été conduite (Riesgo et al., 2010), les protocoles étaient relativement homogènes. Compte tenu des autres contraintes liées à l'application de l'approche bayésienne à notre cas particulier (intégration spatiale notamment), je me suis limité dans un premier temps à un sous-ensemble de jeux de données.

Source	Site-Année	Type	Échantillonnage	Données			
				Distance	Vent	Floraison	Estimation AP
SIGMEA	Allemagne 2000	DC1	96 points (8 à 30 épis prélevés par point)	Distance du point d'échantillonnage au bord de la parcelle émettrice (de 3m à 50m)	Direction et vitesse (par 1/2h)	Période de floraison mâle renseignée	Semis des grains échantillonnés et application d'herbicide, => AP = proportion de plants résistants
	Allemagne 2001	DD2	30 points d'échantillonnage (~30 épis prélevés par point)	Distance bord à bord (de 50m à 200m) et distance du point d'échantillonnage au bord parcelle réceptrice (0,75m à 49,25m)		Pas d'élément sur les floraisons	
AIP OGM	France (Montargis) 1998	DC1	2937 points	Distance du point d'échantillonnage au bord de la parcelle émettrice (0.4 à 71.9m)	Direction et vitesse (toutes les 3 heures)	Date de floraison mâle et femelle (sans info sur la méthode)	Comptage des grains bleus dans les épis jaunes
	France (Montargis) 1999	DC1 DD1	DC1 :4430 points DD1 : 4924 points	Distance du point d'échantillonnage au bord de la parcelle émettrice (0 à 140m)	Direction et vitesse horaire	Nombre de plantes en floraison sur 100 plantes	Comptage des grains bleus dans les épis jaunes
WUR	Pays-Bas 2006-2007	DD1	5 épis/Blocs 16 blocs ou 21 blocs	Distance bord à bord (25 et 250m) et distance du bloc d'échantillonnage au bord parcelle réceptrice (de 0 à 50m)	Direction et vitesse horaire	Date de floraison mâle et femelle (40 plantes évalués 3 fois par semaines)	PCR
JKI	Allemagne 2005-2006-2007	DD2	Entre 94 et 280 points d'échantillonnage (20 épis par point)	Distance bord à bord (de 24m à 102m) et distance du point d'échantillonnage au bord parcelle réceptrice (0m à 154m)	Direction et vitesse horaire	Nombre de plantes ayants acquis un certain stade (20 à 60 plants considérés dans l'émetteur et 53 à 96 dans le récepteur, évalués tous les 2 à 3 jours)	PCR
	Allemagne 2007-2008	DD2	20 épis/points 4 point/distance 9 distances =>720 épis	Distance bord à bord (25m) et distance du point d'échantillonnage au bord parcelle réceptrice (0m à 60m)			PCR
IPSS	Portugal 2002-2003	DD2	10 épis par rang, 6 rangs (rang impaires)	Distance bord à bord (de 40m à 250m)	Direction et vitesse (moyenne et maximum sur période de floraison)	Indirectement observé par les fréquences de grains jaunes dans chaque partie de l'épi	Comptage des grains jaunes dans épis blancs + PCR
	Portugal 2005-2007	DC2	2 épis/point sur grille de 9x9m à 36x18m ou 24x12m	Distance du point d'échantillonnage au bord de la parcelle émettrice			PCR
IRTA	Espagne 2004	DC1	*	Distance du point d'échantillonnage au bord de la parcelle émettrice	Mesuré en « wind run values » exprimé en km de vent	Dates de floraison mâle et femelle évaluées visuellement sur 150 plantes (par champ)	Comptage des grains jaunes dans épis blancs + PCR
	Espagne 2004-2008	MSR	**	Distances du point d'échantillonnage au bord des parcelles émettrices	Direction et vitesse (moyenne et maximum par jour)	Dates de floraison relative au 1er champ à fleurir	PCR

TABLE 2.2: Table des caractéristiques des jeux de données rassemblés pour le projet SIGMEA. (Extrait traduit du livrable D4.25 du projet PRICE, disponible en intégralité à l'annexes 5)

Chapitre 3

Modélisation

Table des matières

1 Représentations des échanges de pollen entre parcelles 51
1.1 Mode global 51
1.2 Mode individuel 53
2 Modèle de l'espérance - Noyaux de dispersion 54
2.1 Noyau de Student bivarié (2Dt) 54
2.2 Noyau Normal Inverse Gaussian (NIG) 55
2.3 Noyau Compound Exponential (CEx) 55
3 Intégration de covariables au modèle 56
3.1 Direction et force du vent 56
3.2 Décalage de floraison 57
4 Modèles d'observation 59
4.1 Cas discret 59
4.2 Cas continu 61
5 Modèle d'ensemble et DAG 62
5.1 Modèle d'ensemble 62
5.2 DAG - Définitions et formalisme 63
5.3 Représentations graphiques 63
6 Maillage adaptatif pour l'approche individuelle 68
6.1 Problématique 68
6.2 Grille de référence et recherche d'approximations 68
6.3 Plan d'expérience 70
6.4 Protocole de simulation 70
6.5 Mise en œuvre et exemples de grilles dégradées 72
6.6 Résultats 74
6.7 Application 75

Dans ce chapitre, nous décrivons les éléments constitutifs des modèles de prédiction du taux de pollinisation croisée développés dans cette thèse pour répondre aux objectifs et enjeux définis au chapitre 1. Nous présentons tout d'abord les éléments qui constituent la composante déterministe des modèles de dispersion et qu'on appelle aussi le *modèle de l'espérance* : modes global ou individuel de représentation des échanges de pollen entre parcelles, noyaux de dispersion, covariables liées à la dispersion et à la dynamique de floraison. Puis, nous expliquons comment la nature des données et la façon dont elles ont été récoltées peuvent déterminer un ou plusieurs modèles d'observations, modèles que nous définissons en précisant leurs hypothèses sous-jacentes. Enfin, nous décrivons les méthodes que nous avons déployées pour pallier les problèmes de temps calcul qui se posent lors de l'utilisation du mode de représentation individuel, tout particulièrement dans le cadre d'une approche bayésienne.

1 Représentations des échanges de pollen entre parcelles

Il existe dans la littérature deux grands modes de représentation des échanges de pollen entre parcelles, un mode qualifié de *global*, et un autre qualifié d'*individuel*. La principale différence entre ces deux modes de représentation réside dans la prise en compte explicite ou non de l'ensemble des points des parcelles qui émettent du pollen. La figure 3.1 permet d'illustrer cette différence dans un cas très simple comprenant une parcelle OGM et une parcelle non OGM. Il faut noter que le choix de l'un ou l'autre des deux modes de représentation ne présuppose pas de forme particulière pour le modèle de prédiction sous-jacent. On trouve en effet, pour chacun des deux modes, différents types de formalismes qui peuvent être basés sur :

- des fonctions de dispersion (Colbach and Clermont-Dauphin, 2001; Klein et al., 2003; Damgaard and Kjellson, 2005; Angevin et al., 2008; Allnutt et al., 2008) ;
- des méthodes d'apprentissage ou "Machine Learning" (Ivanovska et al., 2008, 2009; Debeljak et al., 2012) ;
- des modèles statistiques empiriques (Riesgo et al., 2010).

1.1 Mode global

Dans le mode *global*, l'effet de l'émission de pollen par une parcelle sur la pollinisation d'un point récepteur est quantifié en résumant la surface émettrice de la parcelle par quelques variables caractéristiques de cette surface, en particulier sa plus courte distance au point récepteur. L'émission de pollen par la parcelle réceptrice n'est pas prise en compte. L'hypothèse principale de ce mode de représentation peut être formulée ainsi : "L'essentiel du processus d'émission de pollen qui a lieu dans toute la surface émettrice est bien approché par ce qui a lieu au point émetteur le plus proche du récepteur".

Ce mode est utilisé dans plusieurs études considérant des modèles de dispersion (e.g. Damgaard and Kjellson, 2005; Allnutt et al., 2008; Šuštar Vozlič et al., 2010). Ses avantages reposent sur des considérations pratiques en termes de niveau de précision requis

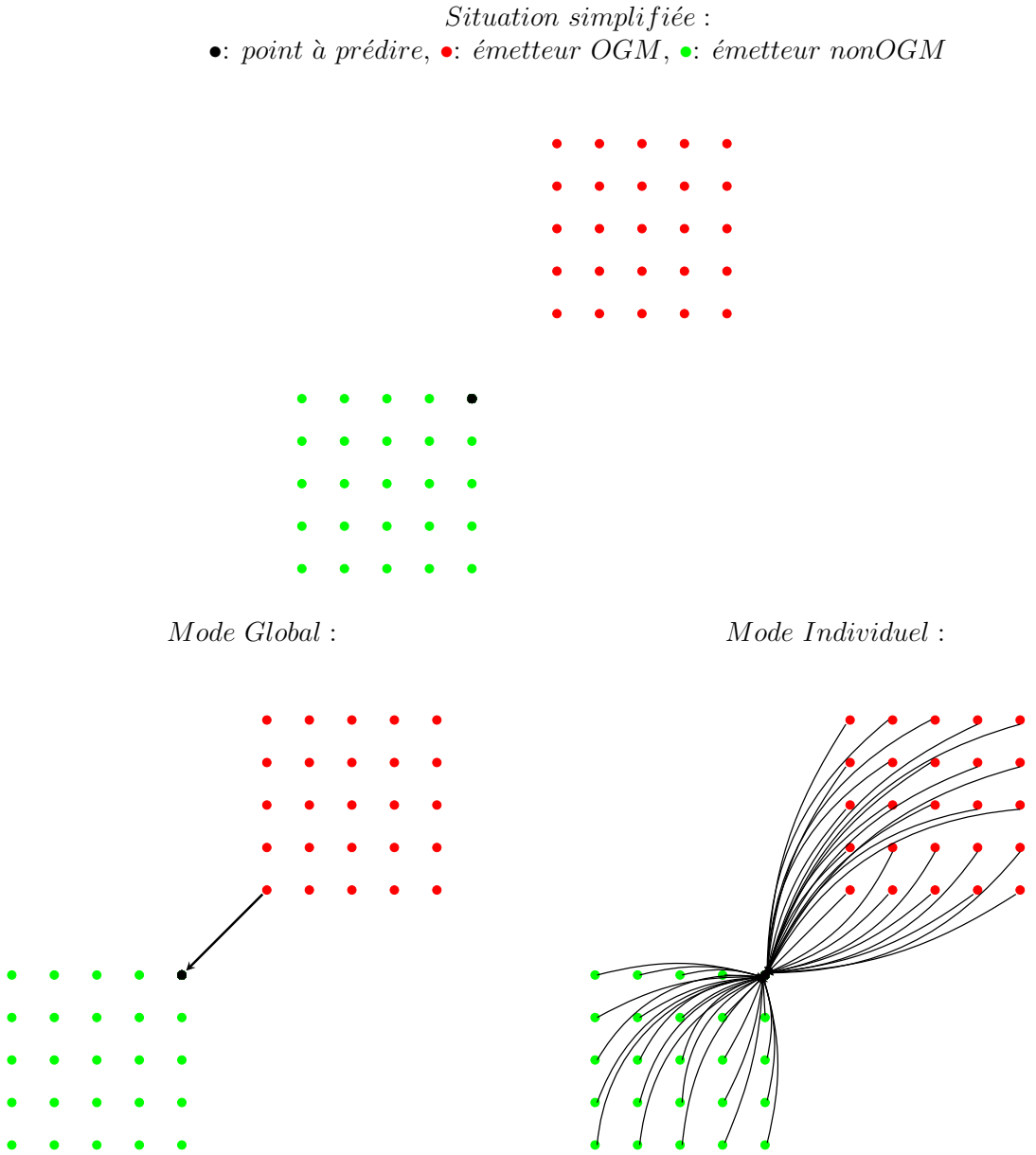


FIGURE 3.1: Représentation schématique des modes global et individuel pour la représentation des échanges de pollen entre parcelles.

et de temps calcul. En effet, sa mise en oeuvre ne nécessite qu'un seul calcul pour un récepteur donné, quelle que soit la taille de la parcelle émettrice. Son inconvénient majeur est sa faible capacité d'adaptation à des contextes différents de ceux utilisés pour la calibration. Ce faible niveau d'adaptation est dû au fait que les surfaces émettrices sont, dans cette approche, résumées par le point le plus proche du récepteur. Cette simplification du problème implique que l'estimation des paramètres se fait pour des tailles et des formes de parcelles émettrices et réceptrices données, celles des parcelles du jeu d'estimation.

Avec le mode global, l'espérance μ_s du taux de pollinisation croisée d'un épi situé en un point s dépend d'une parcelle OGM uniquement au travers des caractéristiques (distance, angle par rapport au vent dominant) de son point le plus proche. En supposant qu'il y a une seule parcelle OGM et en notant s' son point le plus proche de s, on a ainsi :

$$\mu_s = \gamma(s, s'), \tag{3.1}$$

où $\gamma(s, s')$ est la sortie d'une fonction de dispersion et représente un taux de pollinisation croisée. Il faut noter ici que des méthodes ont été développées pour atténuer la dépendance de la qualité de prédiction aux jeux de données utilisés pour l'estimation des paramètres (Klein et al., 2006a). Elles sont basées sur la modulation de l'équation 3.1 par l'intégration de facteurs liés à la taille et à la forme des parcelles émettrices (et parfois réceptrice).

1.2 Mode individuel

Dans le mode de représentation *individuel*, l'effet de l'émission de pollen par une parcelle sur la pollinisation d'un point récepteur est quantifié en intégrant les contributions individuelles des points émetteurs de pollen dans la parcelle en question. De plus, le pollen émis par la ou les parcelles non-OGM est pris en compte, à commencer par celle qui contient le point récepteur s. Cela implique donc d'une part d'avoir une connaissance plus fine des caractéristiques des deux types de parcelles (OGM et non OGM). D'autre part, la prise en compte de l'émission de pollen à partir de chaque point de chaque parcelle, qu'elle soit OGM ou non-OGM, autorise a priori une meilleure adaptation à des configurations de tailles et de formes différentes de celles utilisées pour l'estimation des paramètres.

Avec le mode individuel, l'espérance μ_s du taux de pollinisation croisée d'un épi situé en un point s dépend de la proportion de pollen reçu qui provient de plantes OGM. Pour utiliser le mode individuel dans sa version la plus simple, plusieurs hypothèses (Klein et al., 2003; Larédo and Grimaud, 2007) sont nécessaires :

1. Toutes les plantes (OGM ou non) dispersent leur pollen suivant la même fonction de dispersion γ.
2. Toutes les plantes produisent la même quantité de grains de pollen, quel que soit leur génotype.
3. Toutes les plantes ont les mêmes capacités de fécondation (même durée de viabilité du pollen, même date de floraison, même taux de fécondation).

Sous les hypothèses 1 à 3, la relation liant l'espérance μ_s du taux de pollinisation croisée en un point s aux contributions en pollen de toutes les plantes du paysage s'écrit :

$$\mu_s = \frac{\sum_{s' \in \text{GM}} \gamma(s, s')}{\sum_{s' \in \text{GM}} \gamma(s, s') + \sum_{s' \in \text{nonGM}} \gamma(s, s')}, \tag{3.2}$$

où *GM* représente l'ensemble des plantes OGM dans le paysage considéré, et *nonGM* l'ensemble des plantes non OGM.

En pratique, les hypothèses 2 et 3 peuvent assez facilement être mises en défaut. En effet, la quantité émise et la qualité du pollen varient entre variétés OGM ou non-OGM. Pour s'affranchir de ces hypothèses et prendre en compte des différences pouvant exister, soit entre la quantité de pollen produite soit entre les capacités de fécondation du pollen des variétés, on peut ajouter à la fonction μ_s un paramètre m reflétant ces différences entre variétés non-OGM et variétés OGM (Angevin et al., 2008) :

$$\mu_s = \frac{\sum_{s' \in \text{GM}} \gamma(s, s')}{\sum_{s' \in \text{GM}} \gamma(s, s') + m \sum_{s' \in \text{nonGM}} \gamma(s, s')}. \tag{3.3}$$

2 Modèle de l'espérance - Noyaux de dispersion

Un noyau de dispersion (ou fonction de dispersion) $\gamma(s, s')$ est une fonction de probabilité bidimensionnelle. Dans le contexte des flux de gènes chez le maïs, sa valeur s'interprète comme la densité de probabilité qu'un grain de pollen viable émis en s' tombe à hauteur des soies (fleurs femelles) du maïs au point s (voir Lavigne et al., 1998; Klein et al., 2003, pour plus de détails). Plusieurs noyaux de dispersion sont disponibles dans la littérature. En pratique, et en l'absence d'étude exhaustive sur les avantages relatifs d'un noyau par rapport aux autres, il nous semble utile de comparer plusieurs noyaux ayant des comportements contrastés. En particulier, une des caractéristiques importantes à considérer est le comportement du noyau à longue distance (voir Klein et al., 2006b, pour plus de détails), exhibant des queues lourdes (e.g. la fonction 2Dt), des queues légères (e.g. la fonction CEx) ou des queues que l'on peut qualifier d'intermédiaires (e.g. la fonction NIG).

2.1 Noyau de Student bivarié (2Dt)

Le noyau *Student bivarié* (2Dt) a été défini dans Clark et al. (1999) pour modéliser la dispersion de graines de conifères et repris pour une étude de coexistence par Lavigne et al. (2008). Notre intérêt pour ce noyau provient du fait qu'il présente une queue lourde. Cela signifie que, par rapport aux autres noyaux retenus, les quantités dispersées par le 2Dt seront plus faibles à courtes distances et plus fortes à longues distances. Le noyau

2Dt anisotrope s'écrit mathématiquement :

$$\gamma_{2Dt}(r,\omega) = \frac{b-1}{\pi a^2}\left(1+\frac{r^2}{a^2}\right)^{-b} e^{\kappa\, cos(\omega-\omega_0)}. \tag{3.4}$$

Dans cette formule, on considère que l'émetteur est situé en un point de coordonnées (x,y) et que le récepteur est situé à l'origine $(0,0)$, et l'on note $r=\sqrt{x^2+y^2}$ la distance émetteur-récepteur. Dans le terme associé à l'anisotropie, κ est un coefficient proportionnel à la vitesse du vent, ω représente l'angle entre l'axe des abscisses et le vecteur de coordonnées (x,y) (qui relie les points émetteur et récepteur), et ω_0 l'angle entre l'axe des abscisses et la direction du vent dominant. Il faut noter que l'équation (3.4) a été reprise telle que définie dans Lavigne et al. (2008) ; cette expression est normalisée (son intégrale somme à 1) pour le cas isotrope ($\kappa=0$) mais pas pour le cas général. Notons également que la prise en compte de l'anisotropie due au vent dans les noyaux de dispersion est rediscutée un peu plus bas.

2.2 Noyau Normal Inverse Gaussian (NIG)

Le noyau *Normal Inverse Gaussian* (NIG) a été développé à l'origine pour modéliser la dispersion de grains de sable (Barndorff-Nielsen, 1986, 1997). C'est la fonction de dispersion par défaut dans MAPOD, développée pour le pollen de maïs sur des bases quasi-mécanistes par Klein et al. (2003) et utilisée dans les études de coexistence (Angevin et al., 2002; Messéan et al., 2006). Elle présente une décroissance en loi de puissance à courte distance et une décroissance exponentielle sur de longues distances. Ce comportement représente un compromis entre les noyaux à queues lourdes et les noyaux à queues légères. Avec les mêmes conventions que pour le 2Dt, le noyau NIG anisotrope s'écrit :

$$\gamma_{NIG}(x,y) = \frac{\delta_x\delta_y e^{\lambda_z}}{2\pi}\frac{q(x,y)^{-1/2}+p^{1/2}}{q(x,y)} e^{-\sqrt{pq(x,y)}} e^{x\delta_x\lambda_x+y\delta_y\lambda_y}, \tag{3.5}$$

avec $p=\lambda_z^2+\lambda_x^2+\lambda_y^2$ et $q(x,y)=1+\delta_x^2x^2+\delta_y^2y^2$.
Les quantités $\lambda_z, \lambda_x, \lambda_y, \delta_x$ et δ_y sont des paramètres modulés par les conditions de vent comme suit :

$$\lambda_z = 0.027\tfrac{h}{0.831}\ ;\ \lambda_x = 0.165\tfrac{h}{0.831}\tfrac{\mu\cos(\theta)}{2}\ ;\ \lambda_y = 0.165\tfrac{h}{0.831}\tfrac{\mu\sin(\theta)}{2}\ ;\ \delta_x=\delta_y = 0.499\tfrac{0.831}{h}.$$

Dans ces expressions, le paramètre h représente la différence de hauteur entre les fleurs mâles des plantes GM et les fleurs femelles des plantes non GM. Les paramètres μ et θ sont respectivement la vitesse moyenne et la direction du vent mesurée à dix mètres de hauteur. Voir Klein et al. (2003) pour plus de détails.

2.3 Noyau Compound Exponential (CEx)

Le noyau *Compound Exponential* (CEx) est issu de travaux de modélisation de la dispersion des grains de pollen de colza (Damgaard and Kjellson, 2005). Il existe plusieurs différences entre la dispersion du pollen de colza et celle du pollen de maïs. Elles portent

notamment sur la possibilité d'une dispersion entomophile (transport du pollen par les insectes), quasiment jamais observée sur le maïs (Aly and Hassan, 1999), mais aussi sur la viabilité et les caractéristiques morphologiques des grains de pollen eux-mêmes. Ces différences plaident de façon générale pour des noyaux à queues plus lourdes pour le maïs que pour le colza. Néanmoins, il n'y a pas de raison objective et définitive de ne pas considérer le noyau *Compound Exponential* comme un noyau candidat pour un modèle de dispersion du pollen de maïs. De plus, ce noyau étant à queue légère, son intégration à la liste des fonctions de dispersion candidates nous permet d'avoir un "représentant" de chaque type de queue ; i.e. lourdes, légères et intermédiaires. L'expression mathématique du noyau CEx est :

$$\gamma_{Expo}(r) = \begin{cases} K_e \times e^{-a_1 r} & r \leq D \\ K_e \times e^{-a_1 D - a_2(r-D)} & r \geq D \end{cases} \tag{3.6}$$

Notons que le paramètre K_e, qui s'interpréte comme le taux de pollinisation croisée à une distance 0, n'a pas d'utilité dans le mode individuel. Notons également que le noyau CEx n'est pas un noyau exponentiel classique mais composé (*compound*) par 2 pentes distinctes (a_1 et a_2) ; cette distinction a été établie pour refléter la décroissance très rapide du taux de pollinisation croisée dans les tous premiers mètres et une décroissance plus lente sur des distances plus longues.

3 Intégration de covariables au modèle

Les trois noyaux de dispersion définis ci-dessus utilisent la distance entre l'émetteur et le récepteur comme principale covariable explicative du processus de dispersion. Les deux premiers possèdent l'avantage d'intégrer également la direction et la force du vent comme covariables, permettant ainsi de modéliser une dispersion anisotrope. Dans le but de comparer les trois noyaux, il est indispensable, en premier lieu, d'intégrer la direction et la force du vent dans le troisième noyau défini (CEx).

D'autre part, nous avons vu au Chapitre 1 que les décalages de floraison pouvant survenir entre variétés GM et variétés non GM représentent un des facteurs qui influencent la capacité du pollen OGM à féconder des soies non-OGM (Bateman, 1947b; Messeguer et al., 2006; Angevin et al., 2008; Langhof et al., 2010) et donc, indirectement, le taux de pollinisaton croisée. Nous avons donc cherché à intégrer l'effet de ce décalage potentiel à la définition du modèle de l'espérance.

3.1 Direction et force du vent

Pour intégrer la force et la direction du vent, nous avons défini et testé deux solutions :

- Moduler la distance prise en compte dans le noyau par un facteur d'anisotropie dépendant de la direction et de la force du vent. Cette solution passe par la définition d'une distance dite *efficace* r^* pour laquelle les distances émetteur/récepteur sont

raccourcies sous le vent et allongées contre le vent :

$$r^* = r \times \big(1 - \kappa \cos(\omega - \omega_0)\big), \tag{3.7}$$

où r est la distance euclidienne entre s' et s, κ est un coefficient proportionnel à la vitesse du vent, et $(\omega - \omega_0)$ est l'angle entre la direction du vent dominant et le vecteur de coordonnées (x, y) qui relie les points émetteur et récepteur.

- Moduler le noyau lui-même par un facteur d'anisotropie. Cette solution est celle adoptée par Clark et al. (1999) pour prendre en compte l'anisotropie dans le noyau 2Dt :

$$\gamma_{Aniso}(r, \omega) = \gamma_{Iso}(r) \times e^{\kappa\, cos(\omega - \omega_0)}. \tag{3.8}$$

$\gamma_{Iso}(r)$ une fonction de dispersion isotrope quelconque et $\gamma_{Aniso}(r, \omega)$ sa version anisotrope.

Ces deux solutions ont été implémentées et testées. Cependant, une analyse par simulation a montré que la première solution n'offre pas autant de flexibilité que la seconde. De plus, la seconde solution permet de comparer les noyaux 2Dt et CEx avec la même façon de prendre en compte l'anisotropie, sans revenir sur la formule de Clark pour le 2Dt. Pour ces deux raisons, nous utilisons par la suite la seconde solution pour intégrer l'anisotropie dans le noyau CEx.

3.2 Décalage de floraison

Comme évoqué précédemment, la prise en compte du décalage de floraison dans les configurations multi-sources représente un des facteurs clés pour bien décrire les situations réelles. On rappelle ici que les essais mono-sources sont réalisés le plus souvent dans un contexte de synchronisme total de floraisons, considéré comme représentant la situation de risque maximal. Cependant, ce synchronisme contrôlé en expérimentation relève plutôt de l'exception que de la norme dès lors qu'on s'intéresse à des situations réelles. Nous nous sommes donc penchés sur les études portant sur la dynamique de floraison du maïs et plus particulièrement celles qui intègrent les aspects de floraison à un calcul du taux de pollinisation croisée (principalement Messeguer et al., 2006; Angevin et al., 2008; Palaudelmàs et al., 2012). Ces travaux débouchent sur des modèles très différents aussi bien au niveau de la prise en compte des processus que de la quantité d'information requise pour initialiser le modèle.

Nous avons identifié deux modèles potentiels, intéressants vis-à-vis de nos objectifs :

- D'une part, le modèle MAPOD (Angevin et al., 2008) qui a déjà été évoqué dans le premier chapitre. Comme nous l'avons vu, ce modèle est qualifié d'intégratif car il intègre un très grand nombre de processus autres que la dispersion du pollen en elle-même. En effet, le modèle permet de simuler au pas de temps journalier la croissance des plantes pour déterminer les dates à partir desquelles les mâles émettent du pollen et les soies femelles sont réceptives. Une fois ces dates déterminées, le modèle simule la quantité de pollen produite par chaque variété à partir de caractéristiques variétales, puis ces quantités sont dispersées via une approche individuelle et matricielle basée sur une fonction de dispersion NIG. Ce modèle est très attractif si on

considère le nombre de processus pris en compte et la qualité des prédictions qui en résulte. Cependant il demeure très exigeant en termes de données nécessaires à son initialisation. De plus l'intégration de la fonction de dispersion sur des paysages réels requiert une forte puissance de calcul. Surtout, le temps calcul nécessaire rend son utilisation quasi impossible dans un contexte d'aide à la décision où le modèle doit pouvoir tourner rapidement pour permettre de tester des configurations et d'en proposer de nouvelles à partir des prédictions du modèle. Pour toutes ces raisons et malgré les efforts consentis pour le développement d'un tel modèle, je me suis concentré sur une solution moins mécaniste et plus pragmatique.

- D'autre part, le modèle Global Index (GI) développé par Messeguer et al. (2006) et qui constitue un bon exemple de pragmatisme appliqué à la modélisation. Il est basé sur le calcul d'un indice global considérant comme deux seules covariables la distance minimale entre deux parcelles et les dates de floraison de ces parcelles. Il s'écrit très simplement en trois lignes :

$$GI = \sum ECP, \tag{3.9}$$

$$ECP = \frac{DF}{DI^2}, \tag{3.10}$$

$$DF = 10 - |f_{OGM} - f_{nonOGM}|, \tag{3.11}$$

où f_{nonOGM} et f_{OGM} sont les dates de floraison des parcelles nonOGM et OGM, respectivement ; DI représente la distance minimale entre les deux parcelles ; ECP (pour *Estimated Cross Pollination*) représente la contribution d'une parcelle OGM au taux de pollinisation croisée de la parcelle conventionnelle réceptrice.

Comme on l'a dit précédemment, en restant dans une logique de proposer des modèles simplifiés vis-à-vis de l'existant, notamment par rapport à MAPOD, tout en permettant une prise en compte satisfaisante des processus clés ainsi qu'une évalution de l'incertitude associée aux prédictions, nous avons choisi de nous concentrer sur une modélisation pragmatique dans la même idée que celle du GI.

L'équation que nous proposons ici reprend l'idée de Messeguer et al. (2006) de définir l'effet du décalage de floraison par une relation simple entre les dates de floraison des parcelles émettrices et réceptrices. Néanmoins, la non linéarité de l'effet du décalage de floraison observé dans certaines expérimentations et les simulations conduites avec le modèle MAPOD nous pousse à considérer une relation légèrement plus complexe :

$$DF = \frac{1}{1 + \alpha|f_{OGM} - f_{nonOGM}|} \tag{3.12}$$

où f_{nonOGM} et f_{OGM} sont les dates de floraisons de la parcelle nonOGM et OGM, respectivement et α est un paramètre à estimer.

4 Modèles d'observation

Un modèle d'observations, également appelé modèle d'échantillonnage, permet d'établir le lien entre les données observées et la composante du modèle qui décrit le phénomène étudié, que nous venons de décrire. Il requiert une description mathématique probabiliste du processus qui a généré les observations (Boreux et al., 2009). Poser un modèle d'observations sur les données suppose que l'on considère que le processus qui y conduit est, au moins pour partie, un processus stochastique. Ce modèle prend généralement la forme d'une distribution statistique associée à des hypothèses sur les observations. Le choix de ce modèle fait clairement partie du travail de modélisation. Plusieurs considérations sont à prendre en compte dans ce choix, en premier lieu le réalisme mais aussi la parcimonie des paramètres. Il est important de rappeler que tous les résultats obtenus seront nécessairement conditionnels à l'adoption de ce modèle.

Dans les travaux sur la coexistence, on rencontre principalement deux types de données. Dans le cas discret, les observations sont des comptages visuels des grains portant le transgène dans des épis de maïs prélevés en différents sites s. Ces données seront décrites en détail dans le Chapitre 4. Dans le cas continu, les données sont des mesures PCR (Polymerase Chain Reaction) effectuées sur un broyat de grains issus de ces épis.

4.1 Cas discret

Pour définir un modèle d'observations, il faut faire des hypothèses sur les données que l'on cherche à modéliser.

Soit y_s le nombre de grains de maïs portant le transgène dans un épi situé au point s de la parcelle réceptrice et K le nombre total de grains dans l'épi. Soit $\mu_s^{'}$ le taux de pollinisation moyen en s prédit par le modèle de dispersion (la distinction entre $\mu_s^{'}$ et μ_s est expliquée plus bas). Supposons que :

1. La probabilité qu'un grain de pollen OGM féconde un ovule situé au point récepteur s est égale à $\mu_s^{'}$.
2. Les grains de pollen fécondent les ovules d'une même plante de façon indépendante de leur génotype et indépendante entre ovules.
3. Les variables aléatoires y_s sont indépendantes.

Ces hypothèses d'indépendance proviennent du fait que les grains de pollen dans le nuage pollinique sont très nombreux et considérés comme non limitants.

Alors les variables aléatoires y_s sont distribuées selon une loi binomiale :

$$y_s|K, \mu_s^{'} \sim \mathcal{B}(K, \mu_s^{'}). \tag{3.13}$$

L'événement d'intérêt, ici le succès d'une fécondation entre un grain de pollen OGM et un ovule non-OGM, peut être considéré comme un événement *rare* si $y_s \ll K$, c'est-à-dire

$\mu_s^{'} \to 0$. En posant $K \to \infty$ et $\mu_s^{'} \to 0$, on obtient alors la distribution de Poisson :

$$y_s|K, \mu_s^{'} \sim \mathcal{P}(K\mu_s^{'}). \tag{3.14}$$

L'hypothèse 1. peut être facilement mise en défaut : si la fécondation est entomophile, c'est-à-dire assurée par des insectes pollinisateurs (ce qui n'est pas le cas du maïs), un insecte peut féconder plusieurs ovules en une seule visite ce qui introduit une corrélation entre les génotypes des grains issus d'une même plante. Si la fécondation est anémophile, c'est-à-dire assurée par le vent, et s'il existe une hétérogénéité temporelle du vent pendant la période de floraison accompagnée d'une variabilité dans la période de fertilité des ovules, les grains issus d'une même plante pourraient être fécondés dans les mêmes conditions de vent et donc avoir des génotypes corrélés.

Une simple représentation des données dont on dispose ainsi qu'une revue de littérature sur la modélisation statistique des données de comptage montrent que les comptages de grains GM dans un épi de maïs conventionnel présentent une très grande variabilité, avec un peu d'épis exceptionnels (près de 100% de grains GM) et de nombreux épis avec zéro grain GM (Goedhart et al., 2014). On observe donc de la sur-dispersion par rapport à la distribution de Poisson, et en particulier un excès de zéros (Kuhnert et al., 2005b; Sileshi, 2008). Ces éléments nous a conduisent à chercher un moyen d'intégrer la sur-dispersion dans le modèle d'observations. La distribution de Poisson zéro inflatée (ZIP) nous est un bon candidat pour faire face à l'excès de zéro (Kuhnert et al., 2005b; Goedhart et al., 2014) puisqu'elle consiste en un mélange d'une distribution de Poisson et d'une distribution de Dirac en zéro. La ZIP suppose que, avec probabilité p, la seule observation possible est un zéro et, avec probabilité $q = 1 - p$, le modèle d'observation est une distribution de Poisson (voir Lambert, 1992, pour une description complète).

$$y_s|q_s, K, \mu_s^{'} \sim ZI\mathcal{P}(1 - q_s, K\mu_s^{'}) \tag{3.15}$$

$$\Leftrightarrow$$

$$\begin{aligned} y_s|Z_s = 0 &\sim \delta(\{0\}) \\ y_s|Z_s = 1 &\sim \mathcal{P}(K\mu_s^{'}). \end{aligned}$$

Pour estimer le poids des zéros dans le mélange de la ZIP, on définit Z comme une variable cachée distribuée selon une loi de Bernoulli :

$$Z_s|q_s \sim \mathcal{B}ern(q_s), \tag{3.16}$$

avec un lien logistique à la distance r à l'émetteur OGM le plus proche :

$$\text{logit}(q_s) = \beta_1(\beta_2 - r),$$

où β_2 représente l'abscisse du point d'inflexion et β_1 est proportionnel à la pente de la tangente au point d'inflexion. En tout, le modèle ZIP a deux paramètres (β_1 et β_2).

Par ailleurs, les travaux de Larédo and Grimaud (2007) ont montré qu'il pouvait exister une variabilité résiduelle entre plantes qui n'est prise en compte ni par la variabilité de la loi de Poisson ni par celle ajoutée par le modèle ZIP. Afin d'en tenir compte, nous

proposons d'introduire un terme de variance supplémentaire en considérant que $\mu_s^{'}$ est une variable aléatoire centrée en la valeur μ_s donnée par le modèle de dispersion. On obtient alors deux cas possibles selon que l'on considère $\mu_s^{'}$ comme fixé ou aléatoire :

$$\mu_s^{'} = \mu_s \tag{3.17}$$

$$\mu_s^{'} \sim \mathcal{N}(\mu_s, \sigma^2) \tag{3.18}$$

Pour le cas discret, nous avons retenu deux principaux modèles d'observations possibles ; le modèle de Poisson et le modèle de Poisson inflaté en zéro. Nous retenons aussi les deux variantes ci-dessus sur la définition de l'espérance $\mu_s^{'}$, qui sont applicables dans les deux cas (Poisson et ZIP).

4.2 Cas continu

Le cas des données continues correspond à des mesures PCR d'épis échantillonnés dans un champ non OGM. En un point donné, cette mesure correspond à la fréquence d'allèles transgéniques détectée dans une farine de grains de maïs issue du mélange puis du broyage de trois épis voisins. Le modèle d'observations pour les données PCR s'écrit :

$$log(y_s) \sim \mathcal{N}(\mu_s, \sigma^2), \tag{3.19}$$

avec y_s le taux de PCR mesuré au point s et μ_s la valeur de la fonction de dispersion en ce point. Ce modèle d'observations fait apparaître un paramètre supplémentaire σ^2 par rapport au modèle de Poisson. Il faut noter ici que l'erreur de mesure sur la PCR est, dans un premier temps et pour simplifier le problème, volontairement négligée. En effet, si on note z_s la vraie fréquence d'allèles transgéniques au point s, on aurait :

$$log(y_s) \sim \mathcal{N}(z_s, \sigma^2_{PCR}) \tag{3.20}$$

et

$$log(z_s) \sim \mathcal{N}(\mu_s, \sigma^2) \tag{3.21}$$

Il est important de noter ici que la PCR est une méthode d'analyse qui permet *i)* de détecter la présence d'un certain type d'ADN dans une séquence donnée à partir d'un seuil : le seuil de détection S_d ; *ii)* de quantifier la proportion de cet ADN à partir d'un autre seuil, le seuil de quantification S_q. On comprend aisément que $S_d \ll S_q$ mais aussi et surtout que les zéros contenus dans les observations ne sont pas forcément de *vrais* zéros, une partie de ces derniers correspondant simplement à une valeur inférieure à l'un des deux seuils. Les modèles de censure permettent de prendre en compte cette structure. Malheureusement et principalement par manque de temps, nous n'avons pas pu mettre en place ce type de modèle et ne chercherons donc pas ici à le définir plus formellement.

5 Modèle d'ensemble et DAG

5.1 Modèle d'ensemble

Dans les précédentes sections de ce chapitre, nous avons défini plusieurs éléments constitutifs des modèles de prédiction du taux de pollinisation croisée. Nous les rappelons brièvement ici :

- le mode de représentation de la dispersion (global ou individuel) ;
- la fonction (ou noyau) de dispersion γ ;
- les covariables X à intégrer aux fonctions de dispersion ;
- le modèle d'observations (composante stochastique du modèle).

Le modèle de prédiction que nous proposons ici peut être vu comme une sorte de métamodèle, au sens d'un modèle dont une partie des entrées est représentée par des sous-modèles et dont la sortie est un modèle de prédiction particulier. Les métavariables de ce métamodèle sont les éléments de la liste ci-dessus, pour chacun desquels plusieurs choix sont possibles. Ainsi, un modèle de prédiction, tel que proposé ici, se définit par une combinaison des modalités de facteurs définis tout au long de ce chapitre. La figure 3.2 présente un schéma de ce métamodèle pour aider à la compréhension de sa construction.

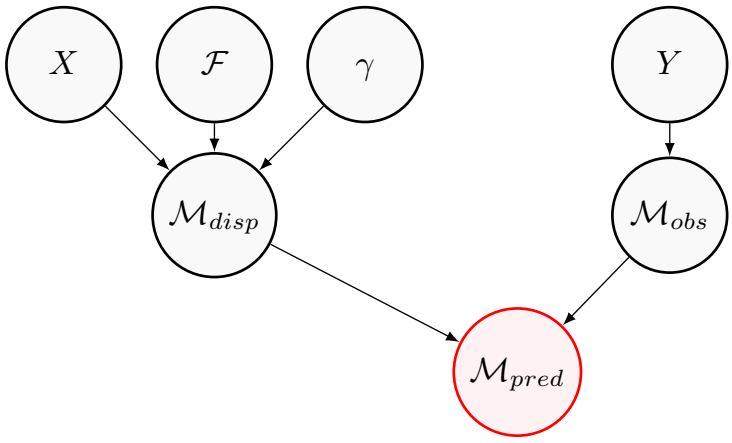


FIGURE 3.2: Réprésentation schématique du métamodèle de prédiction.

En remontant le schéma de bas en haut, on constate que notre modèle de prédiction peut-être vu comme une fonction de deux variables principales : un modèle de dispersion $\mathcal{M}_{disp}$ et un modèle d'observations $\mathcal{M}_{obs}$. Le modèle de dispersion $\mathcal{M}_{disp}$ se construit à partir de trois éléments : une fonction de dispersion γ (cf section 2 de ce chapitre), un mode de représentation de la dispersion $\mathcal{F}$ (cf section 1 de ce chapitre), et un jeu de covariables X (cf section 3 de ce chapitre). Le modèle d'observations $\mathcal{M}_{obs}$ se détermine à partir de la nature des données à prédire Y (cf section 4 de ce chapitre).

Dans cette thèse, et tout au long de ce chapitre, nous avons défini et utilisé un ensemble de choix possibles pour chacune des métavariables entrant dans la définition du métamodèle de prédiction, il est important de noter ici que ces ensembles n'ont pas un caractère exhaustif et pourraient tout à fait être augmentés par des éléments nouveaux. Nous rappelons à présent les ensembles dans lesquels les métavariables du métamodèle de prédiction prennent leur valeurs :

$$X \in \{d; (d, \omega_0, \kappa);\ (d, f_{flo});\ (d, \omega_0, \kappa, f_{flo})\},$$

$$\gamma \in \{\text{2Dt},\ \text{NIG},\ \text{CEx}\},$$

$$\mathcal{F} \in \{individuel;\ global\},$$

$$\mathcal{M}_{obs} \in \{\mathcal{P};\ \mathcal{P}+\mathcal{N};\ ZI\mathcal{P};\ \ ZI\mathcal{P}+\mathcal{N};\ \mathcal{N};\ \mathcal{N}+\mathcal{N}\}.$$

5.2 DAG - Définitions et formalisme

Le DAG pour *Directed Acyclic Graph* (Graphe acyclique orienté, en français) est un outil de visualisation largement utilisé en modélisation bayésienne (Spiegelhalter, 1998; Parent and Bernier, 2007). Il permet de représenter de façon intuitive les relations entre les observations, les variables et les paramètres des modèles (Makowski et al., 2008; Boreux et al., 2009). Il nous semble donc très utile, notamment à des fins de compréhension du modèle par tout public, y compris non statisticien, de représenter l'ensemble des modèles définis et décrits dans les parties précédentes à l'aide d'un DAG.

Quelques éléments doivent êtres expliqués au préalable pour permettre ou faciliter la compréhension du DAG. Les quantités incertaines constituent des nœuds stochastiques et sont représentées par des cercles ◯. Les paramètres du modèle sont les nœuds sans parent, ce sont les inconnues du problème. Les observations sont les nœuds sans enfants. Il existe aussi des quantités déterministes qui influencent le système mais qui ne nous intéressent pas en tant que telles ; ce sont les covariables des modèles de dispersion (distance et vent) que l'on suppose connaître de façon précise et sur lesquelles on n'exerce pas de contrôle direct. De telles quantités sont représentées à l'intérieur de carrés □. Les flèches représentent une relation entre deux quantités. Ces relations peuvent êtres déterministes ou stochastiques. Une relation déterministe entre deux quantités est réprésentée par une flèche en pointillets, une relation stochastique est réprésentée par une flèche en trait plein. Ainsi, l'élément du graphe $A \rightarrow B$ s'interprète comme : B dépend de A par une relation stochastique (e.g. $B \sim \mathcal{B}ern(A)$). L'élément $A \dashrightarrow B$ s'interprète comme : B dépend de A par une relation déterministe (e.g. $B = \frac{1}{1+A}$).

5.3 Représentations graphiques

Nous présentons ici les DAG pour les modèles d'observations dans le cas discret. Il faut noter que, dans ce type de représentation, la partie déterministe du modèle n'apparaît

qu'en filigrane. En effet, on ne peut y entrevoir la fonction de dispersion qu'au travers de ses paramètres. De la même façon, les différents modes de prise en compte de la dispersion n'apparaissent pas clairement, ils sont en fait cachés par la relation qui lie μ aux paramètres de la fonction de dispersion.

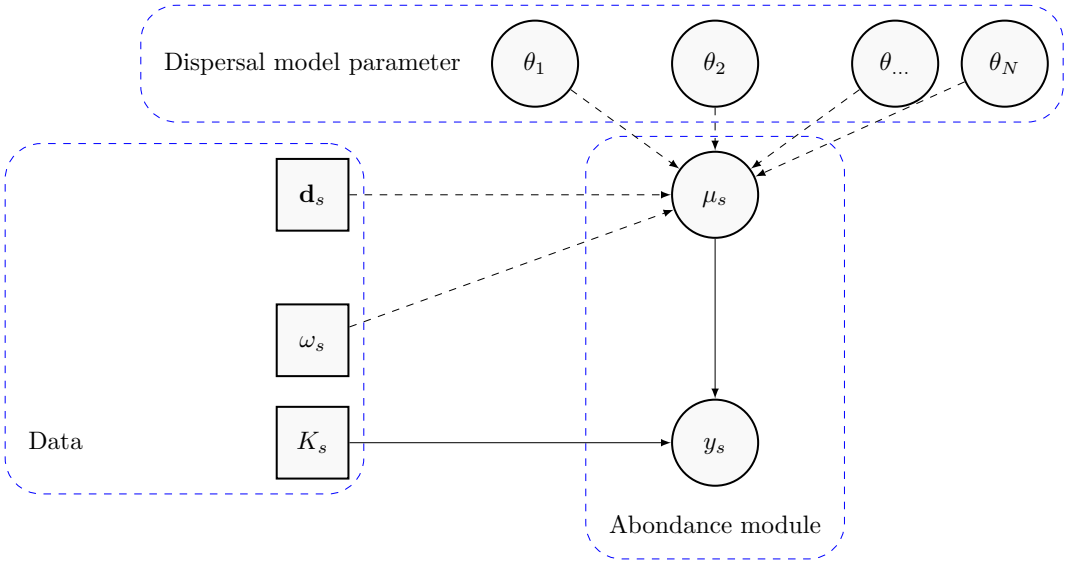


FIGURE 3.3: Graphe acyclique orienté du modèle de Poisson à espérance fixe.

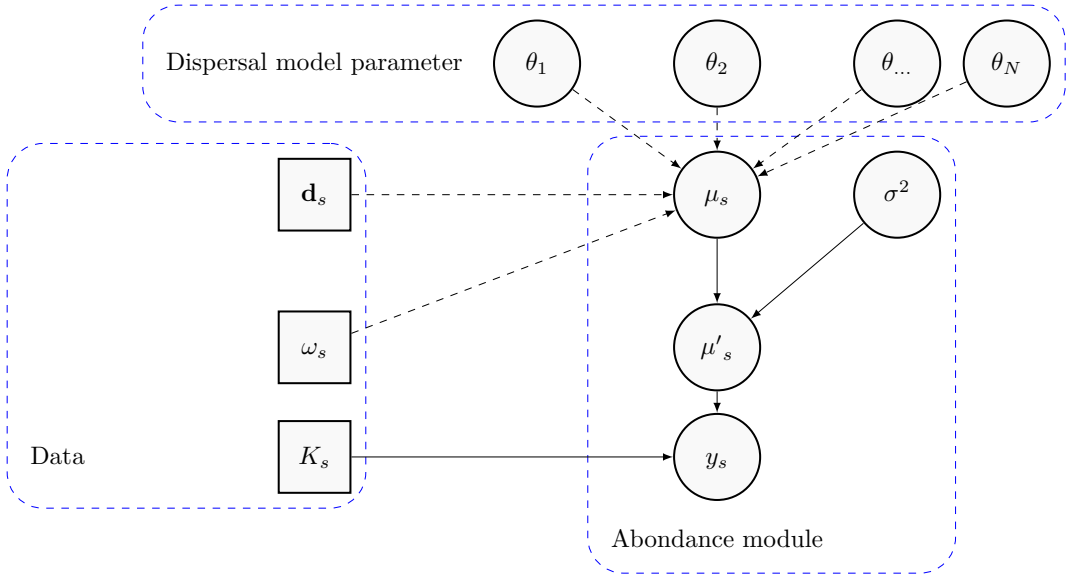


FIGURE 3.4: Graphe acyclique orienté du modèle de Poisson à espérance normale.

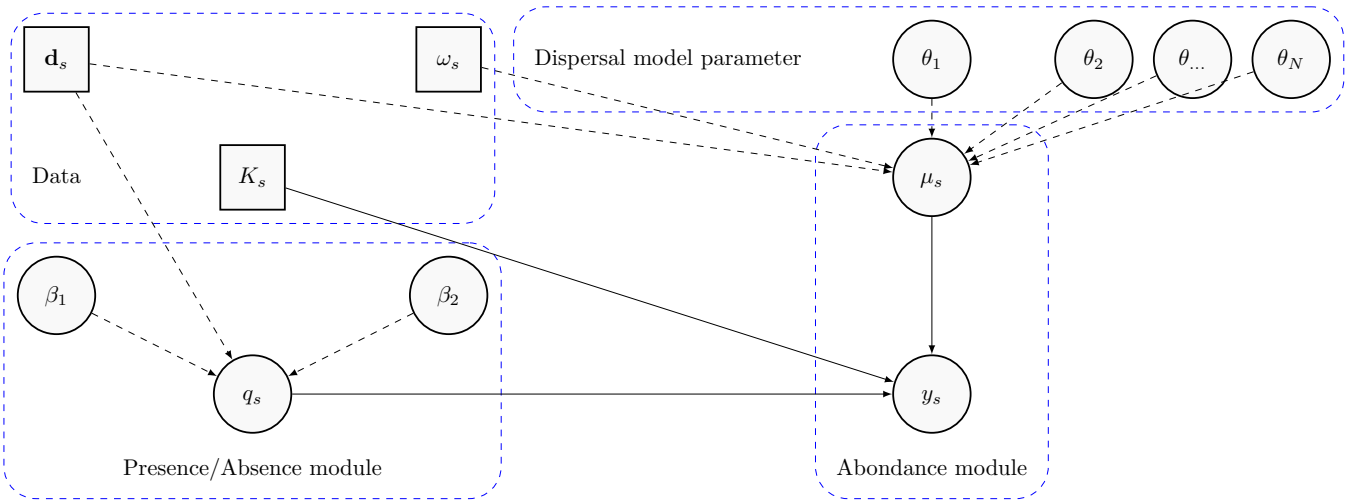


FIGURE 3.5: Graphe acyclique orienté du modèle ZIP à espérance fixe.

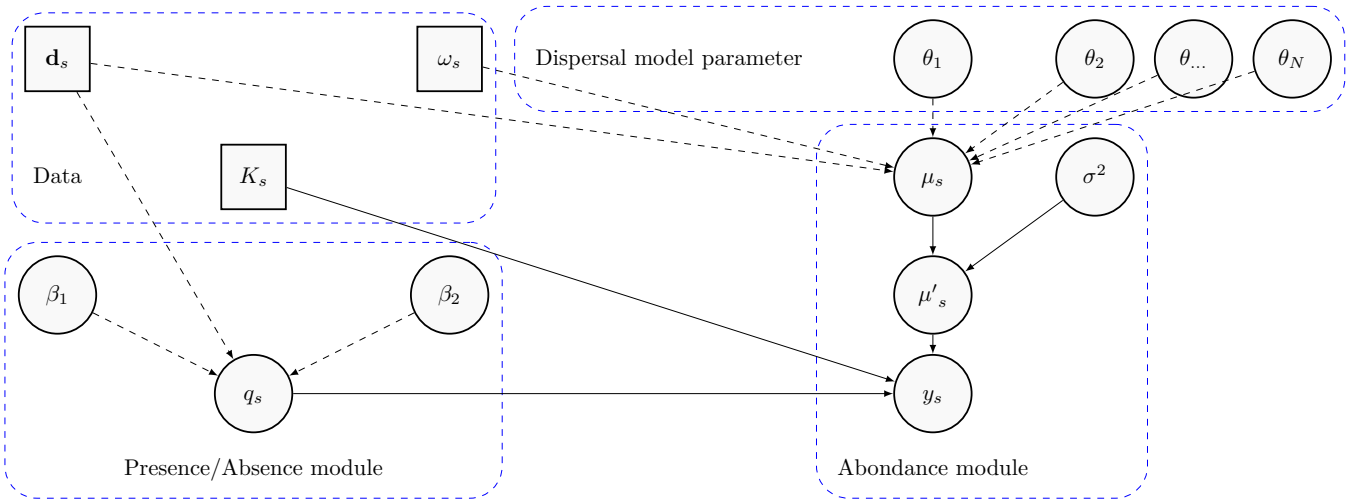


FIGURE 3.6: Graphe acyclique orienté du modèle ZIP à espérance normale.

6 Maillage adaptatif pour l'approche individuelle

6.1 Problématique

Dans le mode individuel de représentation de la dispersion (voir la section 1), la prédiction du taux de pollinisation croisée en chaque point récepteur nécessite le calcul du noyau de dispersion entre tous les points émetteurs de pollen (OGM ou non) du paysage considéré et ce point récepteur. De plus la prédiction à l'échelle d'une parcelle réceptrice nécessite de répéter ce calcul pour un nombre élevé de points à l'intérieur de cette parcelle.

Le temps nécessaire pour accomplir ces calculs s'avère limitant si on veut estimer les paramètres du modèle. En effet l'estimation requiert un très grand nombre d'appels au modèle avec différentes valeurs des paramètres. Chacun de ces appels inclut le calcul du noyau de dispersion individuelle pour toutes les paires émetteurs/récepteurs du paysage considéré. De plus dans un cadre bayésien, le nombre d'évaluations du modèle est considérablement augmenté par rapport aux méthodes classiques.

Le temps de calcul devient donc un problème rédhibitoire si on veut estimer les paramètres du noyau de dispersion individuelle par une approche bayésienne. Cette section présente une méthode originale d'approximation numérique du taux de pollinisation croisée μ_s, développée dans le cadre de la thèse pour réduire les temps de calcul du mode individuel de dispersion sans perdre trop en qualité de prédiction.

La méthode proposée s'appuie sur certaines caractéristiques communes aux fonctions de dispersion. On sait notamment que, du fait de la très rapide décroissance de ces fonctions à courte distance, ce sont les émetteurs les plus proches du récepteur qui contribuent significativement à sa pollinisation. La contribution des émetteurs varie plus lentement à moyenne distance et elle devient très faible voire négligeable à longue distance.

Nous présentons dans un premier temps la grille de référence , c'est-à-dire la grille de point émetteurs à prendre dans le calcul en absence d'approximation, et le principe de l'approximation recherchée. Nous définissons ensuite les facteurs de réglages de cette approximation et explicitons les hypothèses qui permettent d'établir ces facteurs. Puis, nous décrivons les protocoles établis pour *i)* générer des grilles dégradées par rapport à la grille de référence ; *ii)* réaliser les calculs avec de telles grilles. Enfin nous présentons l'approximation retenue par la suite pour réaliser les calculs.

6.2 Grille de référence et recherche d'approximations

Nous avons fait implicitement plusieurs hypothèses dans l'introduction de cette section, nous les explicitons ici :

1. La fonction de dispersion décroît très rapidement ; cela signifie que les émetteurs les plus proches du récepteur contribuent le plus à sa pollinisation. De ce fait, il est nécessaire de prendre en compte une fraction élevée d'émetteurs dans le voisinage proche du récepteur pour le calcul (reste à définir ce que les termes "proche" et "fraction élevée" signifient).

2. Pour les mêmes raisons, plus la distance à l'émetteur augmente, plus la contribution à la pollinisation diminue et le fait de façon lente. Cela nous amène à considérer qu'au delà de d'une certaine distance, la fraction d'émetteurs à prendre à compte dans le calcul peut être significativement diminuée.

3. A partir d'un certain point il n'est plus nécessaire de prendre en compte les émetteurs individuellement ; les surfaces peuvent alors êtres "résumées" par des points.

Cela nous amène à poser le problème mathématique d'approximation de la façon suivante : nous considérons comme méthode de référence, le calcul de la dispersion s'appuyant sur une discrétisation du paysage par une grille régulière $\mathcal{S}$ considérée comme suffisamment fine pour la précision recherchée en pratique. Les différentes approximations envisagées consistent à utiliser des maillages plus grossiers, dont le pas décroît avec la distance.

Plus précisément, on appelle $\mathcal{S}$ la grille rectangulaire de points s' dans $\mathbb{R}^2$ centrée sur s et de pas δ_x et δ_y. On note GM l'aire occupée par les parcelles de maïs cultivées en OGM et nonGM l'aire occupée par les parcelles de maïs conventionnelle (non cultivées en OGM). Le ratio que l'on cherche à approximer est :

$$\mu_s = \frac{\rho_1}{\rho_1 + \rho_2}, \tag{3.22}$$

où

$$\rho_1 = \sum_{s' \in S \cap \mathrm{GM}} \gamma(s, s') \quad \text{et} \quad \rho_2 = \sum_{s' \in S \cap \mathrm{nonGM}} \gamma(s, s').$$

Notons maintenant :

$\mathcal{S}_u$, pour $u = 1, \ldots, U$, une sous-grille de $\mathcal{S}$, composée de points t' et de pas $\sqrt{F_u}\delta_x$ et $\sqrt{F_u}\delta_y$,

$\mathcal{C}_u$ un anneau centré sur s de rayons (r_u, R_u). Notons que $\mathcal{C}_1$ est systématiquement le disque de rayon R_1 ($r_1 = 0$).

Nous recherchons une approximation de la forme suivante :

$$\mu_s^* = \frac{\sum_{u=1}^{U} \rho_{1,u}^* + \rho_3}{\sum_{u=1}^{U} \rho_{1,u}^* + \sum_{u=1}^{U} \rho_{2,u}^* + \rho_3}, \tag{3.23}$$

où

$$\rho_{1,u}^* = \sum_{t_u' \in \mathcal{C}_u \cap \mathcal{S}_u \cap \mathrm{GM}} \gamma(s, t_u') \times F_u,$$

$$\rho_{2,u}^* = \sum_{t_u' \in \mathcal{C}_u \cap \mathcal{S}_u \cap \mathrm{nonGM}} \gamma(s, t_u') \times F_u$$

et

$$\rho_3 = \gamma(s, z') \times \tau$$

avec

z' le point OGM le plus proche du récepteur, au delà du dernier anneau $\mathcal{C}_U$,

τ le nombre de points OGM de $\mathcal{S}$ au delà de $\mathcal{C}_U$.

À partir de cette définition générale de l'approximation recherchée, plusieurs paramètres de réglage sont à considérer pour obtenir une bonne approximation de μ_s, en particulier :

- Quel nombre de zones U ?
- Quels rayons (r_u, R_u) pour les anneaux $\mathcal{C}_u$?
- Quel pas relatif F_u pour la grille $\mathcal{S}_u$ par rapport aux pas de $\mathcal{S}$, dans l'anneau $\mathcal{C}_u$?

6.3 Plan d'expérience

Pour répondre aux questions que pose la recherche d'une approximation efficace sur les paramètres de réglage, nous avons choisi de fixer $U = 3$ et défini un plan d'expérience factoriel complet avec les facteurs et leurs modalités suivants :

- $\mathcal{C}_1$: disque de rayons R_1 centré sur le récepteur dans lequel on considère la sous-grille $\mathcal{S}_1$ de pas $\sqrt{F_1}\delta_x$ et $\sqrt{F_1}\delta_y$.
  $\mathcal{C}_1 \in \{(0,2)\ ;\ (0,3)\ ;\ (0,4)\ ;\ (0,5)\}$
- $\mathcal{C}_2$: anneau de rayons (r_2, R_2) centré sur le récepteur dans lequel on considère la sous-grille $\mathcal{S}_2$ de pas $\sqrt{F_2}\delta_x$ et $\sqrt{F_2}\delta_y$.
  $\mathcal{C}_2 \in \{(R_1, R_1+5)\ ;\ (R_1, R_1+10)\ ;\ (R_1, R_1+20)\ ;\ (R_1, R_1+30)\}$
- $\mathcal{C}_3$: anneau de rayons (r_3, R_3) centré sur le récepteur dans lequel on considère la sous-grille $\mathcal{S}_3$ de pas $\sqrt{F_3}\delta_x$ et $\sqrt{F_2}\delta_y$.
  $\mathcal{C}_3 \in \{(R_2, R_2+20)\ ;\ (R_2, R_2+30)\}$
- F_1 : pas relatif de la sous-grille $\mathcal{S}_1$.
  $F_1 \in \{2\ ;\ 4\ ;\ 6\}$
- F_2 : pas relatif de la sous-grille $\mathcal{S}_2$.
  $F_2 \in \{50\ ;\ 100\ ;\ 200\}$
- F_3 : pas relatif de la sous-grille $\mathcal{S}_3$.
  $F_3 \in \{200\ ;\ 500\ ;\ 1000\}$

Les six facteurs considérés ayant respectivement 4, 4, 2, 3, 3 et 3 modalités, on a au total $4^2 \times 3^3 \times 2 = 864$ grilles différentes. Chaque combinaison de modalités des facteurs représente une grille candidate pour l'approximation. Le nombre de facteurs ainsi que leurs modalités respectives ont été définis par expertise (Antoine Messéan, Frédérique Angevin, communication personnelle, 12 septembre 2012). Différents exemples de grilles candidates sont illustrées dans les parties qui suivent.

6.4 Protocole de simulation

Nous présentons à présent le protocole de simulation défini pour mener les calculs à partir des grilles candidates générées par le plan d'expérience. On dispose de 864 grilles

différentes ou approximations candidates. Chaque grille est composée de 4 zones, dans lesquelles les procédures de calcul du taux de pollinisation croisée sont légérement différentes :

- Zone A : Cette zone correspond au disque $\mathcal{C}_1$ de rayons R_1 autour du récepteur, dans lequel on considère la sous-grille $\mathcal{S}_1$ de $\mathcal{S}$, composée par les points $t_1^{'}$ et de pas $\sqrt{F_1}\delta_x$ et $\sqrt{F_1}\delta_y$. Pour chacun des émetteurs sélectionnés dans cette zone, on applique la fonction de dispersion individuelle et on pondère le résultat par la fraction de points conservés dans la zone (par rapport à la grille de référence). On obtient donc $\gamma(s, t_1^{'}) \times F_1$ pour chaque émetteur. On note $\rho^*_{1,1}$ (pour rester cohérent avec les notations de l'équation 3.23) la somme des contributions de tous les points émetteurs OGM du disque $\mathcal{C}_1$, et $\rho^*_{2,1}$ la somme des contributions de tous les points émetteurs non OGM de ce disque. Ces deux grandeurs correspondent aux approximations des sommes des contributions à la pollinisation du récepteur s, des émetteurs OGM et non OGM de la zone A :

$$\rho^*_{1,1} = \sum_{t_1^{'} \in \mathcal{C}_1 \cap \mathcal{S}_1 \cap \mathrm{GM}} \gamma(s, t_1^{'}) \times F_1,$$

$$\rho^*_{2,1} = \sum_{t_1^{'} \in \mathcal{C}_1 \cap \mathcal{S}_1 \cap \mathrm{nonGM}} \gamma(s, t_1^{'}) \times F_1.$$

- Zone B1 : Cette zone correspond à l'anneau $\mathcal{C}_2$ de rayon (r_2, R_2) autour du récepteur dans lequel on considère la sous-grille $\mathcal{S}_2$ de $\mathcal{S}$, composée par les points $t_2^{'}$ et de pas $\sqrt{F_2}\delta_x$ et $\sqrt{F_2}\delta_y$ Dans cette zone, la démarche est identique à celle de la zone A. On obtient $\gamma(s, t_2^{'}) \times F_2$ pour chaque émetteur. On note $\rho^*_{1,2}$ la somme des contributions de tous les points émetteurs OGM de l'anneau $\mathcal{C}_2$, et $\rho^*_{2,2}$ la somme des contributions de tous les points émetteurs non OGM de cet anneau :

$$\rho^*_{1,2} = \sum_{t_2^{'} \in \mathcal{C}_2 \cap \mathcal{S}_2 \cap \mathrm{GM}} \gamma(s, t_2^{'}) \times F_2,$$

$$\rho^*_{2,2} = \sum_{t_2^{'} \in \mathcal{C}_2 \cap \mathcal{S}_2 \cap \mathrm{nonGM}} \gamma(s, t_2^{'}) \times F_2.$$

- Zone B2 : Cette zone correspond à l'anneau $\mathcal{C}_3$ de rayon (r_3, R_3) autour du récepteur dans lequel on considère la sous-grille $\mathcal{S}_3$ de $\mathcal{S}$, composée par les points $t_3^{'}$ et de pas $\sqrt{F_3}\delta_x$ et $\sqrt{F_3}\delta_y$ Dans cette zone, la démarche est identique à celle de la zone A et B1. On obtient $\gamma(s, t_3^{'}) \times F_3$ ppur chaque émetteur. On note $\rho^*_{1,3}$ la somme des contributions de tous les points émetteurs OGM de l'anneau $\mathcal{C}_3$, et $\rho^*_{2,3}$ la somme des contributions de tous les points émetteurs non OGM de cet anneau :

$$\rho^*_{1,3} = \sum_{t_3^{'} \in \mathcal{C}_3 \cap \mathcal{S}_3 \cap \mathrm{GM}} \gamma(s, t_3^{'}) \times F_3,$$

$$\rho^*_{2,3} = \sum_{t_3^{'} \in \mathcal{C}_3 \cap \mathcal{S}_3 \cap \mathrm{nonGM}} \gamma(s, t_3^{'}) \times F_3.$$

- Zone C : Cette zone correspond à l'ensemble du paysage considéré au delà de l'anneau $\mathcal{C}_3$. Dans cette zone, on applique la fonction de dispersion individuelle au point émetteur OGM z' le plus proche du récepteur pondéré par le nombre τ de points émetteurs OGM au delà de l'anneau $\mathcal{C}_3$. On obtient $\gamma(s, z^{'}) \times \tau$. On note ρ_3 cette quantité.

Le taux de pollinisation pour un récepteur situé en s est donné par :

$$\mu_s^* = \frac{\rho_{1,1}^* + \rho_{1,2}^* + \rho_{1,3}^* + \rho_3}{\rho_{1,1}^* + \rho_{1,2}^* + \rho_{1,3}^* + \rho_{2,1}^* + \rho_{2,2}^* + \rho_{2,3}^* + \rho_3}. \tag{3.24}$$

6.5 Mise en œuvre et exemples de grilles dégradées

Les données utilisées pour la calibration de l'approximation sont celles décrites et publiées dans Klein et al. (2003). Elles sont issues d'une expérience en condition contrôlée de type *mono-source*. La grille de référence correspond à une grille où chaque plante est un point. Il y a une parcelle de 120 m x 120 m avec :

- en ordonnées : 1 rang tous les 0.8 m ;
- en abcisses : 1 plante tous les 0.15 m.

Une parcelle centrale de 20 m x 20 m est semée avec du maïs bleu (considéré comme la source OGM). Il y a au total 122553 points dans la grille de référence dont 3225 considérés comme OGM (les points de la parcelle centrale). La grille de référence est illustrée en figure 3.7.

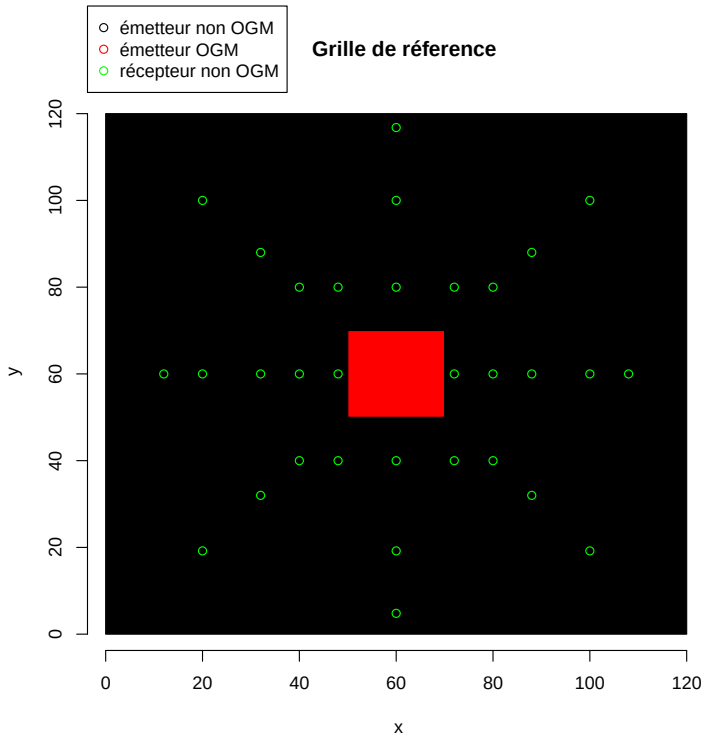


FIGURE 3.7: Grille de référence.

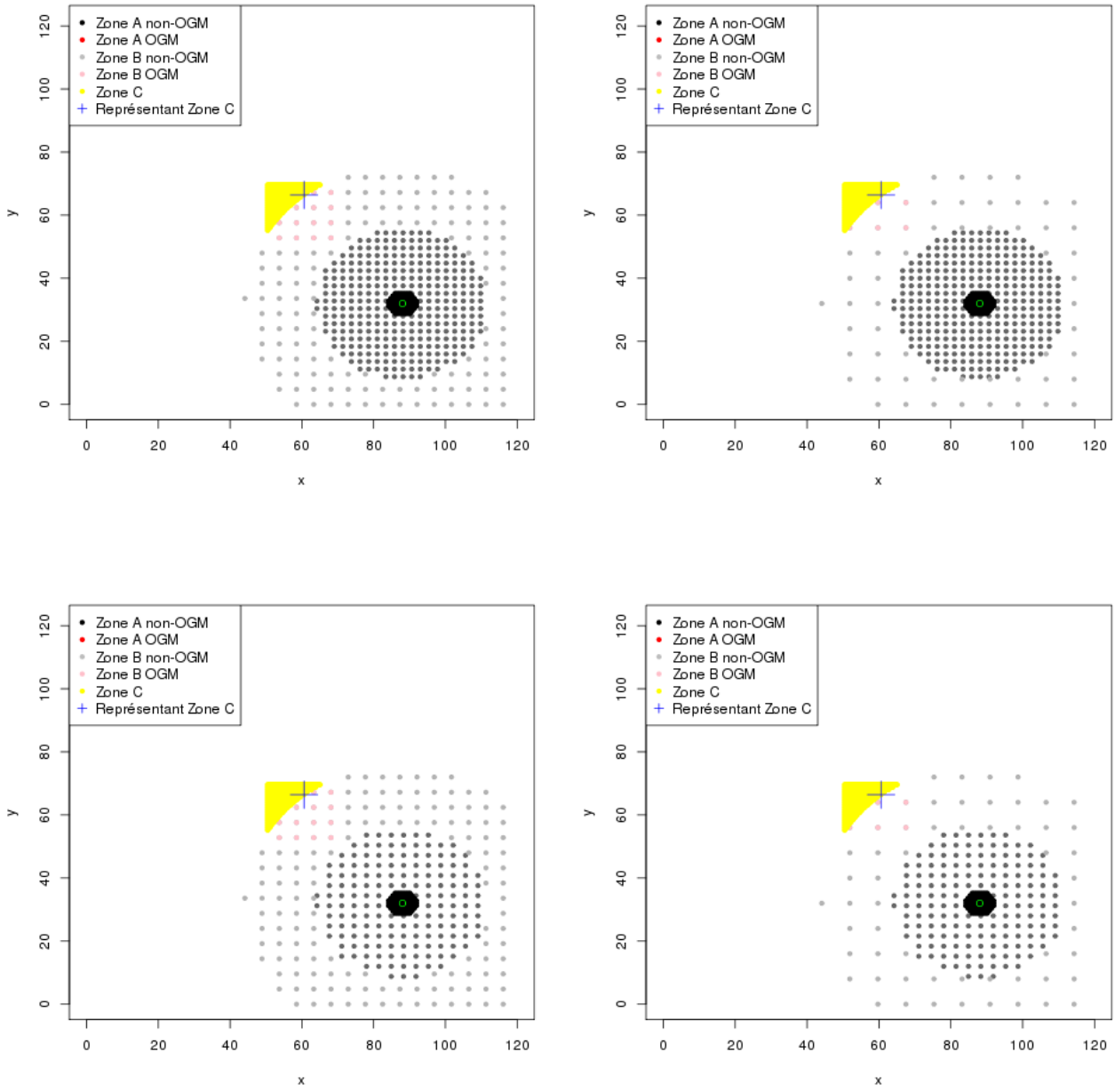


FIGURE 3.8: Exemples de grilles dégradées pour un récepteur avec $\mathcal{C}_1 = (0, 4)$, $\mathcal{C}_2 = (4, 20)$, $\mathcal{C}_3 = (20, 24m)$, $F_1 = 4$, et avec de gauche à droite $F_3 = 200,\ 500$ et de haut en bas $F_2 = 50,\ 100$.

Afin d'analyser le comportement des différentes approximations de la grille de référence et d'en évaluer les facteurs de réglage, on considère un certain nombre de points récepteurs dans cette grille sur lesquels seront menés les calculs. Ces récepteurs sont choisis pour remplir suffisament l'espace sans que leur nombre soit limitant pour les temps de simulation. On superpose ensuite à la grille de référence des émetteurs la grille de récepteurs. Nous avons défini 32 points récepteurs au total. Ces points, et leur répartition dans la parcelle sont visibles sur la figure 3.7.

On réalise dans un premier temps le calcul de μ_s par la formule exacte de l'équation (3.22) pour les 32 points récepteurs s en considérant la totalité des émetteurs s' de la grille de référence. On obtient un vecteur de référence Ω de 32 valeurs auquel nous pourrons comparer les prédictions qui s'appuient sur une grille d'émetteurs dégradée par rapport à la grille de référence.

On réalise ensuite le calcul de μ_s^* par la formule approchée de l'équation (3.24) pour les 32 points récepteurs s et pour chaque grille candidate générée par le plan d'expérience défini plus haut (864 au total). On obtient un vecteur Ω_i^* ($i \in \{1, 2, \ldots, 864\}$) de 32 valeurs pour chacune des grilles candidates.

Pour déterminer un critère permettant de classer les qualités de prédiction obtenues avec chaque grille candidate, on calcule les vecteurs Δ_i des valeurs absolues des différences entre le vecteur de référence et les vecteurs de prédictions issues du calcul approché par les grilles candidates $\Delta_i = |\Omega - \Omega_i^*|$ ($i \in \{1, 2, \ldots, 864\}$). On retient ensuite la valeur maximum de chacun de ces vecteurs pour le reste de l'analyse et la recherche de la meilleure approximation.

6.6 Résultats

Dans l'objectif d'obtenir une approximation dans le cas du mode individuel de représentation de la dispersion, il y a, pour chaque grille candidate (864 au total), deux quantités qui nous intéressent :

- la qualité de prédiction obtenue par rapport à celle de la grille de référence ;
- le temps de calcul associé à son utilisation (directement proportionel au nombre de points conservés dans la grille par rapport à la référence).

On se trouve donc dans un cas où il y a deux critères à optimiser conjointement et plus précisément dans une situation de recherche d'optimum de Pareto. Pour un tel optimum, l'amélioration d'une des deux quantités ne peut se faire que par la dégradation de l'autre. À l'origine, cette théorie provient du domaine de l'économie et s'exprime comme *un état de la société dans lequel on ne peut pas améliorer le bien-être d'un individu sans détériorer celui d'un autre* (Roger and Eeckhoudt, 1998). Cependant, elle s'applique aussi très bien à notre situation. Ici les "individus" sont *i)* la qualité de prédiction ; *ii)* le temps de calcul. Et l'amélioration de la qualité de prédiction obtenue par une grille dégradée optimale au sens de Pareto implique forcément la prise en compte de plus de points et donc l'augmentation du temps de calcul (l'augmentation des temps de calcul étant, bien entendu, considérée comme une détérioration).

À partir de ce constat et plutôt que de chercher ou d'élaborer un système d'optimisation multicritères plus ou moins sophisitiqué, nous nous sommes contentés de représenter le front de Pareto défini comme l'ensemble des compromis possibles entre les deux critères de cette étude. La figure 3.9 offre une représentation de ce front de Pareto. Elle contient un point pour chacune des 864 grilles testées, avec en abscisse le nombre de points de la grille et en ordonnée sa qualité de prédiction sur le jeu de données. Sur ce graphique, les

points qui nous intéressent et qui correspondent au front de Pareto sont les points les plus proches des deux axes. Le point idéal serait à l'origine (i.e. une qualité parfaite pour un effort de simulation nul), cependant la situation nous oblige à faire un compromis entre le nombre de points dans la grille et la qualité de prédiction qui résulte de son utilisation. On cherche donc les points pour lesquel la qualité de prédiction reste acceptable pour un nombre de points dont le temps de calcul est raisonnable.

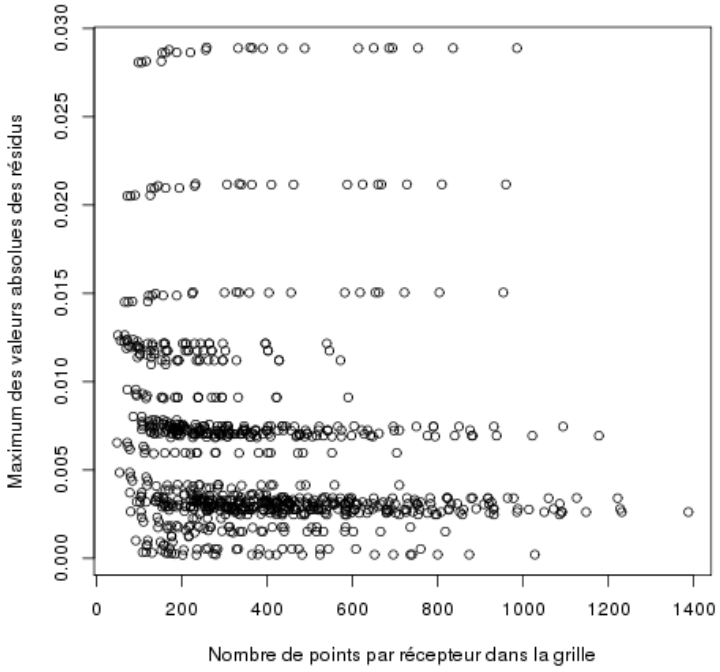


FIGURE 3.9: Front de Pareto - Qualité de prédiction en fonctions du coût de la simulation (nombre de points $\propto$ temps de calculs)

L'examen de la figure 3.9 et des grilles proches du front de Pareto montre qu'une grille de seulement 118 points permet, sur les données utilisées, d'obtenir une qualité de prédiction similaire à celle obtenue par une grille de plus de 1000 points. C'est cette grille (dont l'illustration est disponible à la figure 3.10) qu'on utilisera pour réaliser les calculs dans toute la suite de ce travail.

6.7 Application

Nous présentons désormais deux figures pour illustrer le fait que la grille que nous venons de sélectionner peut s'adapter à différents cas en fonction de la position du récepteur par rapport à la surface cultivée en OGM. En effet l'intérêt de la grille adaptative n'est pas uniquement de réduire le nombre de points pour le calcul de la pollinisation d'un récepteur mais aussi de réduire le nombre de fois où le calcul de dispersion est réalisé ; ainsi, à partir de la grille sélectionnée et quel que soit la taille de la surface sur laquelle on veut prédire,

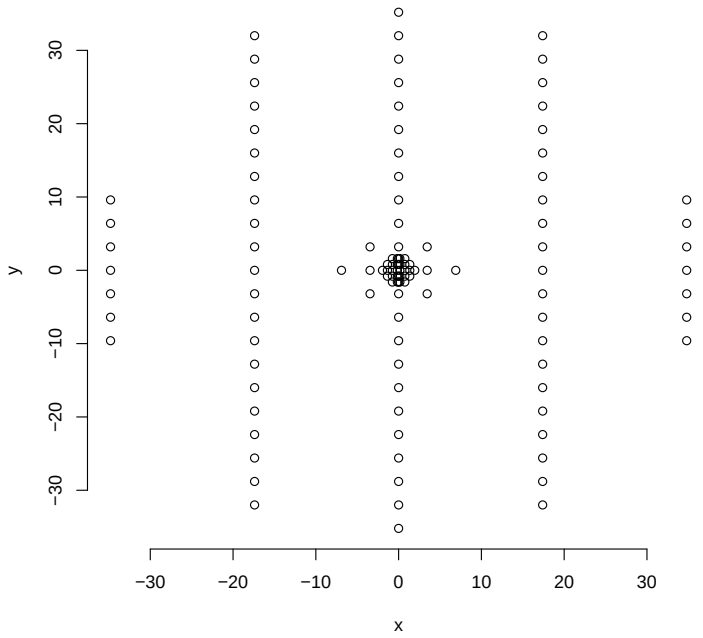


FIGURE 3.10: Grille minimisant le maximum de la valeur absolue des résidus $\mathcal{C}_1 = (0, 2)$, $\mathcal{C}_2 = (2, 7)$, $\mathcal{C}_3 = (7, 37)$, $F_1 = 2$, $F_2 = 100$, $F_3 = 1000$.

on ne fait le calcul de dispersion à proprement parler que 118 fois en tout, et pour tous les points récepteurs (on ne répète pas 118 calculs pour chaque récepteur). Le calcul effectif de pollinisation croisée en un récepteur donné se fait simplement en adaptant la grille au récepteur d'intérêt, c'est-à-dire en identifiant les points de la grille sélectionnée qui, pour le récepteur d'intérêt, correspondent à des émetteurs OGM ou non.

Les figures 3.11 et 3.12 présentent deux cas. Le cas d'un récepteur dont l'anneau $\mathcal{C}_3$ recouvre toute la surface cultivée en OGM et le cas d'un récepteur dont l'anneau $\mathcal{C}_3$ ne recouvre pas toute la surface cultivée en OGM.

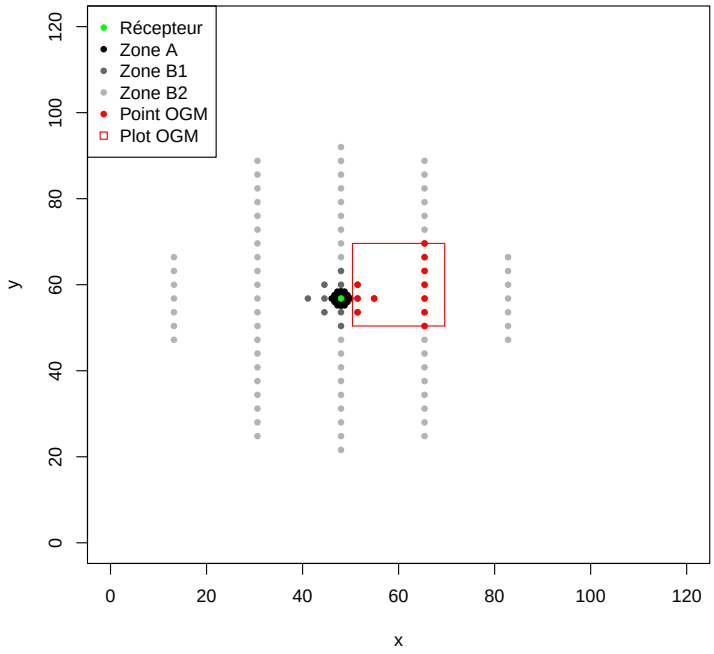


FIGURE 3.11: Illustration de l'adaptation de la grille à un récepteur dont les zones B1 et B2 recouvrent toute la surface cultivée en OGM.

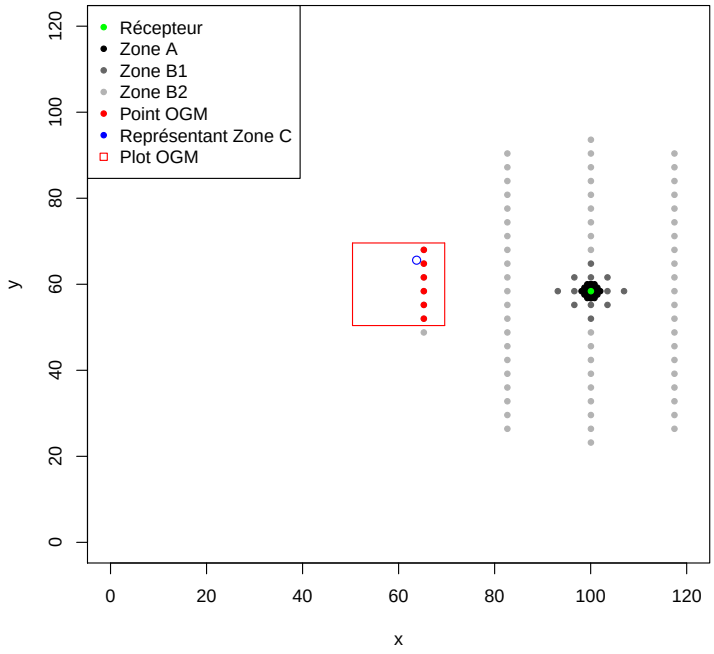


FIGURE 3.12: Illustration de l'adaptation de la grille à un récepteur dont les zones B2 et C recouvrent toute la surface cultivée en OGM.

L'approximation définie dans cette section et la réduction drastique du nombre de points émetteurs à prendre en compte dans le calcul nous permettent d'appliquer la démarche d'estimation Bayésienne sur tout un ensemble de modèles de dispersion basés sur le mode individuel (voir chapitre 4). Nous rappelons qu'il s'agit d'une approche pragmatique que nous avons développée pour des raisons pratiques de faisabilité technique. Cependant, elle nous semble tout à fait applicable à d'autre domaines pour lesquels des calculs de dispersion entre surfaces émettrices de particules interviennent. L'approximation développée ici se prête à plusieurs améliorations, en particulier la définition d'une procédure pour adapter la grille aux valeurs des paramètres du modèle de dispersion plutôt que de choisir une grille une fois pour toutes.

Chapitre 4

Estimation et Mise en œuvre

Table des matières

1 Contexte mono-source 81

1.1 Données 81

1.1.1 Jeu d'entrainement - Montargis 98 et Montargis 99 . 81

1.1.2 Jeu de validation - Mas Cebria 82

1.2 Estimation 82

1.2.1 Distribution a priori 84

1.2.2 Convergence 84

1.3 Prédiction 92

2 Contexte multi-sources 98

2.1 Données 99

2.2 Intégration du décalage de floraison 102

2.3 Stratégies pour l'estimation 103

2.4 Prédiction avec modèle ajusté en *mono-source* 106

3 Évaluation et Sélection de modèles 112

3.1 Critères statistiques 112

3.1.1 Critères classiques 112

3.1.2 Critères de scoring 113

3.1.3 Critères décisionnel 113

3.2 Analyse en composantes principales 114

3.2.1 Méthode 115

3.2.2 Données 115

3.2.3 Résultats 115

3.2.4 Discussion 116

Ce chapitre est consacré à la mise en œuvre des méthodes décrites précédemment et à l'estimation des paramètres des modèles de prédiction du taux de pollinisation croisée définis au chapitre 3, section 5. Nous présentons dans un premier temps la démarche employée pour estimer les paramètres et prédire le taux de pollinisation croisée dans des contextes mono-source, c'est-à-dire comprenant une seule parcelle source de pollen OGM. La section 1 reprend en partie un article soumis à *Environmental Modelling & Software* et qui figure en annexe. Cet article a été rédigé dans une optique très générique permettant l'application des méthodes à un très large spectre d'applications. De ce fait, la construction de la section 1 n'est pas la même que celle de l'article ; nous y décrivons tout d'abord les données utilisées et la structure factorielle des modèles testés (les facteurs et leurs modalités étant décrits au chapitre 3). Puis nous décrivons la mise en œuvre de la procédure d'estimation et discutons la qualité des prédictions. Dans un second temps (section 2), nous synthétisons ce qui a été fait et ce que nous proposons dans les contextes multi-source. Dans la section 3, nous revenons sur la stratégie que nous avons développée et les méthodes employées pour répondre à une des questions essentielles en statistique et en modélisation, et centrale dans ce chapitre, à savoir l'évaluation et la sélection de modèles.

1 Contexte mono-source

1.1 Données

Pour la calibration des modèles proposés dans le chapitre 3, nous avons sélectionné trois jeux de données issus d'expériences en conditions contrôlées de type *mono-source*, dont les mesures sont des comptages de grains et sont donc discrètes. Nous présentons tout d'abord ces jeux de données.

1.1.1 Jeu d'entrainement - Montargis 98 et Montargis 99

Ces deux expérimentations ont été réalisées durant les étés 1998 et 1999 près de Montargis (France, département du Loiret). En 1998, un champ de maïs de 120×120 m a été semé : 160 lignes espacées de 0.8 m avec 800 plantes par ligne espacées de 0.15 m. Une parcelle centrale de 20×20 m a été semée avec des plantes produisant des graines de couleur bleue, le reste du champ ne contenant que du maïs produisant des graines de couleur jaune de la variété hybride Adonis. Le maïs bleu était d'une variété proche d'Adonis et homozygote pour l'allèle "grain bleu".

La dispersion du pollen a commencé le 18 juillet. Les deux types de plantes (à grains bleus ou jaunes) ont fleuri de manière presque synchrone : le maïs bleu a commencé sa floraison le 19 juillet (floraison mâle) et le 20 juillet (floraison femelle) tandis que le maïs jaune a commencé le 18 juillet (floraison mâle) et 19 juillet (floraison femelle). La dispersion a duré 14 jours et s'est terminée le 1er août. Les épis ont été récoltés et analysés le 16 octobre. Un total de K=2937 épis ont été échantillonnés selon une grille rectangulaire : 101 lignes ont été sélectionnées (chacune des 36 lignes de part et d'autre de la parcelle centrale et une ligne sur trois ailleurs) et 31 épis sur chaque ligne sélectionnée

(un épi tous les 4 m). Soixante-quatre épis n'ont pas pu être échantillonnés dans le coin ouest du champ. La figure 4.1 illustre ces expérimentations et le plan d'échantillonnage réalisé.

1.1.2 Jeu de validation - Mas Cebria

Une expérience similaire aux deux décrites précédemment à été conduite sur la plaine de Foixà près de Girone (Espagne, Catalogne). Ces données sont décrites et utilisées par Palaudelmàs et al. (2012) Quatre variétés de maïs Bt différentes, à grains jaunes et issues de l'événement MON810 (Monsanto Co), ont été semées dans le centre du champ, en formant des rectangles de surfaces croissantes le long d'un axe suivant la direction du vent dominant. Ces zones ont été reçu les quatre variétés suivantes : Aristis Bt (Nickerson-SENASA) sur environ 0.25 hectares (ha) ; DKC6575 (Monsanto Co) sur environ 0.75 ha ; PR33P67 et PR32P76 (Pioneer Hi Bred) sur environ 1.25 et 1.75 ha, respectivement. Un hybride non transgénique et produisant des grains blancs (hybride PR32Y52 de Pioneer Hi Bred) a été semé tout autour de la parcelle centrale pour remplir une superficie totale d'environ 27 ha (Fig. 1). L'essai était situé à quelques kilomètres de la mer, sur la plaine à l'embouchure de la rivière Ter. Toutes les variétés génétiquement modifiées ont été semées le 23 avril, tandis que les variétés conventionnelles environnantes étaient semées du 24 au 26 avril. Les champs ont été travaillés suivant les pratiques agricoles normales dans la zone. Durant les premiers stades de développenent, les repousses occasionnelles dans l'ensemble du domaine ont été arrachées à la main.

Trois méthodes d'échantillonnage ont été utilisées dans le champ récepteur (voir figure 4.2) : *(i)* une méthode stratifiée selon une grille de 10×10 m dans la direction principale du vent, dans la partie nord-ouest du champ conventionnel (champ A1). Les échantillons (trois épis chacun) ont été recueillis tous les 10 m dans les deux orientations, intra-ligne et inter-lignes, et en outre, dans le champ récepteur à 0, 2, 5 et 10 m de la bordure la plus proche du champ OGM dans la direction sous le vent dominant ; *(ii)* une méthode stratifiée, similaire à celle décrite précédemment, mais utilisant une grille de 20×20 m (champs A2, B, C et D, c'est à dire tout le reste du maïs conventionnel qui n'est pas dans la direction principale du vent).

1.2 Estimation

Dans le chapitre 3 et dans le cas de données discrètes, nous avons présenté et formalisé :

- 2 modèles d'observations (Poisson et ZIP) ;
- 2 options pour l'espérance (fixe et normale) ;
- 3 noyaux de dispersion (Exponentielle, 2Dt, NIG) ;
- 2 modes de représentation de la dispersion (global et individuel).

Tous ces aspects et variantes dans la modélisation de la dispersion ont été trouvés dans la littérature, cependant aucune étude suffisamment exhaustive ne permet de quantifier les bénéfices relatifs de l'une ou l'autre de ces variantes. C'est pourquoi, sans réelles informations a priori sur les performances respectives des différentes combinaisons des quatre facteurs définis ci-dessus (modèle d'observations, modèle d'espérance, noyau de dispersion

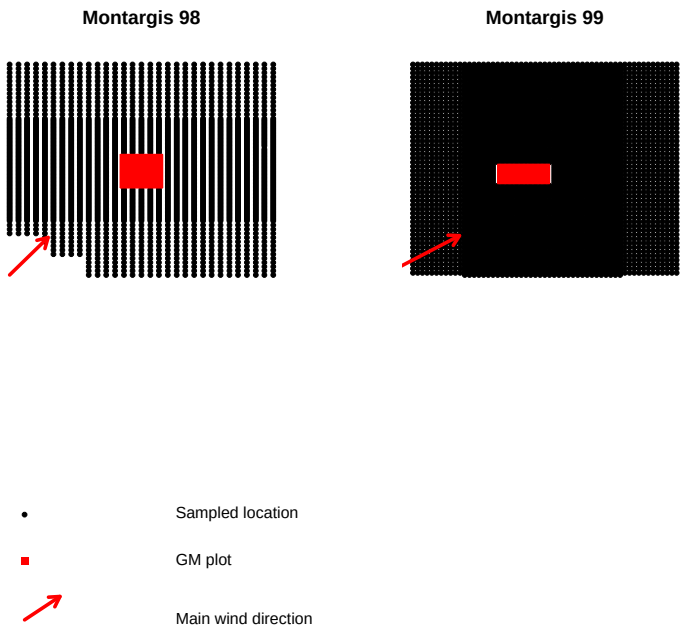


FIGURE 4.1: Plans d'échantillonnage utilisés pour les 2 expérimentations *mono-source* du jeu d'entrainement. Les zones OGM sont représentées en rouge.

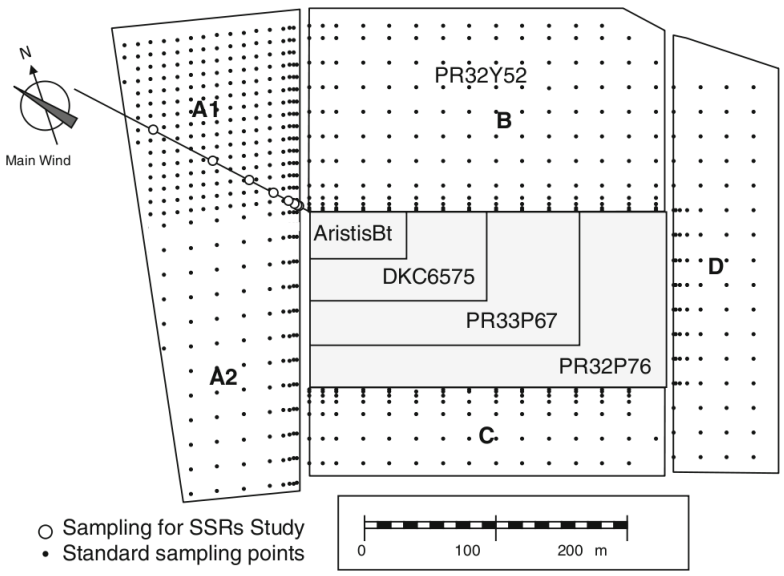


FIGURE 4.2: Plans d'échantillonnage utilisés pour l'expérimentation *mono-source* du jeu de validation. (Source : Palaudelmàs et al. (2012))

et mode de représentation de la dispersion), nous avons fait le choix de les comparer à l'aide d'un plan factoriel complet et donc de tester toutes les combinaisons de modalités des quatre facteurs. Au total, il y a $2 \times 2 \times 3 \times 2 = 24$ modèles. Le plan d'expérience complet figure dans le tableau 4.1. Pour chaque combinaison du plan factoriel, nous avons mis en œuvre l'estimation bayésienne et le calcul des critères décrits dans la section suivante.

Pour l'estimation des paramètres et la validation des modèles, nous avons suivi trois stratégies en parallèle :

1. Les paramètres sont estimés sur le premier jeu de données (Montargis 98). Les deux autres jeux (Montargis 99 et Mas Cebria) sont utilisés pour l'évaluation de la qualité prédictive des modèles estimés.

2. Les paramètres sont estimés sur le premier jeu de données (Montargis 99). Les deux autres jeux (Montargis 98 et Mas Cebria) sont utilisées pour l'évaluation de la qualité prédictive des modèles estimés.

3. Les paramètres sont estimés sur les deux premiers jeux de données (Montargis 98 et Montargis 99). Le jeu restant (Mas Cebria) est utilisé pour l'évaluation de la qualité prédictive des modèles estimés.

1.2.1 Distribution a priori

Comme nous l'avons vu au chapitre 2, section 3, dans le cadre statistique bayésien des distributions a priori doivent être attribuées à chaque paramètre pour réaliser l'estimation de la distribution a posteriori. Les distributions a priori des paramètres des modèles d'observation, à savoir β_1, β_2 et σ^2 sont listées dans le Tableau 4.2. Aucune information n'était disponible pour les paramètres de dispersion à l'exception de valeurs nominales et des bornes sur leurs valeurs potentielles, nous avons donc choisi des distributions a priori uniformes que l'on peut qualifier de non-informatives dans la mesure où aucune valeur, à l'intérieur de l'intervalle défini, n'est privilégiée. Notons par exemple que a_2 dans la fonction *Compound Exponential* est contraint à être inférieur à a_1 pour refléter la rapide décroissance du taux de pollinisation croisée dans les premiers mètres et une diminution plus lente après, comme c'est souvent observé dans les jeux de données Damgaard and Kjellson (2005). De la même façon, le paramètre K_e dans la fonction *Compound Exponential*, qui peut s'interpréter comme le taux de pollinisation croisée à une distance 0, a été contraint à se situer entre 0 et 1, par le simple fait que c'est un taux. Les distributions a priori pour tous les paramètres des noyaux de dispersion sont résumées dans le Tableau 4.3.

1.2.2 Convergence

En estimation bayésienne, et comme on l'a vu au chapitre 2 section 3, il est important de s'assurer de la convergence de l'algorithme d'estimation avant de réaliser l'inférence sur les paramètres. Ici, nous n'avons pas eu de problèmes particuliers de convergence des algorithmes MCMC ni de trop fortes autocorrélations entre les valeurs de paramètres

retenues dans la chaîne. La statistique de Gelman-Rubin (Gelman and Rubin, 1992) a été calculée pour les 24 modèles et dans les trois situations d'estimation et est toujours inférieure à la valeur 1, 1, ce qui indique qu'on peut considérer que les chaînes ont convergé. On considère donc à partir d'ici que l'estimation s'est déroulée convenablement et qu'on peut utiliser nos chaînes de Markov pour faire de l'inférence. Les figures 4.3, 4.4, 4.5 et 4.6 présentent des chaînes pour quelques-uns des 24 modèles et permettent d'illustrer leur convergence. Pour ne pas alourdir inutilement ce mémoire nous nous concentrons ici sur les paramètres de la fonction de dispersion CEx et considérons les deux modèles d'observations (Poisson et ZIP) et les deux modes de représentation de la dispersion entre parcelles (global et individuel).

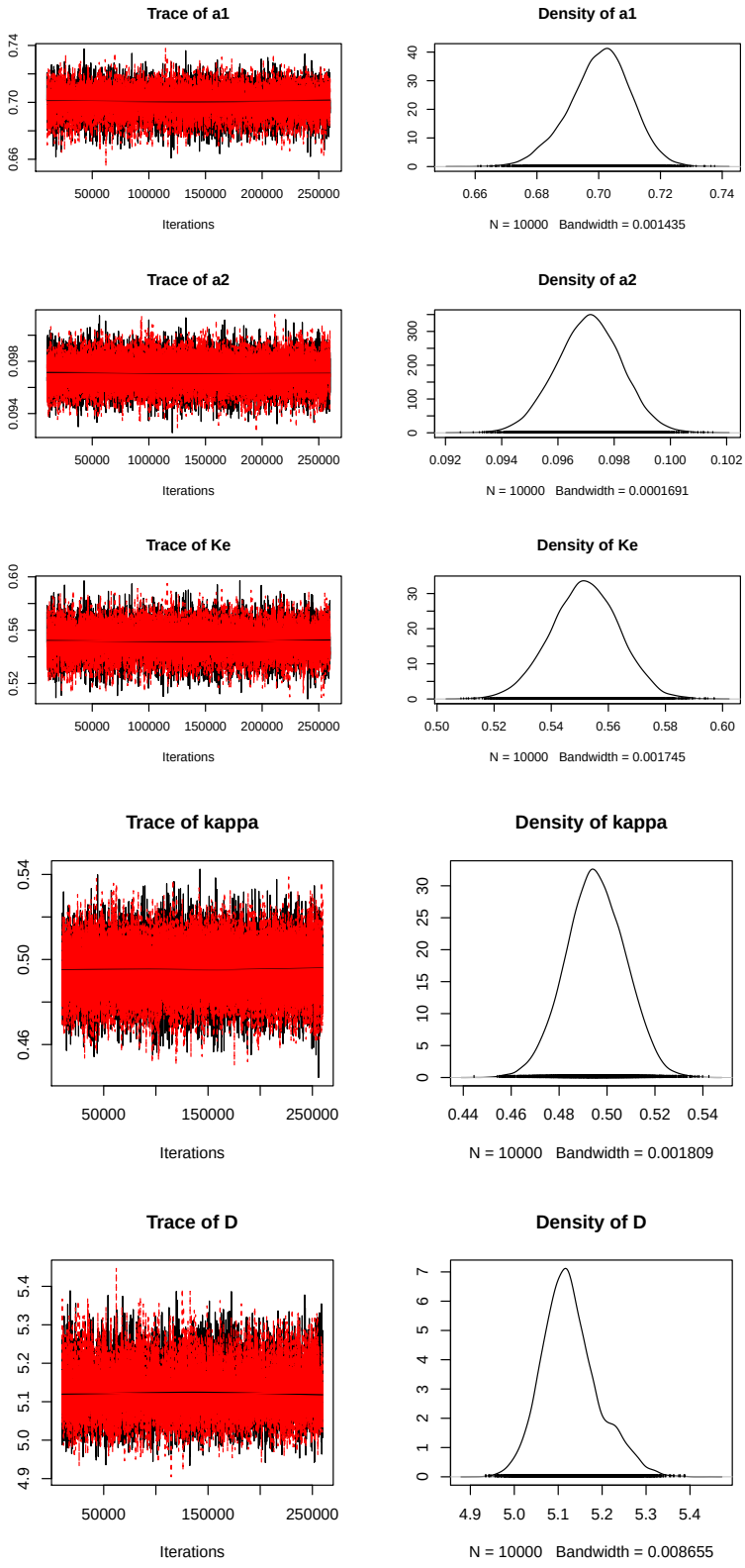


FIGURE 4.3: Traces des chaînes de Markov et distributions a posteriori des paramètres du modèle $GPExpoA$, estimés sur les données *mono-source*.

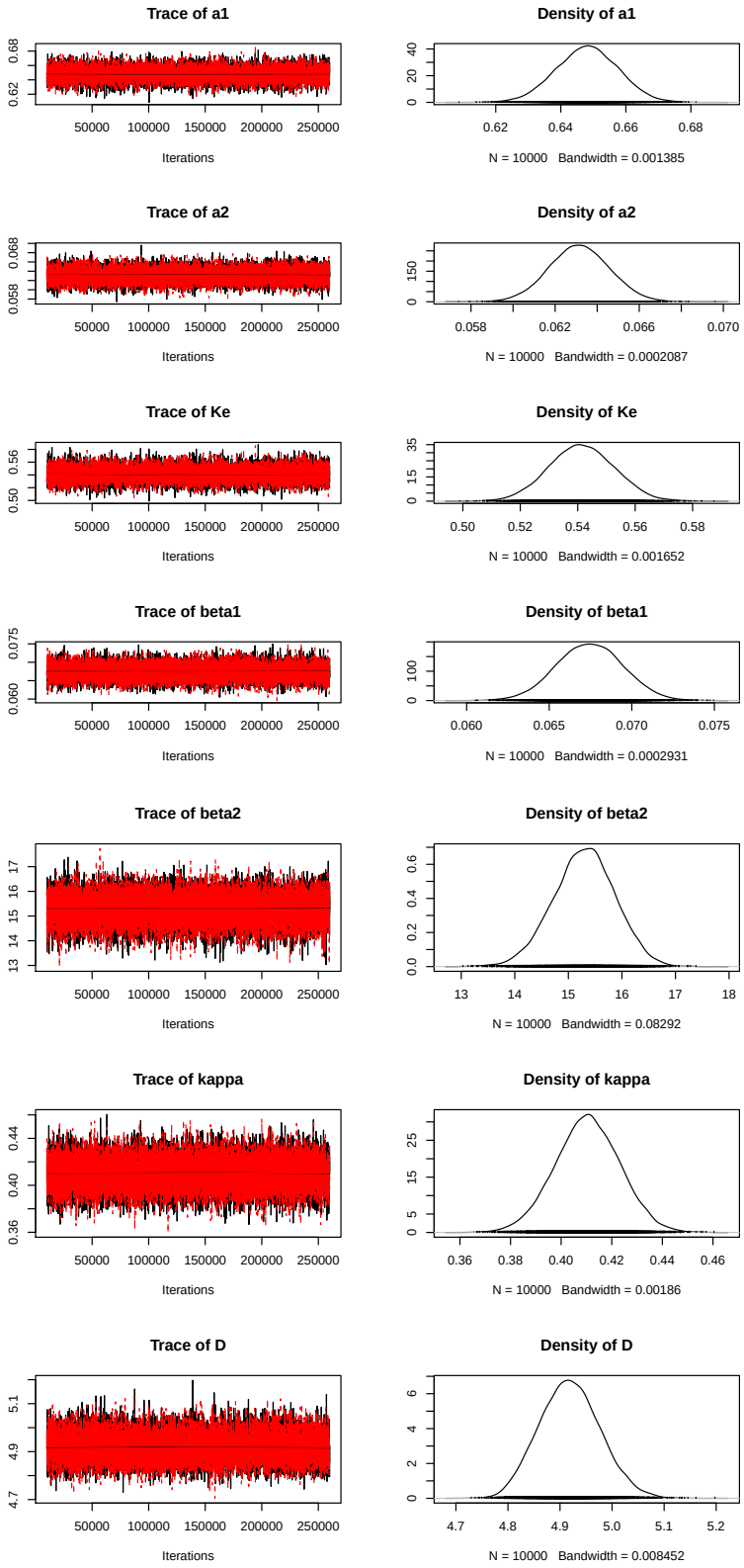


FIGURE 4.4: Traces des chaînes de Markov et distributions a posteriori des paramètres du modèle $GZExpoA$, estimés sur les données *mono-source*.

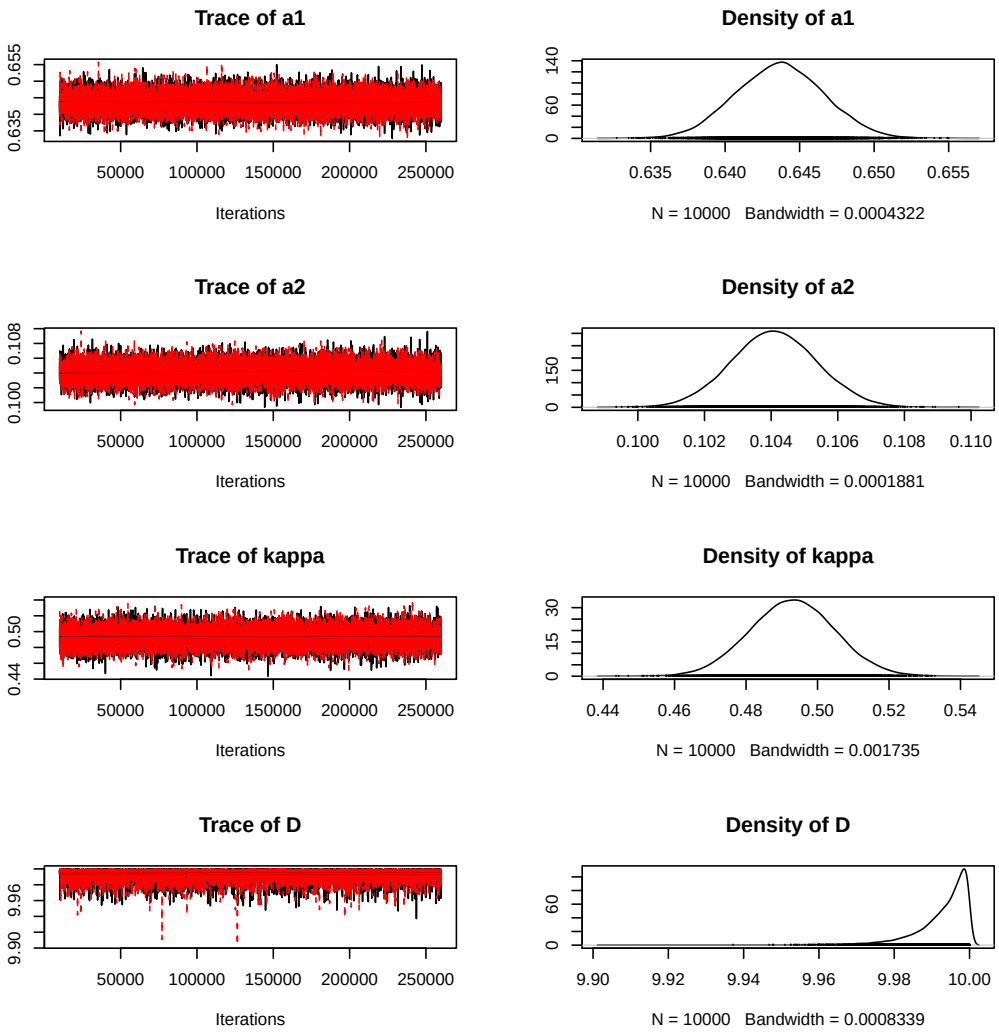


FIGURE 4.5: Traces des chaînes de Markov et distributions a posteriori des paramètres du modèle $IPExpoA$, estimés sur les données *mono-source*.

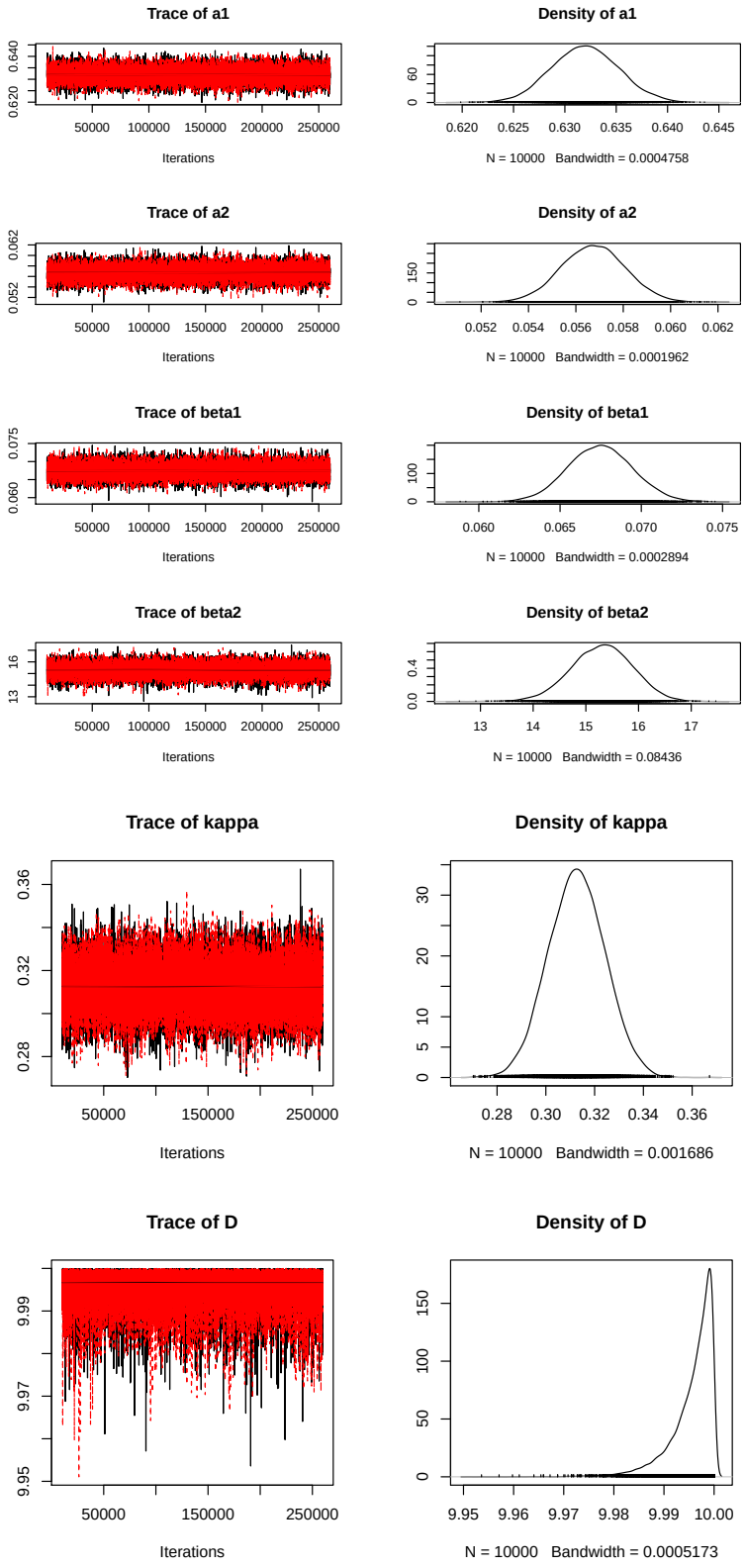


FIGURE 4.6: Traces des chaînes de Markov et distributions a posteriori des paramètres du modèle $IZExpoA$, estimés sur les données *mono-source*.

Mode	Distribution Y	Noyau γ	Distribution μ'	Code
Global	$\mathcal{P}(\mu')$	Exponential	μ	GPExpoA
Global	$\mathcal{P}(\mu')$	Exponential	$\mathcal{N}(\mu, \sigma^2)$	GPExpoB
Global	$\mathcal{P}(\mu')$	2Dt	μ	GP2DtA
Global	$\mathcal{P}(\mu')$	2Dt	$\mathcal{N}(\mu, \sigma^2)$	GP2DtB
Global	$\mathcal{P}(\mu')$	NIG	μ	GPNIGA
Global	$\mathcal{P}(\mu')$	NIG	$\mathcal{N}(\mu, \sigma^2)$	GPNIGB
Global	$ZI\mathcal{P}(1-q, \mu')$	Exponential	μ	GZExpoA
Global	$ZI\mathcal{P}(1-q, \mu')$	Exponential	$\mathcal{N}(\mu, \sigma^2)$	GZExpoB
Global	$ZI\mathcal{P}(1-q, \mu')$	2Dt	μ	GZ2DtA
Global	$ZI\mathcal{P}(1-q, \mu')$	2Dt	$\mathcal{N}(\mu, \sigma^2)$	GZ2DtB
Global	$ZI\mathcal{P}(1-q, \mu')$	NIG	μ	GZNIGA
Global	$ZI\mathcal{P}(1-q, \mu')$	NIG	$\mathcal{N}(\mu, \sigma^2)$	GZNIGB
Individual	$\mathcal{P}(\mu')$	Exponential	μ	IPExpoA
Individual	$\mathcal{P}(\mu')$	Exponential	$\mathcal{N}(\mu, \sigma^2)$	IPExpoB
Individual	$\mathcal{P}(\mu')$	2Dt	μ	IP2DtA
Individual	$\mathcal{P}(\mu')$	2Dt	$\mathcal{N}(\mu, \sigma^2)$	IP2DtB
Individual	$\mathcal{P}(\mu')$	NIG	μ	IPNIGA
Individual	$\mathcal{P}(\mu')$	NIG	$\mathcal{N}(\mu, \sigma^2)$	IPNIGB
Individual	$ZI\mathcal{P}(1-q, \mu')$	Exponential	μ	IZExpoA
Individual	$ZI\mathcal{P}(1-q, \mu')$	Exponential	$\mathcal{N}(\mu, \sigma^2)$	IZExpoB
Individual	$ZI\mathcal{P}(1-q, \mu')$	2Dt	μ	IZ2DtA
Individual	$ZI\mathcal{P}(1-q, \mu')$	2Dt	$\mathcal{N}(\mu, \sigma^2)$	IZ2DtB
Individual	$ZI\mathcal{P}(1-q, \mu')$	NIG	μ	IZNIGA
Individual	$ZI\mathcal{P}(1-q, \mu')$	NIG	$\mathcal{N}(\mu, \sigma^2)$	IZNIGB

TABLE 4.1: Plan d'expérience utilisé pour la comparaison des modèles.

Paramètre	Distribution
β_1	$\mathcal{U}(0, 10)$
β_2	$\mathcal{U}(-150, 150)$
σ^2	$Inv\mathcal{G}amma(0.001, 0.001)$

TABLE 4.2: Distribution *a priori* des paramètres des modèles d'observations. $\mathcal{U}$: loi uniforme ; $Inv\mathcal{G}amma$: loi Gamma inverse.

FD	Paramètre	Borne inférieure	Borne supérieure	Moyenne	ET	CV
2Dt	a	0	20	10	100/3	10/3
2Dt	b	1	2	1.5	0.29	0.19
2Dt	κ	0	3	1.5	0.75	0.5
NIG	λ_x	0	1	0.5	0.29	0.58
NIG	λ_z	0	1	0.5	0.29	0.58
NIG	δ_x	0	1	0.5	0.29	0.58
CEx	K_e	0	1	0.5	0.29	0.58
CEx	a_1	0	2	1	0.58	0.58
CEx	a_2	0	a_1	$a_1/2$	$a_1/\sqrt{12}$	0.58
CEx	D	1	10	5.5	2.6	0.47
CEx	κ	0	3	1.5	0.75	0.5

TABLE 4.3: Caractéristiques des distributions *a priori* des paramètres des fonctions de dispersion. FD : fonction de dispersion, ET : écart-type, CV : coefficient de variation (= ET/Moyenne).

1.3 Prédiction

Les qualités de prédiction des différents modèles testés ont fait l'objet d'un important travail de comparaison pour cette thèse. Les méthodes et critères statistiques employés seront présentés dans la section 3, ainsi que les résultats obtenus pour les trois stratégies évoquées dans la section 1.2. L'article soumis à *Environmental Modelling & Software* et disponible en annexe contient également des comparaisons détaillées sur les qualités prédictives des modèles.

Dans cette section consacrée à la prédiction, nous n'insistons donc pas plus sur les comparaisons et la sélection entre les modèles. Nous discutons par contre de leur aptitude à représenter les différentes sources de variabilité et d'incertitude sur les flux de gènes, via les prédictions probabilistes basées sur la distribution a posteriori des paramètres.

Pour ne pas alourdir inutilement le mémoire, nous considérons uniquement les résultats relatifs à la troisième stratégie d'estimation définie dans la section 1.2 (i.e. jeu d'entrainement : Montargis 98 et 99, jeu de validation : Mas Cebria) car elle semble plus réaliste. En effet, cette stratégie est celle qui utilise le plus de données pour l'estimation ; hors dans un contexte où on cherche des modèles prédictifs, on préférera toujours calibrer les modèles avec le maximum de données possibles. Par ailleurs, nous limitons la présentation des résultats à un petit nombre de modèles et, lorsque c'est possible, nous généralisons nos interprétations à l'ensemble des modèles. Nous nous concentrons dans un premier temps sur la fonction de dispersion CEx avec le modèle d'observation ZIP et le mode de représentation *global*. Nous avons dans notre plan d'expérience deux modèles qui satisfont ces critères : $GZExpoA$ et $GZExpoB$. C'est donc sur l'analyse des prédictions de ces deux modèles que nous concentrons nos efforts. Dans un second temps, nous tentons d'analyser l'effet du mode de représentation de la dispersion sur la qualité des prédictions, pour cela nous nous concentrerons sur les deux modèles $GZExpoA$ et $IZExpoA$.

On rappelle qu'une partie de nos objectifs est de permettre une modélisation satisfaisante de l'espérance et de la variance du taux de pollinisation croisée aussi bien pour de très courtes que pour de très longues distances. La figure 4.7 est une représentation des données du jeu d'entrainement. Elle représente les observations du nombre de grains bleus (représentant l'OGM dans les expériences de Montargis 98 et 99), regroupées en boxplot par classes de distances à l'émetteur OGM le plus proche. La décomposition du premier graphe de la figure en deux sous-graphes pour chaque grande direction (sous le vent : *Downwind* et contre le vent : *Upwind*) a été réalisée pour permettre de mieux observer la *vraie* variabilité locale de la réponse. Celle-ci est en effet confondue avec la variabilité due à l'anisotropie du processus lorsqu'on regarde dans toutes les directions simultanément. On précise ici que la même construction sera utilisée pour les graphes des prédictions.

Sous la même forme que la figure 4.7, les figures 4.8 et 4.9 comparent les données observées et les données prédites par les modèles $GZExpoA$ et $GZExpoB$ respectivement. Globalement, si on ne regarde que le graphique du haut dans ces deux figures, les deux modèles permettent une prise en compte et une retranscription de la variabilité observée qui paraît satisfaisante. Néanmoins, la variabilité totale des observations (sous et contre le vent) est bien mieux reproduite par les prédictions du modèle $GZExpoB$. Cette meilleure

retranscription de la variabilité des observations est toujours vérifiée lorsque le modèle d'observations pour l'espérance du taux de pollinisation croisée $\mu_s^{'}$ passe de fixe à aléatoire, et lorsque le modèle d'observations intègre la distribution ZIP pour le nombre de grains OGM. Cela indique que la version aléatoire de l'espérance du taux de pollinisation croisée $\mu_s^{'}$ permet aux prédictions de bien s'adapter à la variabilité des observations non prise en compte par le modèle ZIP.

Il faut noter ici qu'on n'observe pas le même comportement lorsque la loi de Poisson est utilisée pour modéliser la distribution de probabilité du nombre de grains OGM. Dans ce cas, le modèle à espérance $\mu_s^{'}$ fixe donne de meilleurs résultats (en espérance et en variance) que le modèle à espérance aléatoire normale. On rappelle que dans le modèle de Poisson il n'y a pas de procédure spécifique pour estimer le poids des zéros dans les observations. De ce fait, toute la variabilité des observations est gérée par les variations de $\mu_s^{'}$. La version aléatoire pour l'espérance du modèle de Poisson a donc un effet pervers : dit simplement, tout se passe comme si cette option laissait à $\mu_s^{'}$ le soin de gérer les zéros et à σ^2 celui de gérer les taux exceptionnellement élevés. On observe en pratique, avec le modèle de Poisson à espérance aléatoire normale, que $\mu_s^{'}$ tend vers zéro et que σ^2 prend des valeurs anormalement élevées. En revanche, la loi ZIP permet d'avoir des paramètres spécifiques pour les zéros et, de ce fait, $\mu_s^{'}$ et σ^2 ne gèrent pas toutes la variabilité des observations mais uniquement la variabilité des observations non nulles.

La figure 4.10 présente les boxplots des moyennes et écart-types prédits avec les deux modèles pour différentes classes de distances ainsi que les observations correspondantes. On observe que les deux types de modèles ont tendance à sous-estimer la moyenne observée dans les premiers mètres, mais seulle modèle à espérance $\mu_s^{'}$ fixe sous-estime l'écart-type. Inversement, les modèles à espérance $\mu_s^{'}$ aléatoire surestiment l'écart-type dans les premiers mètres. Toutes ces remarques permettent de conclure que le choix d'un modèle d'observation a un très fort impact sur la moyenne et l'écart-type prédits. Ce choix est donc crucial pour permettre une modélisation satisfaisante de l'espérance et de la variance des observations. Nous recommandons donc d'utiliser, pour ce type de données des modèles d'observations qui prennent en compte l'excès de zéros par rapport aux distributions usuelles et tout particulèrement le modèle ZIP.

Pour qualifier l'effet du mode de représentation de la dispersion sur la qualité des prédictions, nous avons regroupé et moyenné les observations sur des surfaces homogènes (micro-parcelles) et représenté les prédictions correspondantes. La figure 4.11 montre les moyennes a posteriori des prédictions en fonction des observations correspondantes sur les données d'ajustement et la figure 4.12 montre les mêmes quantités sur les données de validation.

Dans l'ensemble et contrairement à ce qui était attendu, on observe de faibles performances pour les prédictions issues du mode de prise en compte individuel sur les données de validation. Bien que les corrélations entre les valeurs observées et prédites soient similaires entre les deux modes de représentation, le mode individuel surestime les valeurs observées et, par conséquent, leur variabilité tandis que le mode global a tendance à sous-estimer les valeurs observées. Cela montre plusieurs choses : d'une part, un biais peut toujours survenir même en utilisant des modèles plus complexes qui, a priori, s'adaptent

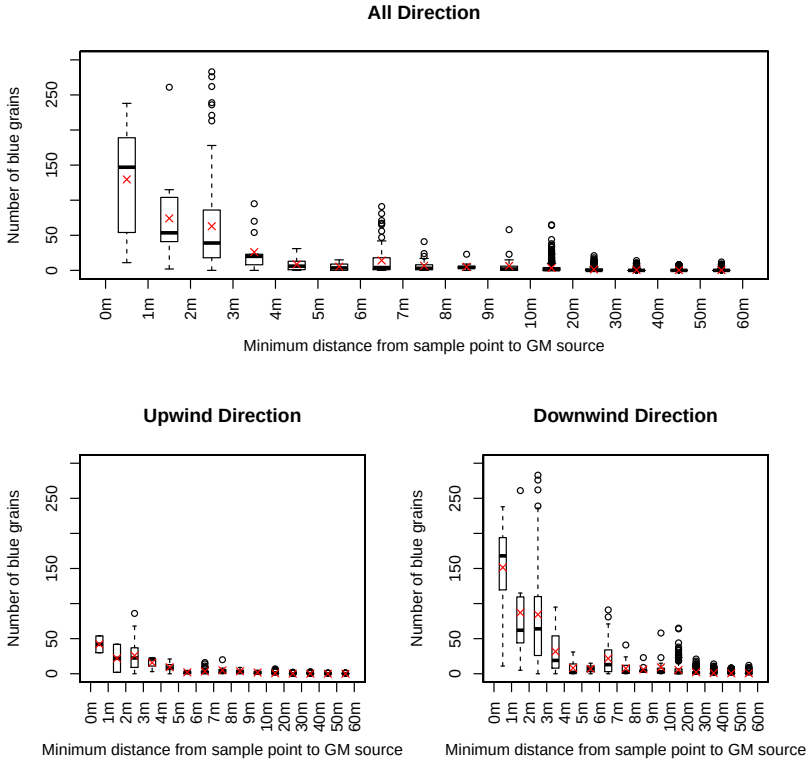


FIGURE 4.7: Boxplots du nombre de grains bleus par épi en fonction de la distance à la source OGM la plus proche. Les croix rouges correspondent à la moyenne du nombre de grains bleus pour chaque classe de distance.

plus facilement à des contextes très différents de ceux utilisés pour l'estimation. D'autre part, l'importance est confirmée d'une validation en bonne et due forme sur un ensemble de données indépendantes, ainsi que de critères appropriés en fonction de l'objectif de la prédiction. En effet, le modèle qui donne le meilleur ajustement au jeu d'entrainement n'est pas nécessairement le meilleur modèle pour faire des prédictions dans une situation indépendante.

Il convient de noter ici que la taille du champ de émetteur est beaucoup plus grande dans le jeu de validation dans le jeu d'entraînement (environ 80 fois plus grande). Ces différences de tailles peuvent expliquer les faibles performances du mode individuel. Si on regarde les résultats des deux autres stratégies d'estimation (tableau 4.7 et 4.9) pour lesquelles il y on a moins de différences de tailles entre les parcelles émettrices du jeu d'estimation et jeu de validation, on constate qu'il y a moins de divergence entre les deux modes de représentation. Cela indique que les différences de tailles de surfaces émettrices entre le jeu d'entrainement et de validation ont vraisemblablement un grand impact sur la capacité prédictive et ce dans les deux modes de prise en compte de la dispersion. Toutes ces observations plaident pour l'utilisation d'un ensemble plus représentatif de la diversité des situations possibles dans le jeu d'entrainement, incluant des caractéristiques de tailles et de formes des parcelles très différentes.

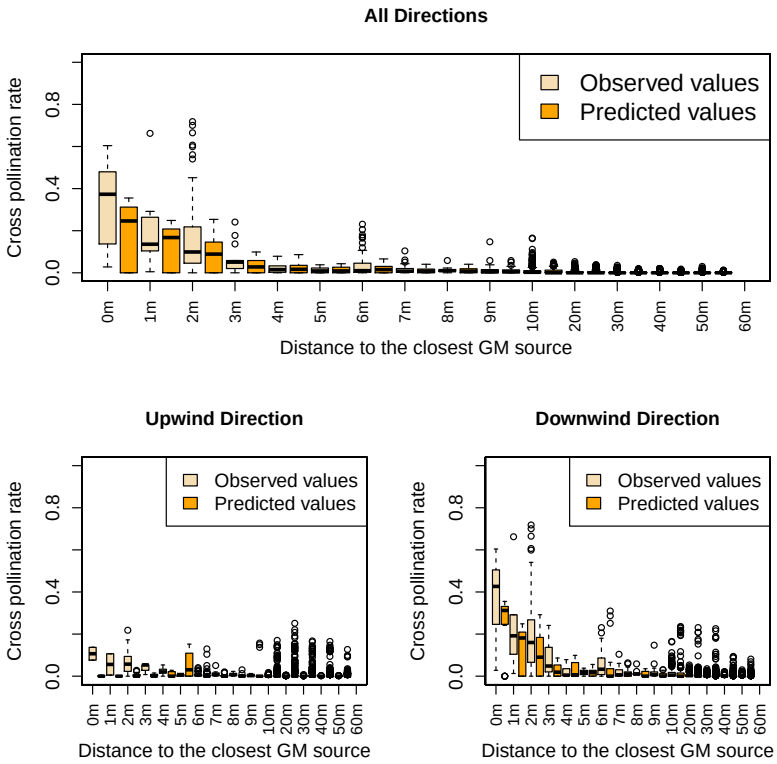


FIGURE 4.8: Boxplots des taux de pollinisation croisée observés et prédits (par le modèle GZExpoA) en fonction de la distance à la source OGM la plus proche.

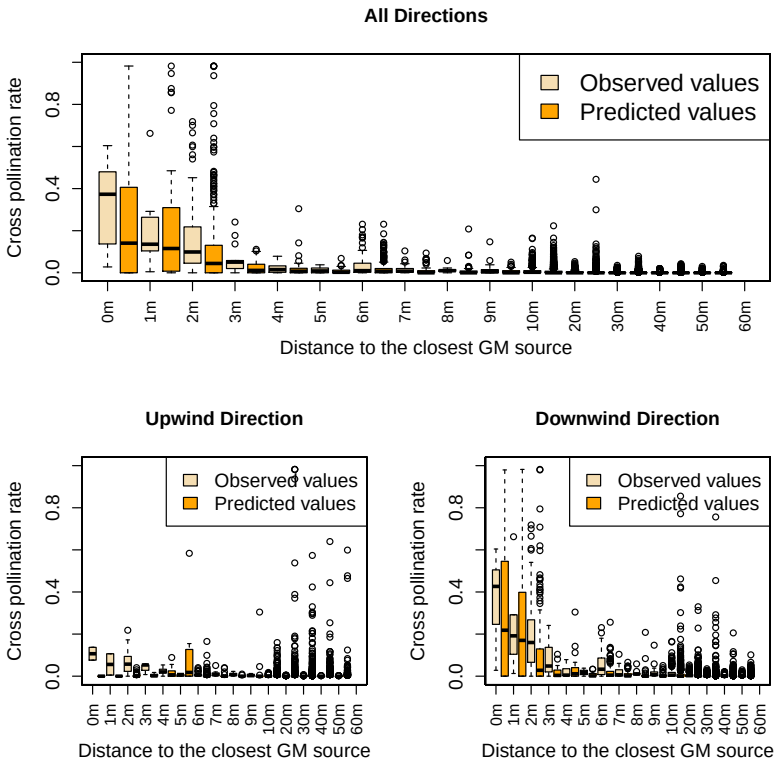


FIGURE 4.9: Boxplots des taux de pollinisation croisée observés et prédits (par le modèle GZExpoB) en fonction de la distance à la source OGM la plus proche

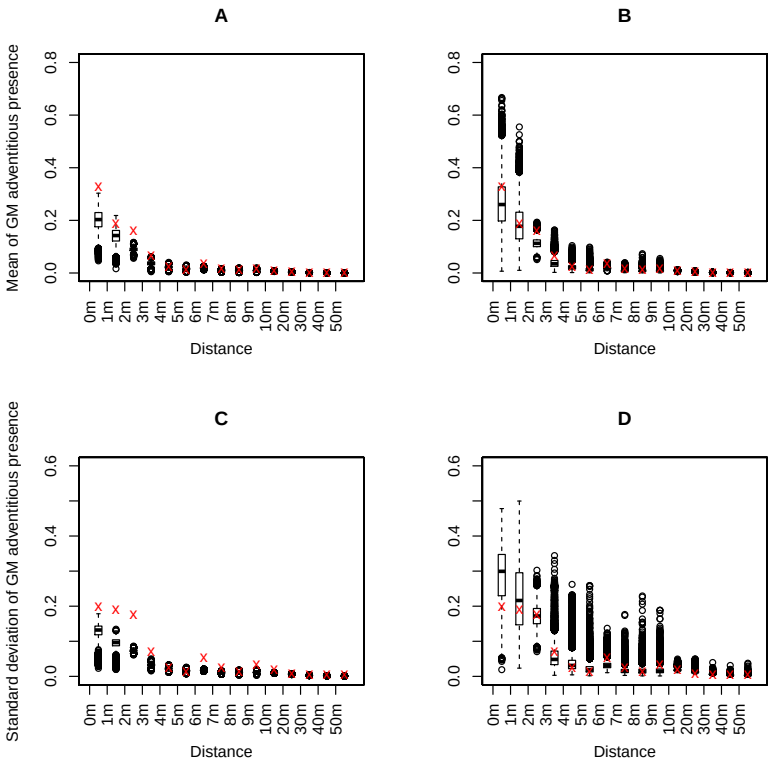


FIGURE 4.10: Boxplot des moyennes (A, B) et des écarts-types (C, D) du taux de pollinisation croisée prédit en fonction de la distance à la source OGM la plus proche. Graphes A et C (respectivement B et C) : prédictions issues du modèle $GZExpoA$ (respectivement $GZExpoB$). Les croix rouges correspondent à la moyenne (A, B) et à l'écart type (C, D) du taux observé par intervalle de distance.

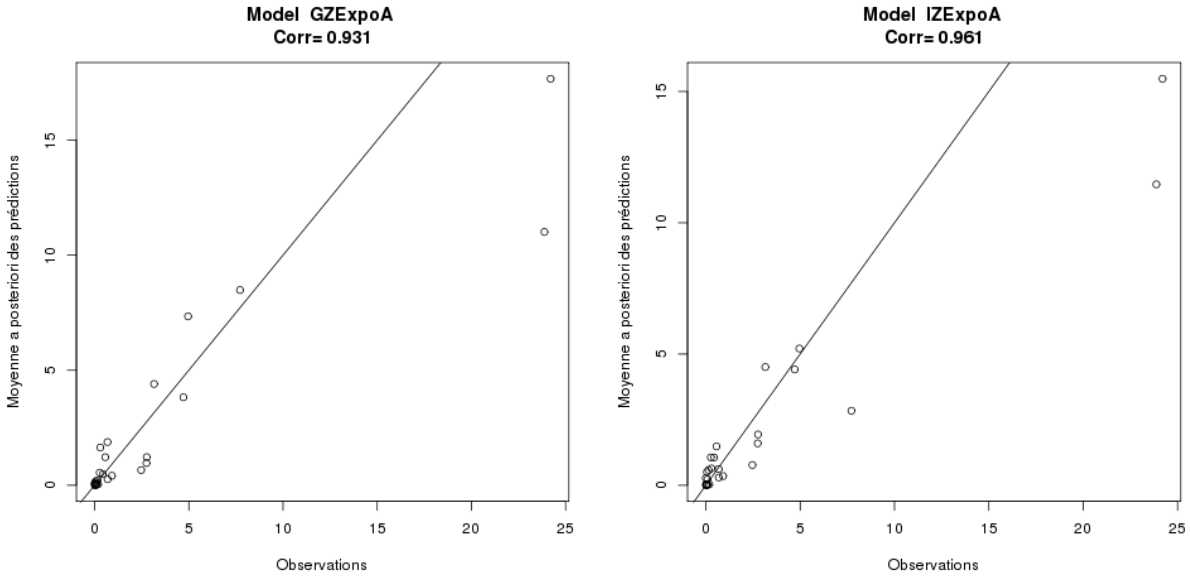


FIGURE 4.11: Moyennes a posteriori des prédictions moyennes par micro-parcelle, en fonction des observations correspondantes sur les données d'ajustement. À Gauche (resp. à droite), prédiction du modèle $GZExpoA$ (resp. $IZExpoA$).

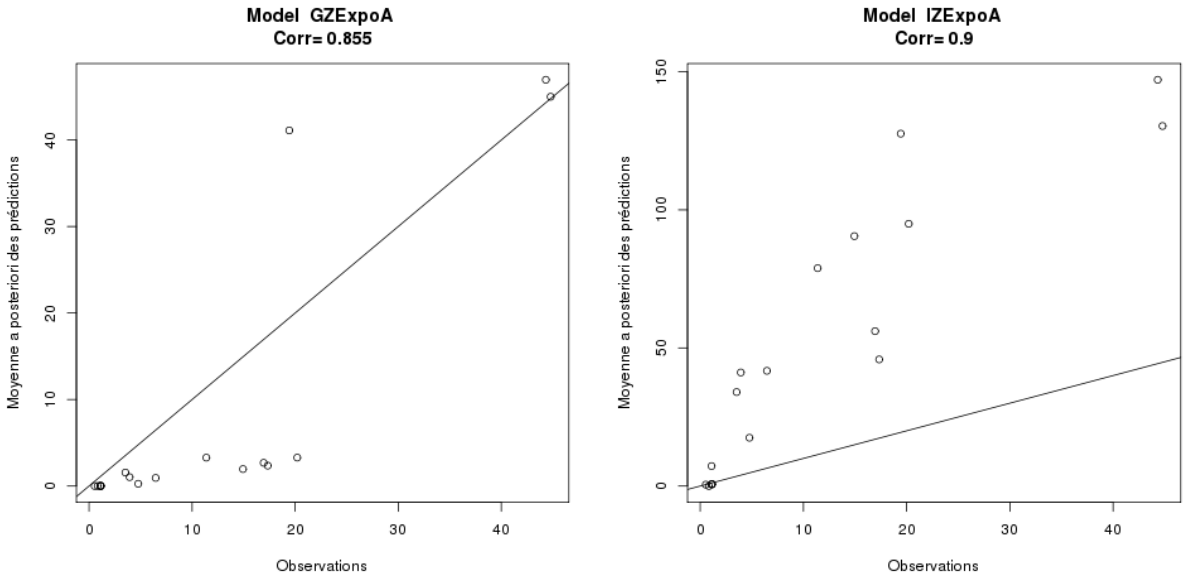


FIGURE 4.12: Moyennes a posteriori des prédictions moyennes par micro-parcelle, en fonction des observations correspondantes sur les données de validation. À Gauche (resp. à droite), prédiction du modèle $GZExpoA$ (resp. $IZExpoA$).

2 Contexte multi-sources

Dans cette section, nous explorons la possibilité de réaliser l'estimation des paramètres de nos modèles de dispersion, définis au Chapitre 3, en utilisant des données de nature très différente de celles utilisées dans la section précédente. Par rapport aux modèles estimés avec les données issues d'essais *mono-source*, il y a deux modifications principales

à apporter à la définition des modèles pour permettre leur estimation sur les données dont nous disposons et issues de dispositifs *multi-sources* :

1. utiliser un modèle d'observation pour des données continues (voir section 4, Chapitre 3) ;
2. permettre la prise en compte des décalages de floraisons potentiels entre deux variétés de maïs (voir section 3, Chapitre 3).

Nous décrivons dans un premier temps les données utilisées. Puis nous abordons l'intégration des décalages de floraison et de la prise en compte de la multiplicité des sources sur la définition du modèle de l'espérance. Dans un second temps nous présentons les stratégies qui ont été déployées pour réaliser l'estimation des paramètres sur ces données. Enfin, nous évaluons la qualité prédictive des modèles estimés sur des données issues de contextes *mono-source* lorsqu'ils sont utilisés en prédiction dans des contextes *multi-sources*.

2.1 Données

Les données dont nous avons disposé pour le contexte *multi-sources* sont celles décrites dans Messeguer et al. (2006). Contrairement aux jeux de données utilisés dans la section précédente, celui-ci n'est pas construit sur la base d'une expérience en condition contrôlée mais issu d'un dispositif de surveillance de parcelles de variétés non GM en situation réelle de coexistence avec des parcelles de variétés GM.

Ces données ont été acquises dans la zone de Foixà sur cinq années consécutives, de 2004 à 2008. Sur toute cette période, un total de de 33 variétés de maïs ont été utilisées, dont 13 étaient étaient des variétés génétiquement modifiées. Une représentation schématique de la répartition des parcelles pour une année donnée est fournie dans la figure 2.4 du Chapitre 2. Les données météorologiques ont été recueillies par l'IRTA dont la station expérimentale se situe à 7 km de la zone de Foixà.

Pendant les cinq années de suivi (2004 à 2008), différentes parcelles de maïs conventionnel ont été échantillonnées afin d'estimer le taux de pollinisation croisée au moyen d'analyse PCR : sept parcelles en 2004, cinq en 2005, cinq en 2006, sept en 2007 et trois en 2008. Ces parcelles ont été sélectionnés selon la proximité des champs OGM environnants mais aussi en fonction des décalages existants dans la dynamique de floraison afin de représenter une gamme variée de situations de coexistence (Messeguer et al., 2006). Selon les parcelles et les années, l'échantillonnage a consisté en trois à dix épis par point de prélèvement. Les points ont été déterminés d'après un plan "standard", suivant le contour, en transect ou selon un plan stratifié. Ces plans d'échantillonnages sont illustrés par la figure 4.13. Pour une même parcelle et une année donnée, il peut y avoir jusqu'à trois de ces différents plans d'échantillonnage superposés. Le nombre de mesures par parcelle et pour une année donnée varie donc entre six et soixante. La figure 4.14 est un tableau regroupant, pour chaque année et chaque parcelle, le nombre d'échantillons prélevés et le ou les plans d'échantillonnage utilisés.

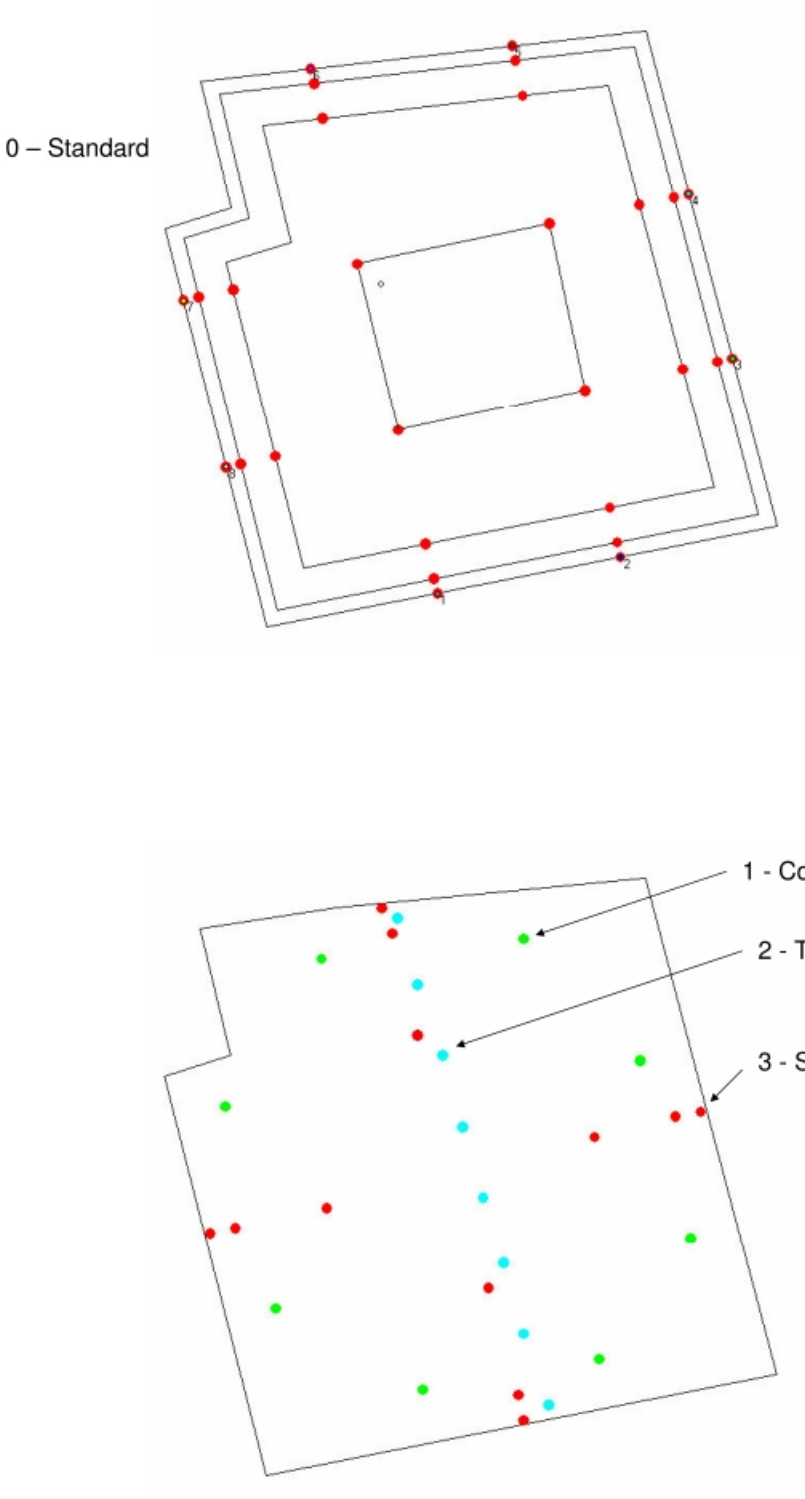


FIGURE 4.13: Plans d'échantillonnage utilisés pour les prélèvements dans la zone de Foixà (Messeguer et al., 2006).

Echantillonnage PCR Foixa

Nombre de points d'échantillonnage

Année	IDFIELD	Total	Standard	Contour	Transect	Stratified
2004	3120	39	39			
2004	3122	20	20			
2004	3194	20	20			
2004	3802	28	28			
2004	4017	28	28			
2004	4030	28	28			
2004	4033	17	17			
2005	3024	37	37			
2005	3114	6	6			
2005	3115	28	28			
2005	3122	28	28			
2005	3801	28	28			
2006	3105	56	28	8	8	12
2006	3192	56	28	8	8	12
2006	4017	56	28	8	8	12
2006	4030	56	28	8	8	12
2006	4040	60	32	8	8	12
2007	3101	44	28	8		8
2007	3103	15		8		7
2007	3192	44	28	8		8
2007	3540	44	28	8		8
2007	3800	16		8		8
2007	4017	16		8		8
2007	4196	44	28	8		8
2008	3117	16		8		8
2008	4016	16		8		8
2008	4017	16		8		8

0-Standard	ech à 0,3,10 et 30m à partir de 8 points d'entrée dans la parcelle - 3 épis par points d'ech
1-Contour	ech entre 3 et 10m - 2 rep sur chaque bord - 10 épis par points d'ech
2-Transect	ech tous les 10m sur une seule route - 10 épis par points d'ech
3-Stratified	ech à 0,3,10 et 30m à partir de chaque bord - 10 épis par points d'ech

2.2 Intégration du décalage de floraison

Comme on l'a vu au Chapitre 3, la prise en compte des décalages de floraison pour la modélisation du taux de pollinisation croisée dans des situations réelles de coexistence est nécessaire. En effet, le synchronisme observé, et même voulu par souci de simplification, dans la quasi totalité des essais *mono-source* relève plus de l'exception que de la règle dans les situations réelles. L'équation définie au Chapitre 3, section 3 pour prendre en compte l'effet des décalages de floraison sur le taux de pollinisation croisée sera donc intégrée aux modèles lorque ceux-ci sont estimés sur des données issues de situations *multi-sources*. On rappelle ici cette équation :

$$DF = \frac{1}{1 + \alpha \left| f_{\text{GM}} - f_{\text{nonGM}} \right|} \tag{4.1}$$

où f_{nonGM} et f_{GM} sont les dates de floraison des parcelles nonOGM et OGM respectivement, et α est un paramètre à estimer. Ilest important de considérer ici qu'il faut prendre en compte la multiplicité des sources émettrices de pollen et de ce fait, la valeur de DF doit être calculée pour chaque paire de parcelles émettrice/réceptrice. D'autre part, en fonction du mode utilisé pour la représentation des échanges entre parcelles (global ou individuel, voir Chapitre 3, section 1), le modèle de l'espérance s'écrit différemment.

Dans le mode global, le modèle de l'espérance du taux de pollinisation croisée d'un épi situé en un point s et dans une configuration *multi-sources* s'écrit :

$$\mu_s = \sum_{i=1}^{L} \gamma(s, s_i^{'}) \times DF_i \tag{4.2}$$

où L est le nombre de parcelles OGM dans l'environnement du récepteur, $s_i^{'}$ est le point de la parcelle i le plus proche du point récepteur s, $\gamma(\cdot, \cdot)$ est une fonction de dispersion quelconque, et DF_i est le décalage de floraison entre la parcelle réceptrice et la parcelle i, tel que défini dans l'équation (4.1).

Dans le mode individuel, le modèle de l'espérance du taux de pollinisation croisée d'un épi situé en un point s et dans une configuration *multi-sources* s'écrit :

$$\mu_s = \frac{\displaystyle\sum_{i=1}^{L} \Big(\sum_{s^{'} \in i} \gamma(s, s^{'}) \Big) \times DF_i}{\displaystyle\sum_{i=1}^{L} \Big(\sum_{s^{'} \in i} \gamma(s, s^{'}) \Big) \times DF_i + \sum_{j=1}^{M} \Big(\sum_{s^{'} \in j} \gamma(s, s^{'}) \Big) \times DF_j}, \tag{4.3}$$

où L est le nombre de parcelles OGM et M le nombre de parcelles non OGM dans l'environnement du récepteur, i (respectivement j) parcourt l'ensemble des parcelles OGM (resp. non OGM) dans le paysage considéré. $\gamma(\cdot, \cdot)$ est une fonction de dispersion quelconque, DF_i (respectivement DF_j) est le décalage de floraison, tel que défini dans l'équation 4.1, entre la parcelle réceptrice et la parcelle i (resp. j).

2.3 Stratégies pour l'estimation

De nombreuses stratégies ont été déployées pour estimer les paramètres avec le jeu de données multi-sources disponible. Disons tout de suite qu'aucune d'entre elles n'a permis d'obtenir une estimation satisfaisante en termes de convergence vers la distribution a posteriori des paramètres. Il nous semble néanmoins très important et scientifiquement intéressant de décrire et discuter ces stratégies ainsi que les hypothèses sous-jacentes, de manière à *i)* éviter que d'autres essuient les mêmes *échecs* ; *ii)* donner des perspectives pour de futurs programmes de recherche.

On précise que dans toutes les stratégies d'estimation déployées dans cette partie, nous avons restreint l'analyse à la fonction de dispersion *Compound Exponential*. Nous avons appliqué les mêmes modèles et les mêmes méthodes (MCMC) que dans le cas *mono-source*, au modèle d'observations et à la prise en compte du décalage de floraison et de la multiplicité des sources près. La dénomination des modèles reprend les *codes* de la section précédente ; la première lettre (G ou I) indique le mode de prise en compte des échanges de pollen entre parcelles, la deuxième (P, Z ou N) représente le modèle d'observation, les quatre suivantes identifient le noyau de dispersion (2Dt, NIG ou Expo) et la dernière (A ou B) le modèle pour le paramètre $\mu_s^{'}$ du modèle d'observation (A pour fixe, B pour aléatoire).

La première stratégie mise en place a consisté à utiliser les données ponctuelles décrites ci-dessus pour estimer les paramètres. Cette stratégie n'a pas permis d'obtenir la convergence des algorithmes d'estimation. Les figures 4.15 et 4.16 montrent les chaînes de Markov issues de ces estimations pour les modèles $GNExpoA$ et $INExpoA$. On voit sur ces figures que, soit l'estimation ne converge pas vers une distribution stationnaire, soit elle converge exactement vers la distribution *a priori*. À ce stade, plusieurs explications peuvent êtres avancées sur l'origine de ce comportement. Néanmoins, celle qui considère que la trop grande variabilité des observations réduit les capacités de convergence des algorithmes d'estimation nous a semblé la plus convaincante. Nous avons donc cherché par la suite et via différentes méthodes ou approches à réduire cette variabilité. Les deux paragraphes suivant décrivent brièvement ces méthodes.

La seconde stratégie déployée a eu pour objectif de diminuer l'influence de la variabilité initiale dans les données. Elle a consisté à regrouper et moyenner les observations suffisamment proches dans l'espace. En effet, un des plans d'échantillonage réalisé par Messeguer et al. (2006) (le plan dit *standard*) présente en bordure de champs des échantillons très rapprochés. La partie supérieure de la figure 4.13 permet de visualiser ces échantillons, ce sont les huit ensembles de trois points en périphérie de la parcelle échantillonnée. Ces points représentent les prélèvemments d'épis réalisés à 0, 3 et 10 mètres de la bordure de la parcelle. La proximité entre ces points suggère des taux de pollinisation croisée assez proches. Nous avons donc fait la moyenne des ensembles de trois points sur la totalité du jeu de données (2004-2008) et considéré la position du prélèvement à 3 mètres pour cette moyenne. L'estimation a été menée sur ce jeu de données modifié mais n'a pas non plus permis d'obtenir de convergence des algorithmes d'estimation.

La dernière stratégie proposée pour l'estimation des modèles sur ces données consiste à changer radicalement l'échelle d'observations en regroupant et moyennant les observations de chaque parcelle. Cette solution à été testée, néanmoins par définition, la quantité

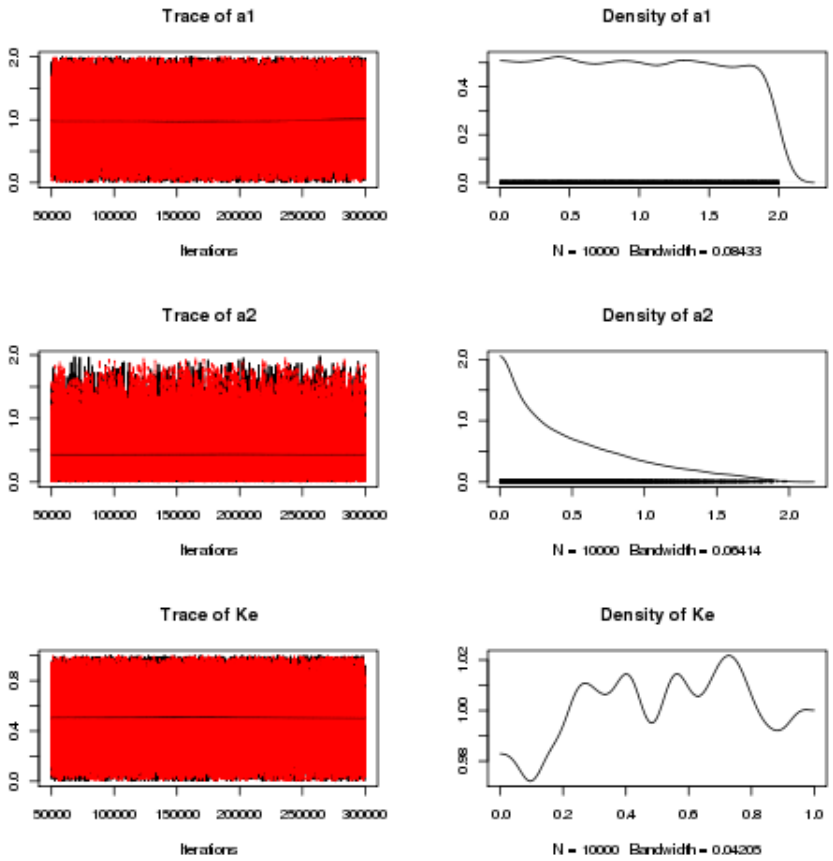


FIGURE 4.15: Traces des chaînes de Markov et distribution a posteriori des paramètres du modèle *GNExpoA* estimés sur les données *multi-sources*.

de données disponibles pour cette solution n'excède pas la somme du nombre de parcelles échantillonnées chaque année soit $7 + 5 + 5 + 7 + 3 = 28$ observations en tout et pour tout. Cette quantité de donnée est très faible, et ce quel que soit la méthode statistique ou le cadre choisi. Nous avons tout de même conduit l'estimation sur ce jeu de 27 observations et, comme indiqué plus haut, cette stratégie n'a pas non plus permis d'obtenir la convergence des algorithmes d'estimation.

Il est important de préciser ici que chacune des trois stratégies présentées a fait l'objet de plusieurs variantes dont aucune n'a permis d'apporter plus d'informations :

- ajout/retrait du module de floraison ;
- ajout/retrait de la prise en compte de sources multiples (retour au cas *mono-source* : tout est résumé par le point le plus proche) ;
- définition d'une fonction de dispersion exponentielle à une pente (au lieu des deux dans la définition initiale) ;
- toutes les combinaisons possibles entre les trois précédentes.

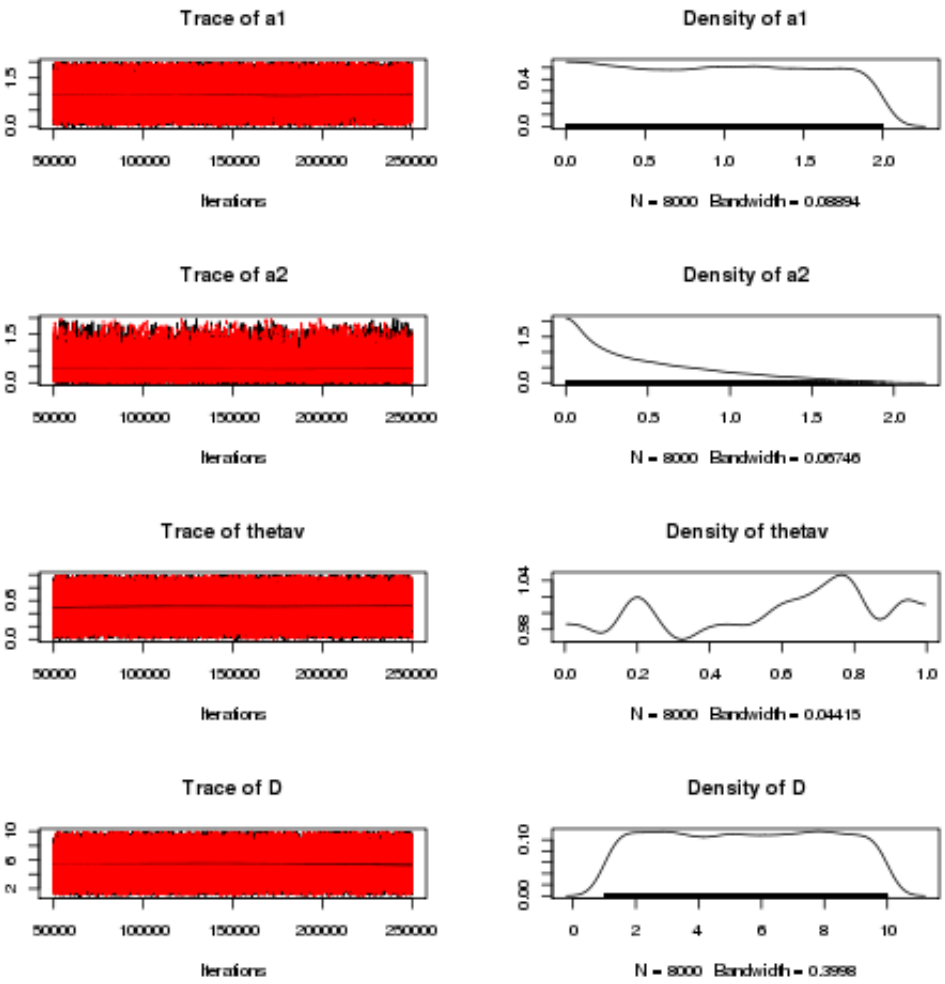


FIGURE 4.16: Traces des chaînes de Markov et distribution a posteriori des paramètres du modèle $INExpoA$ estimés sur les données *multi-sources*

2.4 Prédiction avec modèle ajusté en *mono-source*

Suite aux échecs successifs des tentatives précédemment décrites, nous nous sommes concentrés sur la quantification de la qualité prédictive des modèles estimés dans des contextes *mono-source* lorsqu'ils sont utilisés pour prédire les taux de pollinisation croisée dans des des configurations *multi-sources*. Un des éléments importants à considérer ici est que les modèles d'observations définis pour les modèles de prédiction estimés dans des contextes *mono-source* l'ont été pour des données discrètes de types "comptage de grains", alors que dans les configurations *multi-sources*, les observations sont de types "% de génome" . Les prédictions issues de cette partie seront donc des nombres de grains portant le trangène par épi. L'équation donnée par Pla et al. (2006) a été exploitée ici pour établir la correspondance entre les prédictions du modèle et les observations de Messeguer et al. (2006) issues d'analyse PCR :

$$\%GMnbgrains = \frac{(\%GMgenome + 0.0206)}{0.5085}. \tag{4.4}$$

Où $\%GMgenome$ est la valeur mesurée par PCR qui correspond à un pourcentage de génome transgénique et $\%GMnbgrains$ représente le pourcentage du nombre de grains portant le trangène.

Les prédictions du modèle considérées ici sont les médianes de 1000 simulations du modèle $IPExpoA$. Le tableau 4.5 regroupe les observations et prédictions correspondantes pour les 27 parcelles échantillonnées entre 2004 et 2008, (on rappelle que ces observations sont obtenues en appliquant l'équation (4.4) aux données PCR issues de Messeguer et al. (2006)) la figure 4.17 offre une représentation graphique des prédictions en fonction des observations. Les critères statistiques usuels (racine carrée de l'erreur quadratique moyenne RMSE, coefficient de détermination R^2, et corrélation r) ont été calculés selon les formules 4.6, 4.7 et 4.8 de ce mémoire. Soit $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ le vecteur des données observées et $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ le vecteur des prédictions correspondantes. On note $\sigma_{\mathbf{y}}$ (respectivement $\sigma_{\hat{\mathbf{y}}}$) l'écart-type empirique des observations (resp. des prédictions). D'autre part, de manière à identifier les contributions de différentes sources à l'erreur totale, la décomposition de l'erreur quadratique telle que que donnée par Kobayashi and Salam (2000) a été calculée :

$$\text{MSE} = \text{Biais}^2 + \text{SDSD} + \text{LCS} \tag{4.5}$$

avec

$\text{Biais} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$,

$\text{SDSD} = (\sigma_{\mathbf{y}} - \sigma_{\hat{\mathbf{y}}})^2$,

et $\text{LCS} = 2\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}(1 - r)$.

Cette décomposition offre la possibilité de quantifier différents éléments de l'erreur totale. Elle permet d'identifier les principales sources d'erreur et donne donc des indications sur les éléments à améliorer dans le modèle. Le tableau 4.4 présente ces valeurs.

En premier lieu la RMSE c'est-à-dire la racine carrée de la MSE, qui a la même unité que les observations, est très grande (8.7%), ce qui rend le modèle quasiment inutilisable en pratique, au moins pour prédire les taux de pollinisation croisée avec précision, dans les configurations *mutli-sources*. D'autre part, Le coefficient de détermination R^2 a une valeur négative, cela signifie que l'on prédit moins bien les valeurs observées par le modèle que par la simple moyenne des observations. Ces deux remarques remettent sérieusement en cause l'utilité d'un tel modèle pour prédire de telles situations.

Le biais représente la moyenne de l'erreur du modèle. Contrairement à la RMSE, il peut être positif ou négatif. Il est négatif ici (-4.9) et cela indique que le modèle a tendance à surestimer les valeurs observées. La figure 4.17 permet de visualiser cette surestimation, par exemple par le fait que la majorité des points se situent au-dessus de la bissectrice ou plus simplement par comparaison de l'amplitude des axes. Contrairement aux valeurs des deux autres critères de la décomposition de la MSE, celle-ci est relativement rassurante dans un contexte d'évaluation du risque pour l'aide à la décision. En effet, si les décisions sont prises sur la base des prédictions du modèle, la surestimation des taux de pollinisation croisée conduira à plus de sévérité dans l'établissement de mesure de coexistence et donc à une probabilité plus faible de dépassement des seuils légaux.

Le SDSD (Squared Difference of Standard Deviation) représente la différence entre les écarts-types des observations et des prédictions. Il est très élévé ici et représente à lui seul plus de 50% de la MSE. L'examen des valeurs de ces écart-types ($\sigma_{\mathbf{y}} = 1.44$, $\sigma_{\hat{\mathbf{y}}} = 7.96$) montre que ce sont les prédictions qui présentent trop de variabilité par rapport aux observations. Cette observation pourrait sembler paradoxale avec ce qui a été dit à la section 2.3, mais il faut rappeler que les observations ont été moyennées sur chaque parcelle, ce qui diminue considérablement leur variabilité. D'autre part, le modèle utilisé ici pour les prédictions ayant été calibré sur des données *mono-source* dont l'échantillonnage est très dense et dont la prise en compte de la variabilité a monopolisé une bonne partie de nos efforts, il possède une très grande capacité à *générer* de la variabilité dans les prédictions.

Le LCS est plus délicat à interpréter. En effet, cette quantité fait intervenir la corrélation entre les observations et les prédictions et elle dépend de la capacité du modèle à reproduire les variations des observations d'une situation à l'autre. De ce fait, le LCS peut représenter une multitude de source d'erreurs et il est donc difficile de proposer des améliorations du modèle sur cette base (Wallach et al., 2006). De plus, la contribution du LCS ne représente que 13% de la valeur de la MSE, sa diminution n'est pas une priorité pour l'amélioration du modèle.

Le calcul et l'analyse de ces différents critères permettent finalement de tirer quelques conclusions :

1. le modèle ne prédit pas les taux de pollinisation croisée avec précision ;
2. le biais du modèle est fort et va dans le sens d'une surestimation des observations ;
3. la variabilité des prédictions est trop forte par rapport à celle des observations.

Il faut rappeler ici que le modèle utilisé pour réaliser les prédictions a été calibré sur des données *mono-source* pour lesquelles aucun décalage de floraison n'est observé. De

plus, nous n'avons pas pu obtenir d'estimation satisfaisante des paramètres liés à la prise en compte de ce décalage. De ce fait, les prédictions analysées dans cette partie supposent un synchronisme total, ce qui constitue la situation la plus à risque et peut expliquer une bonne partie de la surestimation des observations par les prédictions. La définition et l'estimation d'un modèle pour la prise en compte du décalage de floraison pourrait donc diminuer significativement le biais du modèle.

En dernier lieu, afin de déterminer si les prédictions de ce modèle peuvent servir d'indicateur pour le classement d'une récolte en OGM ou non, une analyse ROC (Makowski and Monod, 2011) a été réalisée entre les données observées et les prédictions correspondantes. La démarche pour conduire l'analyse est expliquée plus en détail dans la section qui suit. Nous donnons ici les quelques éléments indispensables à sa compréhension ; Ce type d'analyse donne le pourcentage de classement correct vis-à vis d'un seuil de décision donné, ce qui correspond à une qualité de prédiction du dépassement de ce seuil. En d'autres termes, cette qualité s'exprime comme la proportion de prédictions dont les observations correspondantes sont du même *côté* (en-dessous ou au-dessus) par rapport au seuil défini. Ici le seuil a été fixé au seuil légal défini à l'échelle européenne, soit 0.9%. La figure 4.18 présente la courbe ROC déduite de cette analyse. L'AUC est le critère numérique utilisé dans cette analyse, on précise ici qu'un AUC de 50% équivaut à un classement au hasard et qu'un AUC de 100% correspond à un classement parfait. L'AUC calculé entre les prédictions et observations correspondantes pour un seuil de 0.9% est de 86%, cela signifie que dans 86%, le modèle prédit correctement le dépassement ou non du seuil.

Nous sommes conscients du fait que la taille du jeu de données utilisé dans cette évaluation est relativement faible et qu'il faudrait intégrer plus d'observations pour l'analyse soit robuste. Néanmoins, les classements obtenus sur les données *multi-sources* lorsque le modèle utilisé est calibré sur des données *mono-source* sont encourageants et montrent que, malgré la faible capacité du modèle à prédire les taux observés avec précision, on peut tout de même s'en servir pour prendre des décisions.

Critère	Valeur
RMSE	8.696
R^2	-36.923
r	0.514
MSE	75.618
Biais	-4.894
SDSD	40.945
LCS	10.724

TABLE 4.4: Valeurs des critères calculées entre les observations et les prédictions correspondantes du modèle $IPExpoA$ ajusté sur les données *mono-source*.

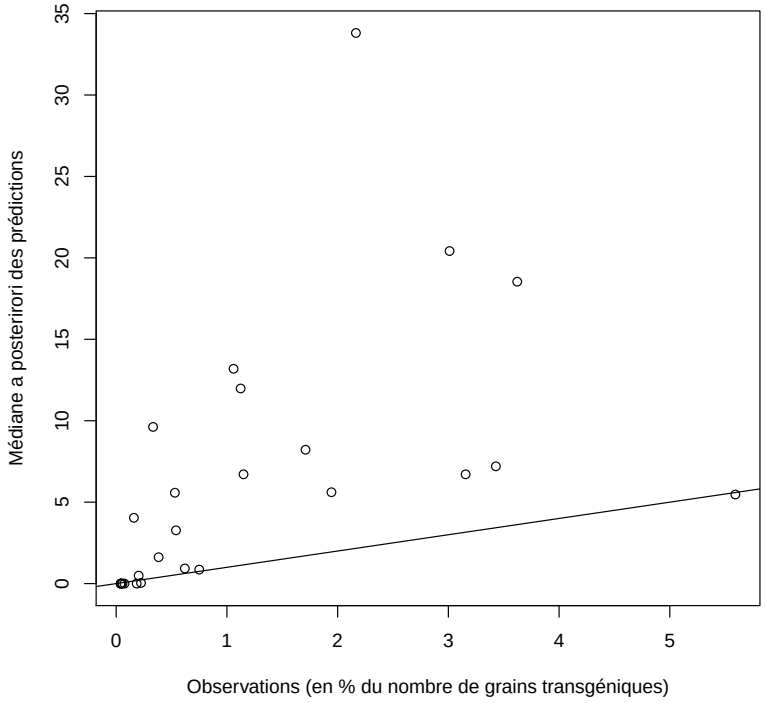


FIGURE 4.17: Médianes a posteriori des prédictions en fonction des observations correspondantes.

Année	IDFIELD	Observation	Prédiction
2004	3120	0.203	0.49
2004	3122	0.042	0.00
2004	3194	0.057	0.00
2004	3802	0.185	0.00
2004	4017	3.429	7.20
2004	4030	3.156	6.71
2004	4033	0.041	0.00
2005	3024	2.165	33.81
2005	3114	3.622	18.54
2005	3115	0.333	9.62
2005	3122	0.041	0.00
2005	3801	0.044	0.01
2006	3105	1.710	8.22
2006	3192	0.750	0.86
2006	4017	0.540	3.27
2006	4030	0.530	5.58
2006	4040	0.620	0.93
2007	3101	0.224	0.04
2007	3103	0.383	1.62
2007	3192	1.944	5.61
2007	3540	1.124	11.98
2007	3800	0.075	0.00
2007	4017	5.592	5.47
2007	4196	3.010	20.42
2008	3117	1.060	13.19
2008	4016	0.160	4.04
2008	4017	1.150	6.71

TABLE 4.5: Comparaison entre observations et prédictions pour les 27 parcelles échantillonnées à Foixà entre 2004 et 2008.

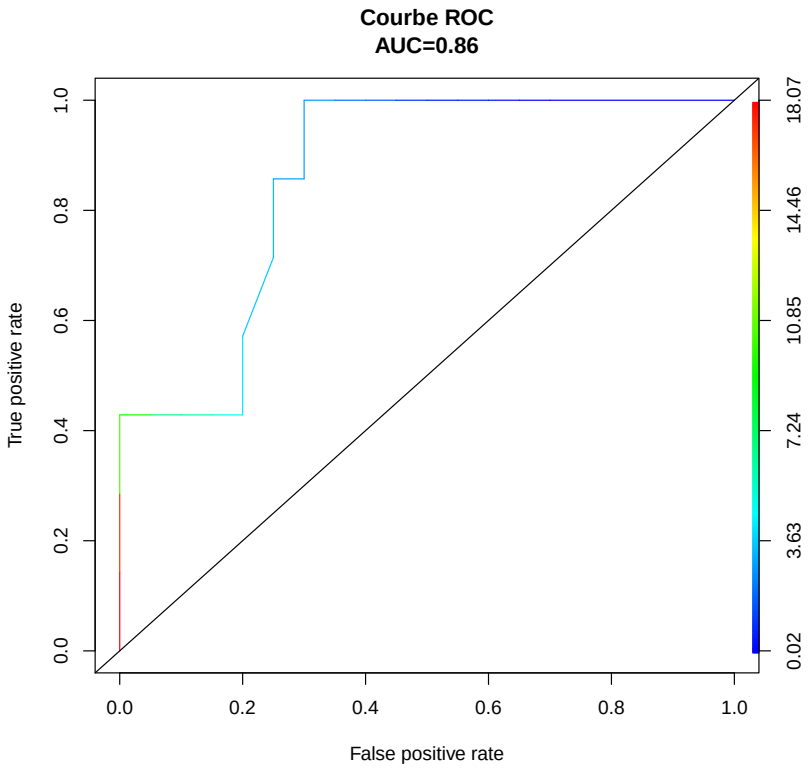


FIGURE 4.18: Courbe ROC pour un seuil de décision de 0.9%.

3 Évaluation et Sélection de modèles

L'évaluation et la sélection de modèles est une étape fondamentale en modélisation et en statistique. Cela a été le cas pour cette thèse, ce qui nous a conduit à comparer plusieurs modèles selon différents critères recommandés dans la littérature statistique. Dans la première section du chapitre, nous nous sommes intéressé aux résultats de notre troisième stratégie d'estimation, qui ne posait pas de problème particulier de cohérence entre les critères. Néanmoins, les résultats des deux premières stratégies d'estimation (lorqu'on utilise uniquement les données de Montargis 98 ou celles de Montargis 99 pour l'estimation) montraient des incohérences et des informations contradictoires entre les valeurs des critères. Par conséquent, il a fallu rechercher une démarche cohérente pour analyser et synthétiser ces informations apparemment contradictoires entre modèles et critères.

Cette section a donc pour objectifs *i)* de définir les critères utilisées pour l'évaluation et la sélection des modèles proposés, quelle que soit la façon dont sont composés les jeux d'entrainement et de validation ; *ii)* de développer un cadre d'analyse pour la sélection de modèles lorsque différents critères donnent des réponses très différentes.

3.1 Critères statistiques

Pour évaluer l'ajustement global des modèles proposés, nous utilisons en premier lieu les critères statistiques classiques sur la réponse moyenne. Ces critères présentent l'avantage d'être facilement calculables et de donner des indications sur le comportement moyen des modèles. Cependant, il existe des critères plus raffinés qui peuvent être utilisés pour *i)* profiter pleinement de la nature probabiliste des prédictions (règles de scoring) ; *ii)* évaluer la qualité de la décision résultant de prédictions du modèle (analyse ROC). Ces deux derniers types de critères sont particulièrement adaptés à notre démarche qui vise précisément à prendre des décisions sur la base de prédictions probabilistes.

3.1.1 Critères classiques

Nous nommons ici critères classiques les critères qui considèrent uniquement la moyenne de la distribution prédictive *a posteriori* pour l'évaluation de la qualité d'ajustement ou de prédiction des modèles. Soit $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ le vecteur des données observées et $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$ le vecteur des réponses moyennes prédites. La racine carrée de l'erreur quadratique moyenne ($RMSE$), le coefficient de détermination (R^2) et la corrélation (r) ont été calculés comme suit :

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{4.6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{4.7}$$

$$r = \frac{\sigma_{\mathbf{y}\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}, \tag{4.8}$$

avec $\sigma_{\mathbf{y}}$, $\sigma_{\hat{\mathbf{y}}}$ et $\sigma_{\mathbf{y}\hat{\mathbf{y}}}$ les écarts-types empirique et la covariance entre $\mathbf{y}$ et $\hat{\mathbf{y}}$.

3.1.2 Critères de scoring

Les critères de scoring permettent d'évaluer la qualité de prédictions probabilistes, en attribuant un score numérique basé sur la distribution prédictive et sur l'événement ou la valeur réellement observée (Gneiting and Raftery, 2007). Par rapport aux critères dits classiques, les critères de scoring ont l'avantage de prendre en compte la nature probabiliste des prédictions.

Comme la prédiction est probabiliste, elle peut être représentée par sa fonction de répartition F pour une observation y_s. Le *Continuous Ranked Probability Score* ($CRPS$) (Gneiting and Raftery, 2007; Lavigne et al., 2012) est défini par :

$$CRPS(F, y_s) = -\int_{-\infty}^{+\infty} (F(x) - \mathbb{1}_{x \geqslant y_s})^2 \mathrm{dx}$$

où $\mathbb{1}_{x \geqslant y_s}$ prend la valeur 1 si $x \geqslant y_s$ et 0 sinon.

Cependant cette intégrale peut être très difficile à calculer. Or, le CRPS peut être exprimé sous une forme facilement calculable à partir des sorties d'un algorithme MCMC comme :

$$CRPS(F, y_s) = \frac{1}{2}\,\mathbb{E}_F|Y - Y'| - \mathbb{E}_F|Y - y_s|$$

où Y et Y' sont deux tirages indépendants dans la distribution prédictive *a posteriori* relative à y_s. Une propriété intéressante du $CRPS$ est qu'en cas de prédiction déterministe son expression se réduit (au signe près) à l'écart absolu entre la valeur prédite et la valeur observée. Il peut donc servir à comparer facilement un système de prédiction probabiliste à un système de prédiction déterministe, ou à l'extraction d'une information ponctuelle (e.g. moments de la distribution) issue d'une prédiction probabiliste. On peut ainsi évaluer l'apport d'information de toute la prédiction probabiliste (exprimée par sa densité de probabilité) par rapport à l'emploi de la seule moyenne et ou médiane.

3.1.3 Critères décisionnel

Nous nommons ici critères décisionnels les critères qui mesurent la qualité d'une décision donnée par rapport à une décision optimale supposée connue. Dans le cas qui nous intéresse, les décisions concernent le classement d'une récolte (ou partie de récolte) en tant qu'OGM (resp. non OGM) si la prédiction du taux de pollinisation croisée dépasse

(resp. ne dépasse pas) la valeur du seuil légal. On est donc dans le cas d'une règle de décision binaire basée sur un indicateur quantitatif, qui est ici la prédiction fournie par un modèle, et sur la valeur du seuil utilisé pour la décision.

La méthode d'analyse appelée *Receiver Operating Characteristics* (ROC) est particulièrement bien adaptée à ce cas. Elle permet en effet à la fois d'évaluer la précision d'une règle de décision binaire basée sur un tel indicateur, de comparer plusieurs indicateurs (c'est-à-dire plusieurs modèles dans notre cas), et d'optimiser leurs seuils de décision (Hanley and McNeil, 1982; Makowski et al., 2008). Ici nous ne nous intéressons qu'à la précision de la règle et à la comparaison des modèles. L'optimisation du seuil n'a pas de sens dans cette application, en effet ce seuil est fixé pour l'Union européenne et il est issu de compromis politiques et non pas d'une quelconque optimisation mathématique.

Le critère numérique utilisé ici est l'aire sous la courbe ROC, notée AUC (Area Under Curve). Cette courbe ROC représente les valeurs de la sensibilité de la règle de décision en fonction des valeurs de "un moins la spécificité", lorsque l'on fait varier le seuil de décision. Rappelons que la sensibilité d'une règle de décision est égale au taux de vrais positifs, soit ici la probabilité de détecter correctement une récolte OGM, alors que la spécificité est égale au taux de vrais négatifs, soit ici la probabilité de détecter correctement une récolte non OGM. Pour un indicateur donné, la sensibilité (resp. la spécificité) est une fonction décroissante (resp. croissante) du seuil de décision. Dit simplement, le critère AUC est d'autant plus élevé qu'il existe une gamme de valeurs du seuil de décision pour lesquelles la sensibilité et la spécificité de la règle de décision sont toutes les deux élevées.

3.2 Analyse en composantes principales

L'évaluation des modèles à l'aide de plusieurs critères présente un avantage certain. Elle permet de comprendre et d'analyser la ou les spécificités de chacun pour une fonction particulière et de sélectionner un modèle en fonction de l'objectif de son utilisation. Cependant la difficulté majeure de ce type d'approche est que des critères de sélection différents peuvent conduire à des conclusions différentes.

Il n'y a pas de réponse claire et tranchée à la question de savoir quel critère devrait être utilisé dans un cas donné. Une possibilité est d'adapter le critère aux objectifs particuliers de l'étude, c'est-à-dire de le construire de telle manière que la sélection favorise le modèle qui estime le mieux la quantité d'intérêt (par exemple le critère AUC dans un contexte d'aide à la décision) et à se focaliser sur ce dernier. Cependant cette dernière approche ne convient pas à notre problématique étant donné la diversité de cas d'utilisation et donc d'objectifs différents qui sont attendus.

L'approche alternative à laquelle nous nous sommes intéressé et que nous décrivons dans cette section repose sur une analyse en composantes principales du tableau de critères et une analyse de sensibilité des critères aux facteurs structurant nos modèles. Elle reprend le cadre des méthodes développées par Lamboni et al. (2009).

3.2.1 Méthode

L'idée principale de la démarche est de considérer, dans les tableaux des critères, que les modèles sont des individus statistiques sur lesquels un certain nombre de variables (les critères) sont mesurées. À partir de cette vision du problème, les méthodes d'analyse en composante principale sont tout indiquées pour extraire une information utile sur les relations entre critères. On applique donc une analyse en composantes principales (ACP) au tableau ayant en lignes les différents modèles et en colonnes les critères d'ajustement et de prédiction.

La seconde idée importante est de poser et d'exploiter une structure factorielle sur les modèles. La situation la plus simple est lorsque ceux-ci résultent de l'ensemble des combinaisons possibles entre les modalités de différentes composantes, ou facteurs. Notons que c'est bien le cas pour les modèles exposés dans le chapitre 3 et étudiés dans le chapitre 4, qui forment un plan factoriel complet pour les facteurs "mode de représentation de la dispersion", "noyau de dispersion", "distribution des observations", "nature de l'espérance"). Dans un tel cadre, l'analyse de la variance apporte de l'information sur le rôle individuel et les interactions entre les différentes facteurs. Nous proposons donc de l'appliquer aux premières composantes principales de l'ACP effectuée sur le tableau des critères.

D'un point de vue technique, la méthode proposée est identique à celle décrite par Lamboni et al. (2009) dans un contexte d'analyse de sensibilité. Elle fournit pour chaque composante principale retenue, une décomposition de sa variance sous la forme d'indice de sensibilité associés aux différents facteurs et à leurs interactions.

3.2.2 Données

Les critères obtenus avec les deux premières stratégies d'estimation de la section 1 sont présentés dans les tableaux 4.6, 4.7, 4.8 et 4.9. Dans ces tableaux, les valeurs en gras sont les trois meilleures valeurs pour un critère donné. On constate qu'on a très rarement une ligne complète de *meilleurs valeurs* (c'est le cas uniquement dans le tableau 4.9, pour le modèle $IP2DtB$) et que lorsque c'est le cas, le classement des modèles n'est jamais conservé lorsqu'on passe de l'ajustement à la prédiction ou lorsqu'on change le jeu de données utilisé pour l'ajustement. La situation de contradiction entre critères est donc effectivement observée avec ces deux stratégies d'estimation, et on constate des réponses différentes *i)* entre les modèles pour un contexte (estimation ou validation) donné ; *ii)* entre les contextes.

L'ACP a été réalisée sur le tableau résultant de la concaténation des tableaux 4.6, 4.7, 4.8 et 4.9 de ce mémoire. Un facteur a été défini et ajouté à l'analyse pour indiquer si les critéres sont calculés en ajustement ou en validation.

3.2.3 Résultats

L'analyse en composantes pricipales montre que 90% de la variabilité entre critères est expliquée par les deux premières composantes. Comme le montre le cercle des corrélations fourni par l'ACP (figure 4.19), le premier axe s'interprète comme une composante de

qualité globale des modèles. Le deuxième axe oppose la corrélation et l'AUC, d'une part, au groupe formé par RMSE, CRPS et le coefficient de détermination R^2, qui varient toujours dans le même sens et apportent donc la même information sur la qualité des modèles.

On identifie ainsi deux classes de critères :

1. une classe de critères représentant un écart à l'observation (RMSE, R^2 et CRPS) ;
2. une classe de critères représentant une corrélation entre observations et prédictions (r et AUC).

D'autre part, la figure 4.20 présente les contributions des critères aux deux premières composantes principales ainsi que les indices de sensibilité des facteurs sur chacune de ces composantes. L'analyse des indices de sensibilité montre que le facteur DF (*Fonction de dispersion*) est le plus sensible sur les deux composantes ; c'est donc le choix de la fonction de dispersion qui a le plus d'impact sur la qualité des modèles. Il faut noter ici que les effets totaux de tous les facteurs ont une grande part d'interactions et il est donc délicat de pousser beaucoup plus loin l'interprétation. On peut néanmoins noter par exemple, que le mode de représentation des échanges de pollen (individuel ou global) a plus d'influence sur la corrélation et l'AUC que sur les autres critères, ce facteur étant très sensible sur la deuxième composante et pas du tout sur la première. Notons aussi que le choix du modèle d'observations est relativement important pour la qualité globale du modèle mais que ce choix doit être consdidérer en interactions avec d'autre facteurs. En outre, le choix de la distribution pour l'espérance du modèle d'observation n'a quasiment pas d'influence sur la réponse en critères des modèles et son choix n'est donc pas primordial. Enfin, le facteur supplémentaire (E) qui indique si les critéres sont calculés en ajustement ou en validation a un indice de sensibilité non négligeable sur la première composante, cela signifie qu'un modèle peut avoir des comportements très différents sur les données d'ajustement et celles de validation. Cette information montre donc l'importance de réaliser une validation systématique des modèles sur des données indépendantes. En effet, les bonnes performances d'un modèle en ajustement n'implique pas de forcément de telles performances sur des données indépendantes.

3.2.4 Discussion

Nous précisons tout de suite que cette approche ne permet pas de déterminer le critère ou groupe de critères le plus pertinent pour faire la sélection ni de sélectionner directement le où les meilleurs modèles. Cependant elle permet dans un premier temps de mieux comprendre comment se comportent les critères les uns par rapport aux autres (i.e si certains apportent ou non une information très différente des autres) et, dans un second temps, de savoir quelles composantes des modèles (les facteurs de l'analyse) influencent le plus la valeur des critères (i.e. à quels facteurs les critères sont-ils le plus sensible ?). Cette information peut permettre de déterminer les facteurs qui ont le plus d'impact sur les critères et d'identifier a contrario les facteurs dont la définition a peu d'importance pour l'établissement d'un modèle de prédiction.

Nous rappelons que les méthodes décrites ici (ACP et analyse de sensibilité) sont assez classiques en statistique. Cependant nous considérons que leur applications au contexte de l'évaluation et de la sélection de modèle permet d'extraire une information utile et pertinente dans ce cadre. L'application de ce type de méthodes dans ce type de contextes constitue donc une des originalités de ce travail qui mériterait d'être approfondie afin d'être valorisée sous forme de publication scientifique dans une revue d'applications statistiques en agronomie/écologie.

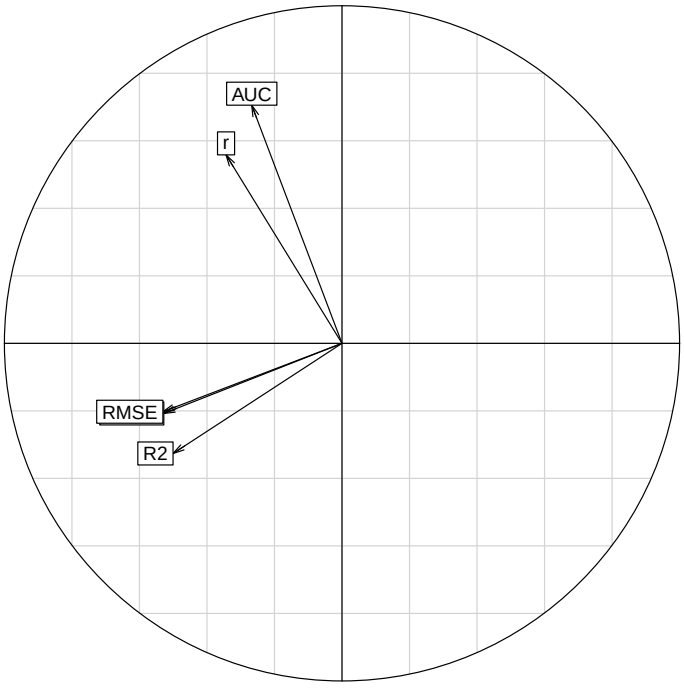


FIGURE 4.19: Cercle des corrélations entre critéres

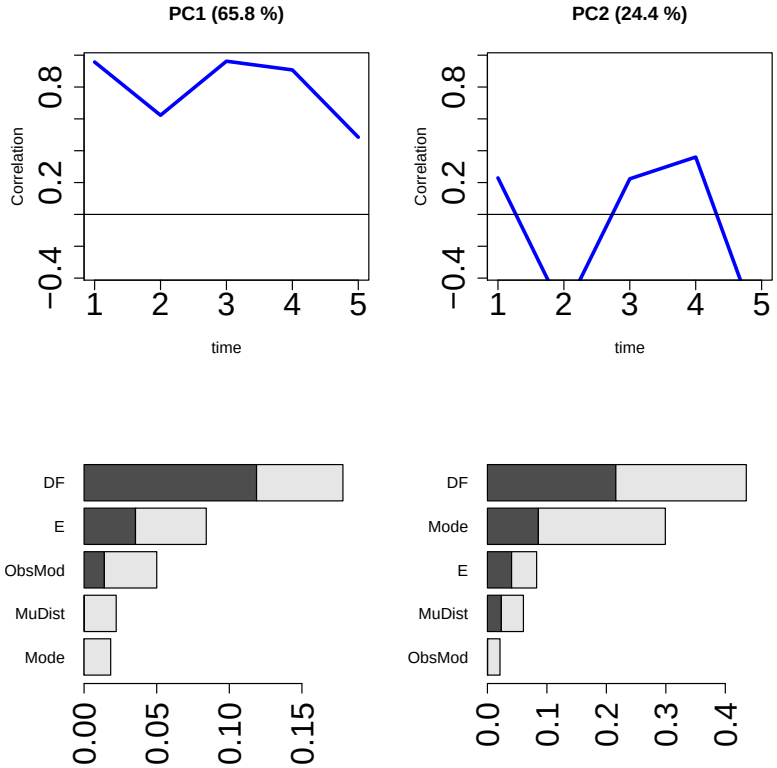


FIGURE 4.20: Contributions des critères aux deux premières composantes principales et indices de sensibilités des facteurs. En abscisse des deux graphiques du haut : 1 : CRPS, 2 : r, 3 : RMSE, 4 : R^2, 5 : AUC. Dans les deux graphiques du bas : indices de sensibilité des facteurs sur les deux premières composantes. La partie noire indique l'indice de sensibilité du premier ordre, l'ensemble de la barre indique l'indice de sensibilité totale. DF :Fonction de dispersion, ObsMod : Modèle d'observations, MuDist :Modèle pour μ, E : Ajustement ou Validation

Model Name	CRPS	r	RMSE	R^2	AUC
GPExpoA	-2.979	0.74	**13.09**	**0.55**	0.9814
GPExpoB	-3.167	**0.751**	**12.9**	**0.56**	0.9799
GP2DtA	-3.379	0.672	15.75	0.34	0.9769
GP2DtB	-4.612	0.627	15.39	0.37	0.9519
GPNIGA	-3.919	0.424	19.83	-0.04	0.9239
GPNIGB	-7.168	0.496	20.43	-0.1	0.8652
GZExpoA	-2.723	0.739	13.22	0.54	0.9814
GZExpoB	-2.977	**0.744**	13.51	0.52	0.9816
GZ2DtA	-3.327	0.667	16.21	0.31	0.9802
GZ2DtB	-3.343	0.651	15.41	0.37	**0.9825**
GZNIGA	-4.083	0.441	19.89	-0.04	0.9318
GZNIGB	-4.636	0.39	19.63	-0.02	0.9688
IPExpoA	-3.185	0.716	14.51	0.44	0.9626
IPExpoB	-3.819	0.713	15.07	0.4	0.9767
IP2DtA	-3.265	0.698	17.08	0.23	0.971
IP2DtB	-3.942	0.697	17.71	0.17	0.9742
IPNIGA	-3.688	0.663	14.76	0.42	0.9716
IPNIGB	-3.542	0.721	15.06	0.4	0.9618
IZExpoA	-2.894	0.71	14.79	0.42	0.9559
IZExpoB	-3.309	0.655	16.53	0.28	0.9816
IZ2DtA	-2.913	0.691	16.68	0.27	0.971
IZ2DtB	-3.234	0.684	16.94	0.24	**0.9822**
IZNIGA	-3.3	0.654	14.93	0.41	0.9725
IZNIGB	-3.139	0.708	15.32	0.38	**0.9834**

TABLE 4.6: Valeurs des critères pour toutes les combinaisons de facteur en ajustement sur Montargis 98

Model Name	CRPS	r	RMSE	R^2	AUC
GPExpoA	-1.145	0.658	6.21	0.31	0.985
GPExpoB	-1.661	0.666	6.23	0.31	**0.9915**
GP2DtA	-0.645	0.644	5.8	0.4	0.9852
GP2DtB	-3.097	0.633	7.61	-0.04	0.9689
GPNIGA	-0.764	0.327	7.32	0.04	0.9609
GPNIGB	-5.658	0.368	11.71	-1.46	0.8537
GZExpoA	-0.957	0.651	5.77	0.4	0.9851
GZExpoB	-1.131	0.628	6.54	0.23	0.985
GZ2DtA	**-0.622**	0.645	5.95	0.37	0.9846
GZ2DtB	-0.841	0.561	6.37	0.27	0.9847
GZNIGA	-0.752	0.341	7.42	0.01	0.9635
GZNIGB	-1.575	0.204	9.1	-0.49	0.9808
IPExpoA	-0.803	0.695	**5.44**	**0.47**	**0.9887**
IPExpoB	-2.172	0.672	6.43	0.26	0.9725
IP2DtA	-0.841	0.659	5.76	0.41	**0.9887**
IP2DtB	-2.26	0.675	6.5	0.24	0.982
IPNIGA	-0.814	**0.713**	**5.29**	**0.5**	00.968
IPNIGB	-1.871	0.675	6.14	0.33	0.9605
IZExpoA	**-0.599**	**0.711**	5.55	0.45	0.9866
IZExpoB	-0.805	0.628	5.96	0.36	0.9836
IZ2DtA	-0.689	0.649	5.68	0.42	0.9861
IZ2DtB	-0.814	0.583	6.2	0.31	0.9833
IZNIGA	-0.721	0.703	**5.36**	**0.49**	0.9796
IZNIGB	-0.806	0.609	6.07	0.34	0.9831

TABLE 4.7: Valeurs des critères pour toutes les combinaisons de facteur en validation (Montargis 99 et Mas Cebria)

Model Name	CRPS	r	RMSE	R^2	AUC
GPExpoA	-1.08	0.73	**7.32**	**0.53**	0.9952
GPExpoB	-1.26	0.482	11.2	-0.1	0.9953
GP2DtA	-1.085	0.649	8.29	0.4	**0.9955**
GP2DtB	-0.937	0.681	7.85	0.46	**0.9954**
GPNIGA	-1.32	0.352	10.49	0.04	0.9530
GPNIGB	-1.346	0.274	10.38	0.06	0.9365
GZExpoA	**-0.93**	0.736	7.77	0.47	0.9947
GZExpoB	**-0.935**	0.731	7.81	0.47	0.9953
GZ2DtA	-1.032	0.653	8.83	0.32	**0.9956**
GZ2DtB	-1.016	0.659	8.18	0.41	0.9952
GZNIGA	-1.318	0.354	10.65	0.01	0.9603
GZNIGB	-1.285	0.293	10.5	0.04	0.9611
IPExpoA	-1.031	**0.76**	**7.14**	**0.55**	0.9917
IPExpoB	-1.317	0.34	10.21	0.09	0.9879
IP2DtA	-1.13	0.684	7.8	0.47	0.9904
IP2DtB	-0.964	0.678	8.31	0.4	0.9913
IPNIGA	-1.121	0.738	**7.45**	**0.51**	0.9935
IPNIGB	-1.358	0.255	10.36	0.06	0.9660
IZExpoA	-0.947	**0.755**	8.14	0.42	0.9907
IZExpoB	**-0.92**	**0.759**	7.63	0.49	0.9872
IZ2DtA	-0.988	0.676	8.31	0.4	0.9892
IZ2DtB	-1.034	0.63	8.54	0.36	0.9875
IZNIGA	-0.949	0.738	7.94	0.45	0.9939
IZNIGB	-1.673	0.391	15.89	-1.21	0.9852

TABLE 4.8: Valeurs des critères pour toutes les combinaisons de facteur en ajustement sur Montargis 99

Model Name	CRPS	r	RMSE	R^2	AUC
GPExpoA	-3.874	0.713	**15.38**	**0.49**	0.9575
GPExpoB	-4.784	0.462	19.7	0.17	0.8817
GP2DtA	-4.441	0.687	17.68	0.33	0.9416
GP2DtB	-4.04	**0.723**	16.34	0.43	0.9483
GPNIGA	-5.255	0.306	21.97	-0.03	0.6610
GPNIGB	-5.382	0.213	21.28	0.03	0.6453
GZExpoA	-3.787	0.688	15.94	0.46	0.9580
GZExpoB	-3.944	0.721	16.21	0.44	0.9581
GZ2DtA	-4.624	0.681	18.98	0.23	0.9500
GZ2DtB	-4.494	0.71	17.61	0.34	0.9435
GZNIGA	-5.379	0.293	22.15	-0.05	0.6737
GZNIGB	-5.214	0.209	21.92	-0.03	0.6826
IPExpoA	-7.008	0.6	22.6	-0.09	0.9393
IPExpoB	-4.592	0.518	18.99	0.23	0.9356
IP2DtA	-3.804	**0.735**	**14.98**	**0.52**	**0.9660**
IP2DtB	**-3.419**	**0.75**	**14.34**	**0.56**	**0.9654**
IPNIGA	-39.78	0.043	141.2	-41.7	0.8883
IPNIGB	-5.585	0.186	26.07	-0.46	0.8432
IZExpoA	-5.677	0.504	26.34	-0.49	0.9304
IZExpoB	-4.675	0.513	26.66	-0.52	0.9282
IZ2DtA	**-3.694**	0.694	15.96	0.45	0.9583
IZ2DtB	**-3.68**	0.683	16.25	0.43	**0.9591**
IZNIGA	-4.084	0.536	18.87	0.24	0.9461
IZNIGB	-4.798	0.506	23.74	-0.21	0.9164

TABLE 4.9: Valeurs des critères pour toutes les combinaisons de facteur en validation sur Montargis 98 et Mas Cebria

Model Name	CRPS	r	RMSE	R^2	AUC
GPExpoA	-1.949	**0.728**	**10.22**	**0.53**	**0.9886**
GPExpoB	-2.410	0.423	16.95	-0.29	0.9863
GP2DtA	-2.004	0.654	11.86	0.37	0.9816
GP2DtB	-1.752	0.668	11.1	0.44	0.9832
GPNIGA	-2.363	0.389	14.74	0.02	0.9397
GPNIGB	-2.426	0.304	14.26	0.08	0.9094
GZExpoA	-1.711	**0.727**	**10.51**	**0.5**	**0.9885**
GZExpoB	**-1.612**	**0.727**	10.53	**0.5**	**0.9885**
GZ2DtA	-1.962	0.653	12.39	0.31	0.9851
GZ2DtB	-1.885	0.644	11.46	0.41	0.9812
GZNIGA	-2.39	0.451	14.86	0.004	0.9628
GZNIGB	-2.245	0.413	14.45	0.057	0.9715
IPExpoA	-1.947	0.721	**10.36**	**0.52**	0.9775
IPExpoB	-2.372	0.352	14.23	0.09	0.981
IP2DtA	-2.024	0.693	10.74	0.48	0.9816
IP2DtB	**-1.679**	0.697	11.13	0.44	0.9826
IPNIGA	-2.147	0.69	10.78	0.48	0.9831
IPNIGB	-2.463	0.278	14.33	0.08	0.9495
IZExpoA	-1.743	0.718	11.21	0.43	0.9749
IZExpoB	**-1.656**	0.704	10.75	0.48	0.9739
IZ2DtA	-1.775	0.685	11.19	0.44	0.9811
IZ2DtB	-1.77	0.66	11.55	0.4	0.9797
IZNIGA	-1.854	0.685	11	0.45	0.9839
IZNIGB	-2.749	0.489	20.05	-0.81	0.9807

TABLE 4.10: Valeurs des critères pour toutes les combinaisons de facteur en ajustement sur Montargis 98 et Montargis 99)

Model Name	CRPS	r	RMSE	R^2	AUC
GPExpoA	**-6.9**	0.736	**20.02**	**0.5**	**0.9282**
GPExpoB	-10.115	0.041	30.57	-0.17	0.5004
GP2DtA	-8.807	0.721	23.87	0.29	0.919
GP2DtB	-7.976	0.712	20.72	0.46	0.9216
GPNIGA	-10.758	0.031	30.24	-0.14	0.599
GPNIGB	-10.763	0.038	30.12	-0.13	0.5303
GZExpoA	**-6.806**	**0.741**	**20.49**	**0.48**	**0.9246**
GZExpoB	**-7.408**	**0.738**	**20.58**	**0.47**	**0.9254**
GZ2DtA	-8.938	0.718	25.1	0.21	0.9203
GZ2DtB	-8.623	0.71	21.74	0.41	0.9244
GZNIGA	-10.759	0.036	30.25	-0.14	0.6681
GZNIGB	-10.672	0.054	30.16	-0.14	0.7915
IPExpoA	-29.998	0.661	52.3	-2.42	0.8856
IPExpoB	-9.473	0.57	24.15	0.27	0.8958
IP2DtA	-21.271	**0.747**	42.07	-1.21	0.8835
IP2DtB	-9.206	0.689	39.68	-0.97	0.8848
IPNIGA	-200.97	-0.113	325.08	-131.12	0.5331
IPNIGB	-10.178	0.217	27.75	0.04	0.6668
IZExpoA	-23.096	0.653	68.25	-4.82	0.9184
IZExpoB	-16.072	0.613	70.96	-5.3	0.9123
IZ2DtA	-16.906	0.718	47.06	-1.77	0.9217
IZ2DtB	-14.298	0.619	64.84	-4.26	0.9183
IZNIGA	-11.061	0.541	42.28	-1.24	0.8594
IZNIGB	-10.341	0.62	46.83	-1.74	0.9104

TABLE 4.11: Valeurs des critères pour toutes les combinaisons de facteur en validation sur Mas Cebria

Chapitre 5

Discussion - Perspectives

Table des matières

1 Apports et difficultés des méthodes statistiques 128
2 Influence des choix de modélisation sur la qualité de prédiction 129
3 Adaptation des modèles aux situations de prédiction 131
4 Mise à disposition et applications 132
5 Perspectives 134

1 Apports et difficultés des méthodes statistiques

Dans cette thèse, nous avons adapté les modèles de dispersion utilisés dans le contexte de la coexistence, au cadre bayésien. Les méthodes Bayésiennes permettent d'intégrer la variabilité des observations et l'information existante sur les valeurs des paramètres au processus d'estimation. La forme prise par les prédictions qui découlent de l'estimation (i.e. une distribution de probabilité) donne directement de l'information sur l'incertitude associée à la réalisation d'un évènement. La quantification des incertitudes pour toutes fonctions des paramètres et la possibilité d'intégrer de l'information a priori lorsqu'elle existe, constituent les deux principaux apports des méthodes proposées au contexte de la coexistence.

L'adaptation des modèles de dispersion au cadre bayésien implique de résoudre des difficultés techniques. En particulier elle a nécessité, lorsque les modèles sont basés sur le mode de représentation individuel, d'accélérer les temps de calcul par la recherche d'une approximation (voir dernière section du chapitre 3) satisfaisante. L'approximation définie et retenue, sur la base d'une étude par simulation, permet de réaliser l'estimation des paramètres dans des temps qualifiés de raisonnables (entre un et trois jours sur des serveurs dédiés au calcul). Elle montre donc la faisabilité de l'ajustement des modèles de dispersion basés sur le mode de représentation individuel dans un cadre Bayésien, ainsi que de leur utilisation dans un contexte d'aide à la décision. Cependant, rappelons que la méthode élaborée ici pour diminuer les temps de calcul est une approximation et, de ce fait, qu'elle s'accompagne également d'une dégradation des performances du modèle, aussi bien sur la capacité d'ajustement que sur la capacité prédictive. Nous reviendrons sur l'influence de cette approximation dans la section suivante. Néanmoins, nous mentionnons ces éléments ici car ils font partie des difficultés techniques rencontrées (ici les temps de calculs) dans l'utilisation des méthodes statistiques.

D'autres difficultés ont été rencontrées et nuancent pour le moment les apports des méthodes proposées. En premier lieu, il s'agit des problèmes de convergence des algorithmes rencontrés dans les cas d'estimation sur des données *multi-sources*. Dans ces cas là, aucune des stratégies déployées n'a permis d'obtenir la convergence des algorithmes d'estimation. Or cette convergence est requise pour faire de l'inférence statistique et donc pour faire des prédictions. Cette absence de convergence, au moins pour une partie des cas d'estimations, constitue donc un des points de blocage de ce travail.

D'autre part, une analyse approfondie des résultats issus des estimations dans les cas *mono-source* montre des comportements très différents entre les estimations de paramètres selon le modèle d'observation. En effet, si le modèle d'observation est le modèle de Poisson à espérance aléatoire ($y_s|K, \mu_s^{'} \sim \mathcal{P}(K\mu_s^{'})$ et $\mu_s^{'} \sim \mathcal{N}(\mu_s, \sigma^2)$, on observe que μ_s tend vers zéro et σ^2 prend des valeurs anormalement élevées. Ce comportement plaide pour une plus grande attention à apporter à la définition du modèle d'observation et ce en cohérence avec la structure des données. Il peut en effet s'interpréter comme une confusion d'effets (i.e. entre les effets représentés par μ et σ^2). Plus précisément il peut s'expliquer par l'inadéquation des contraintes entre espérance et variance associées au modèle d'observation de Poisson et par les caractéristiques des données de comptage de grains (principalement grande variabilité et surabondance des zéros) utilisées dans cette partie. Cette explication est confirmée par les bonnes perfomances du modèle ZIP, dans

lequel la variabilité due à l'abondance des zéros est gérée séparément du reste de la variabilité des données.

2 Influence des choix de modélisation sur la qualité de prédiction

Nous revenons dans cette section sur l'influence des différents choix de modélisation qui ont été effectués, sur la qualité prédictive des modèles. En premier lieu, on observe de grandes différences entre les qualités de prédiction obtenues avec les deux modes de représentation des échanges de pollen entre parcelles, i.e. Global et Individuel (voir figures 4.11 et 4.12 du chapitre 4). Les grandes tendances dans ces différences et leur analyse montrent qu'il existe une relation stable entre le mode utilisé pour la représentation des échanges de pollen entre parcelles et la *sur-* ou *sous-* estimation des observations dans les situations indépendantes des situations de calibration : sur-estimation avec le mode individuel et sous-estimation avec le mode global. Cette relation peut s'expliquer par le fait que le mode global associe la surface émettrice à un point alors que le mode individuel considère chaque point source de pollen individuellement. En effet, les parcelles émettrices de pollen OGM dans le jeu d'entrainement ont des tailles similaires et très inférieures (environ 80 fois) à la taille de la parcelle émettrice du jeu de validation. Il serait intéressant de vérifier cette hypothèse en inversant jeu d'entrainement et jeu de validation. Malheureusement, cette comparaison n'a pu être conduite jusqu'à présent.

D'autre part, une estimation fiable des paramètres requiert un nombre de données suffisament grand. Hors les jeux de données présentant un nombre de données suffisant pour la calibration ont le plus souvent des caractéristiques très spécifiques (i.e. mono-source, en conditions contrôlées, échantillonnage très dense et surface émettrice relativement petite). Les situations de prédiction qui intéressent les preneurs de décision ont des caractéristiques très différentes de celles-ci (multi-sources, en conditions réelles, échantillonnage peu dense et surface émettrice relativement grande). De ce fait, il existe un grand décalage entre les situations dans lesquelles les modèles sont calibrés et les situations pour lesquelles il faut fournir des prédictions. De plus, les résultats de l'analyse en composantes principales (dernière section du chapitre 4) et l'examen des tableaux de critères montrent que les qualités d'un modèle peuvent être très différentes entre la qualité d'ajustement (estimée sur les données d'ajustement) et la qualité de prédiction (estimé sur des données indépendantes des données d'ajustement). Ces constatations plaident pour une réflexion approfondie et en amont de toutes procédures de calibration sur la représentativité des jeux de données utilisés pour l'ajustement ainsi que sur les caractéristiques nécessaires et suffisantes que ces jeux doivent avoir pour rapprocher leurs caractéristiques de celles des situations *réelles* de prédictions.

Cette réflexion doit intégrer la notion de taille et de composition du jeu de données idéal et ce à plusieurs niveaux. Dans ce travail, nous avons commencé par nous concentrer sur les quelques jeux de données les mieux renseignés afin de rechercher une modélisation reproduisant de manière satisfaisante l'espérance et la variance de la réponse en fonction de la distance et d'autres caractéristiques telles que l'orientation et la force du vent. De cette façon, nous voulions nous prémunir contre les artefacts pouvant survenir dans

les estimations lorsque le nombre de données est trop faible et/ou lorsque les données exhibent une trop grande variabilité. L'idée initiale était d'intégrer les autres jeux de données dans un second temps, après avoir appréhendé les facteurs permettant une bonne retranscription de la réponse moyenne et de sa variabilité. Or, d'une part, cette phase de mise au point a pris beaucoup plus de temps que prévu et, d'autre part, la dépendance qu'on observe entre la qualité de prédiction et les caractéristiques des jeux de données utilisées dans l'estimation nous poussent à revoir cette stratégie et à privilégier des jeux de données exhibant des situations plus contrastées (tailles et formes des parcelles émettrices, direction et force du vent, période de floraison, ...), et ce en amont du processus de modélisation à proprement parler. Il est donc important de considérer dès le départ un nombre de situations différentes suffisament grand pour consolider un jeu de données d'entrainement. Notons que ces considérations sont connues et qu'un des objectifs de ce travail était de permettre l'intégration de jeux de données très diversifiés, notamment ceux issues du projet SIGMEA. Précisons donc que l'intégration de la totalité de jeux de données disponibles et la prise en compte de la diversité des situations possibles restent des défis à relever dans le domaine de la modélisation des flux de gènes.

Les résultats issus de l'approximation du mode de représentation individuel méritent quelques approfondissements. Dans la section consacrée à cette approximation, nous avons cherché le meilleur compromis entre le nombre de points dans la grille et la qualité de prédiction pour une situation donnée (Montargis 98), avec une fonction de dispersion donnée et un jeu de paramètres fixé. La grille retenue pour l'approximation a ensuite été appliquée aux deux autres situations (Montargis 99 et Mas Cebria), avec d'autres fonctions de dispersion et avec des valeurs de paramètres différentes. Or, la robustesse de l'approximation *i)* à la fonction de dispersion *ii)* aux valeurs de paramètres *iii)* aux caractéristiques de la situation (taille et forme des champs émetteurs et récepteurs, densité du semis, part d'allocation OGM dans le paysage, ...), n'a pas été testée et constitue donc un facteur d'incertitudes supplémentaire sur les origines des erreurs de prédiction des modèles. De plus, la recherche d'une approximation robuste aux caractéristiques d'une situation de prédiction peut permettre de donner des pistes pour la définition de grilles différentes et optimisées pour chaque type de situation. La définition de procédures pour la conception de grilles adaptées aux particularités de chaque situation constitue donc une des perspectives de ce travail. On rappelle ici que, malgré la lourdeur des calculs qu'il impose, l'intérêt du mode de représentation individuel est de permettre une extrapolation relativement aisée des modèles calibrés à des situations très différentes des situations de calibration.

Enfin, un des choix implicites de ce travail était de ne pas recourir à des modèles dynamiques et ce pour deux raisons principales : d'une part, les problèmes de disponibilité de données météorologiques au pas de temps journalier et, d'autre part, les temps de calcul associés aux prédictions qui, dans un contexte d'aide à la décision, doivent être aussi faibles que possible. Ce choix nous oblige à prendre en compte des effets résultant de processus dynamiques (force et direction du vent, décalage de floraison) dans un modèle non dynamique. Il implique donc une diminution de la capacité prédictive des modèles par de grandes simplifications, notamment la prise en compte de la direction et force du vent uniquement pour le vent dominant durant la période de floraison. Dans ce travail nous n'avons pas testé de version dynamique des modèles proposés. Une piste intéressante serait de tester de telles versions en reprenant les équation définies dans Angevin et al.

(2008), pour mieux intégrer les effets de variables dynamiques sur l'établissement du taux de pollinisation croisée.

3 Adaptation des modèles aux situations de prédiction

Dans cette section, nous discutons des capacités d'adaptation des modèles élaborés dans cette thèse aux différentes situations de prédiction. On rappelle qu'un des objectifs de la thèse était précisément d'élaborer des modèles permettant l'adaptation au niveau d'information disponible pour toute situation de prédiction et cas d'utilisation potentiel. On considère ici deux types d'adaptation : d'une part, l'adaptation des modèles au niveau d'information disponible dans une situation donnée. Et, d'autre part, leur adaptation aux objectifs finaux de la prédiction.

La première composante de l'adaptation a été discutée dans la première section du chapitre 2 ("Conséquences opérationnelles"). On rappelle qu'on propose deux grands ensembles de solutions : un premier ensemble basé sur l'utilisation de modèles différents pour des niveaux d'information différents, et un second ensemble basé sur un même modèle dont les hypothèses sur les entrées varient en fonction du niveau d'information disponible. Ici, le choix de l'une ou l'autre de ces solutions est laissé libre à l'utilisateur. En effet, ce choix dépend de plusieurs facteurs, notamment la capacité de l'utilisateur à faire des hypothèses sur les données d'entrées, difficile à estimer *a priori*.

En ce qui concerne la deuxième composante de l'adaptation, la multiplication des critères statistiques pour la sélection a posé problème, comme on l'a vu plus haut, quant au choix de modèles. Cependant, le choix qui a été fait de ne pas se focaliser sur un seul critère pour effectuer la sélection offre un avantage non négligeable ; il permet d'intégrer la notion d'objectifs et situations de prédiction à la sélection de modèles. On a distingué dans ce travail trois familles d'objectifs :

1. prédire, le plus précisément possible, la valeur moyenne sur une surface donnée (parcelle ou ensemble de parcelles) ;
2. établir un classement de parcelles sur la base du taux de pollinisation croisée. Ici la qualité de prédiction importe peu, ce qui compte c'est que le classement prédit soit le plus proche possible du classsement observé ;
3. définir un plan d'échantillonnage optimal pour l'estimation du taux de pollinisation croisée à partir de prélèvements sur le terrain.

Dans le premier cas, on cherche à prédire précisément le taux de pollinisation croisée au sein d'une récolte (ou ensemble de récoltes). Il faut donc un modèle qui minimise l'écart entre les prédictions et les observations. Les critères faisant intervenir cet écart dans leur calcul sont la RMSE et le CRPS ; c'est donc sur la base de ces deux critères que la sélection doit être faite dans ce cas là. De plus, l'analyse en composantes principales a montré que ces deux critères étaient très corrélés et apportaient donc quasiment la même information. Cette observation est confirmée par l'examen du tableau 4.11 de la section 3 du chapitre 4 (tableau des critères calculés pour tous les modèles en validation

avec la troisième stratégie d'estimation) ; les modèles ayant les meilleurs RMSE ont aussi les meilleurs CRPS et inversement.

Pour le deuxième cas, comme indiqué plus haut, la valeur du taux de pollinisation croisée prédite par le modèle importe peu, c'est la qualité du classement qu'il faut maximiser. Dans ce cas, c'est le critère AUC qui est le plus pertinent. En effet, la valeur de ce critère peut s'interpréter comme le pourcentage de classement correct vis-à vis d'un seuil donné et son calcul n'intègre pas de notion d'écart entre l'observation et la prédiction. On peut donc avoir des modèles très peu précis qui classent très bien et inversement. Ce n'est pas vraiment le cas ici (cf tableau 4.11), les modèles qui prédisent précisément (faible RMSE) ont aussi tendance à bien classer (AUC élevé), et les modèles moins précis ont, en général, une faible qualité de classement.

Le dernier cas constitue une application originale d'un modèle de prédiction des taux de pollinisation croisée que nous précisons dans la section suivante. Cette application permet, à l'aide des prédictions du modèle, de déterminer un plan ou une stratégie d'échantillonnage optimal. Dans ce cas particulier, la précision du taux prédit n'est pas importante, les quantités à maximiser étant d'une part, la corrélation entre les prédictions et les observations et, d'autre part, en faisant des strates dans les observations la corrélation entre les écart-types des observations et celui des prédictions intra strates.

De plus, l'analyse de sensibilité des critères aux facteurs de définition des modèles via l'analyse en composante principale et l'ANOVA associée pourra permettre, après approfondissements, de déterminer les régions, du plan défini par les deux premières composantes, dans lesquelles se trouvent les modèles les plus performant pour une fonction (i.e. Prédire précisément, classer ou échantilloner) particulière. Cette détermination permettra de sélectionner un modèle ou groupe de modèles directement en fonction de l'objectif de son utilisation.

4 Mise à disposition et applications

D'un point de vue opérationnel, ce travail de thèse a permis la mise à disposition des résultats obtenus pour deux grands types d'applications :

1. L'aide à la décision pour le choix d'un itinéraire technique (assolement, variétés, date de semis) minimisant le risque de dépasser (pour l'agriculteur ne cultivant pas d'OGM) ou de faire dépasser (pour l'agriculteur cultivant des OGM) le seuil légal de pollinisation croisée.

2. La définition de stratégies d'échantillonage, assistée par les modèles, pour l'estimation du taux de pollinisation croisée à partir de prélèvements sur le terrain.

Pour la première application, les modèles définis aux chapitre 2, permettant l'adaptation au niveau d'information disponible, ont tous été estimés et déployés sur une plateforme informatique. L'outil résultant de ce déploiement permet de

- définir une configuration spatiale de départ en dessinant des parcelles ou en rentrant leurs coordonnées géographiques via un SIG ;

- renseigner les données auxquelles on a accès (direction du vent, date de floraison) ;
- choisir le modèle le plus pertinent vis-à-vis du niveau d'information dont on dispose ;
- faire tourner le ou les modèles ;
- visualiser et sauvegarder les résultats de simulations.

Cet outil d'aide à la décision est disponible à l'adresse suivante :

http ://www.price.preprod.farmsat.com

La définition de situations de prédiction ainsi que l'implémentation et la mise à disposition de l'outil d'aide à la décision ont fait l'objet d'une communication à la 6ème conférence internationale sur la coexistence entre culture OGM et non OGM (GMCC 2013) (Meillet et al., 2013) dont les actes figurent en annexes (annexe 5). De surcroît, un article décrivant plus précisément les composantes informatiques de l'outil d'aide à la décision est en voie de soumission à *Ecological Informatics* pour un numéro spécial *"Information and Decision Support Systems for Agriculture and Environment"*

La deuxième application concerne la définition de stratégies d'échantillonnage assistées par les modèles. L'originalité principale de ce travail repose sur l'utilisation de modèles, non pas pour prédire la valeur moyenne du taux de pollinisation croisée sur une surface donnée (parcelle ou ensemble de parcelles), mais pour déterminer une stratégie d'échantillonnage optimal.

Cette détermination peut être réalisée via deux familles de méthodes selon la configuration :

- Avec un plan d'échantillonnage défini et effectué au préalable, la méthode proposée sert à définir différentes pondérations permettant, en pondérant les échantillons déjà prélevés, de rapprocher la valeur estimée à partir des échantillons de la vraie valeur.
- Sans plan d'échantillonnage défini au préalable, la méthode permet d'utiliser les prédictions du modèle de manière à identifier des zones de variabilité comparable afin de determiner des strates et procéder à un échantillonage stratifié.

Ce travail à fait l'objet de plusieurs valorisations :

1. Des actes à la conférence 6ème Conférence International sur la coexistence entre cultures génétiquement modifiées (GM) et non-GM, *GMCC 2013* :
   Bancal, R., Makowski, M., and Bensadoun, A. Comparison of sampling strategies to evaluate rate of transgenic adventitious presence in agricultural fields. In *GMCC 2013 - 6th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains* (Lisbon, France, November 2013)

2. Un article dans *AgBioForum* correspondant à une extension des actes présentés lors de la conférence *GMCC 2013* :
   Bancal, R. Makowski, M., Bensadoun, A., Monod, H., and Messéan, A. Comparison of Sampling Strategies to Evaluate Rate of Transgenic Adventitious Presence in Agricultural Fields. *AgBioForum 17(2)* (2014), 166–171. [html]

3. Un article soumis à *Risk Analysis* :
   Makowski, M., Bancal, R., Bensadoun, A., Monod, H., and Messéan, A. Sampling strategies to evaluate rate of transgene adventitious presence in non-genetically modified crop fields. Soumis à *Risk Analysis* en Décembre 2014.

Notons que la mise à dispostion d'outils finalisés et le travail sur les applications ont été menés de front avec la partie plus méthodologique de ce travail. Il semble que ce soit un atout pour un travail de thèse dans un contexte où la pluridisciplinarité est de plus en plus recherchée. Néanmoins, ayant fait l'expérience, je considère à présent qu'il aurait été préférable de se concentrer uniquement sur un de ces aspects (méthode ou application) afin d'acquérir une expertise et des compétences plus solides soit sur la méthodologie soit sur les application.

5 Perspectives

Les problèmes rencontrés pour l'estimation des modèles dans les contextes *multi-sources*, même s'ils ont gêné l'atteinte des objectifs initialement fixés, nous permettent de tirer des leçons intéressantes et d'entrevoir des pistes pour les prises en compte de ces contextes dans la modélisation des taux de pollinisation croisée.

Le premier point concerne la quantité et la qualité des données. En effet, les modèles calibrés dans les contextes *mono-source* l'ont été avec un très grand nombre de données (de 3000 à 4500 plantes échantillonnées sur environ 1.5 hectares) obtenues en conditions contrôlées. Or, cette quantité de données n'est pas atteinte dans un contexte *multi-sources* et ce pour deux principales raisons :

1. Ces données sont issues d'échantillonnages dans des champs pour la production agricole ; on ne peut donc pas multiplier les prélèvements à l'infini.
2. La PCR est une technique relativement coûteuse. On ne peut donc pas espérer qu'il y ait plus de quelques dizaines d'échantillons PCR dans une parcelle.

D'une part, la quantité de données est insuffisante pour calibrer les mêmes modèles de la même façon qu'en *mono-source* et nous venons de voir pourquoi cette quantité n'est pas susceptible d'augmenter significativement à l'avenir. D'autre part, Les données *multi-sources* étant récoltées en condition réelles, elles présentent une variabilité supplémentaire non prise en compte par les modèles dont les origines peuvent êtres multiples et ne constituent pas d'intérêt réel pour la modélisation.

Ces raisons nous obligent à revoir notre approche et à dessiner des perspectives pour les programmes de recherche à venir. On envisage à ce stade deux types de perspectives. En premier lieu, il s'agit d'approfondir les causes de non-convergence des algorithmes d'estimations sur les données *multi-sources*. Les causes possible peuvent êtres *i)* la quantité insuffisante de données, *ii)* la trop grande complexité du processus en condition réelles, *iii)* la conjonction des deux raisons précédentes. On propose donc deux solutions, une basée sur la diminition *virtuelle* de la variabilité initiale des données et une basée sur l'augmentation *virtuelle* du nombre de données :

1. Pour diminuer *virtuellement* la variabilité initiale des données : Utiliser un modèle complexe du type MAPOD (Angevin et al., 2008) pour simuler les situations *multi-sources* et tester la calibration de nos modèles sur ces données simulées.

2. Pour augmenter *virtuellement* le nombre de données : Utiliser des méthodes de géostatistique telles que le krigeage pour interpoler les observations ponctuelles, limitées au sein d'une parcelle, à l'ensemble d'une surface (parcelle ou ensemble de parcelles).

Ces deux possibilités doivent permettre de déterminer la ou les causes de non convergence des algorithmes et donc de donner des directives pour les expérimentations et campagnes de récoltes de données futures. Il faut garder à l'esprit qu'un des inconvénients de ces deux pistes est l'ajout d'un facteur d'erreur ou d'approximation soit par le modèle utilisé pour simuler les données en amont soit par la méthode d'interpolation.

Une autre piste, très différente des deux premières évoquées, repose sur la recherche et l'utilisation d'autres formalismes intéressants pour prendre en compte la dispersion. On pense ici, notamment et sans caractère exclusif, aux automates cellulaires (Von Neumann, 1966; Gardner, 1970) et à leurs applications récentes aux problèmes de dispersion (Rasmussen and Hamilton, 2012). Les automates cellulaires (CA) sont des systèmes qui se composent d'un réseau de cellules et d'un processus qui applique un ensemble de règles pour dériver périodiquement les caractéristiques de chaque cellule, à partir des caractéristiques des cellules voisines. Depuis Von Neumann (1966), les CA ont émergé comme des systèmes très utiles pour la modélisation de phénomènes environnementaux complexes et dynamiques. Typiquement, les CA sont utilisés pour faire des prédictions pour divers problèmes environnementaux, tels que la modélisation d'invasion d'espèces (Slimi et al., 2009), la modélisation de feu de forêt (Innocenti et al., 2009) et la gestion des aquifères (Ravazzani et al., 2011).

Références

I. Albert, E. Grenier, J. B. Denis, and Rousseau J. Quantitative risk assessment from farm to fork and beyond : a global bayesian approach concerning food-borne diseases. *Risk Analysis*, 28(2) :557–571, 2008.

T.R. Allnutt, M. Dwyer, J. McMillan, C. Henry, and S. Langrell. Sampling and modeling for the quantification of adventitious genetically modified presence in maize. *Journal of Agricultural and Food Chemistry*, 56 :3232–3237, 2008.

A.M. Aly and A.R. Hassan. Foraging preference of honey bees for pollen of some maize varieties in middle Egypt. *Shashpa*, 125–136, 1999.

F. Angevin, N. Colbach, J.M. Meynard, and C. Roturier. Analysis of necessary adjustements of farming practices. In EUR 20394 EN, editor, *Scenarios for Co-existence of Genetically Modified, Conventional and Organic Crops in European Agriculture*, pages 52–66. Technical Report Series of the Joint Research Center of the European Commision, 2002.

F. Angevin, E. K. Klein, C. Choimet, A. Gauffreteau, C. Lavigne, A. Messéan, and J. M. Meynard. Modelling impacts of cropping systems and climate on maize cross-pollination in agricultural landscapes : The MAPOD model. *European Journal of Agronomy*, 28 (3) :471–484, 2008.

E. Anklam, F. Gadani, P. Heinze, H. Pijnenburg, and G. Van Den Eede. Analytical methods for detection and determination of genetically modified organisms in agricultural crops and plant-derived food products. *European Food Research and Technology*, 214 : 3–26, 2002.

R. Arritt, C. Clark, A. S. Goggi, H. Lopez-Sanchez, and J. Westgate, M.and Riese. Lagrangian numerical simulations of canopy air flow effects on maize pollen dispersal. *Field Crops Research*, 102 :151–162, 2007a.

R.W. Arritt, C.A. Clark, A.S. Goggi, H. Lopez-Sanchez, M.E. Westgate, and J.M. Riese. Lagrangian numerical simulations of canopy air flow effects on maize pollen dispersal. *Field Crops Research*, 102(2) :151 – 162, 2007b. ISSN 0378-4290. doi : http://dx.doi.org/10.1016/j.fcr.2007.03.008. URL http://www.sciencedirect.com/science/article/pii/S0378429007000378.

J. P. Astini, A. Fonseca, C. Clark, J. Lizaso, L. Grass, M. Westgate, and R. Arritt. Predicting outcrossing in maize hybrid seed production. *Agronomy Journal*, 101 :373–380, 2009.

Y. F. Atchadé and J. S. Rosenthal. On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5) :815–828, 10 2005. doi : 10.3150/bj/1130077595. URL http://dx.doi.org/10.3150/bj/1130077595.

D.E. Aylor, N.P. Schultes, and E.J. Shields. An aerobiological framework for assessing cross-pollination in maize. *Agricultural and Forest Meteorology*, 119(3–4) :111 – 129, 2003. ISSN 0168-1923. doi : http://dx.doi.org/10.1016/S0168-1923(03)00159-X. URL http://www.sciencedirect.com/science/article/pii/S016819230300159X.

R. Bancal, D. Makowski, and A. Bensadoun. Comparison of sampling strategies to evaluate the rate of transgenic adventitious presence in agricultural fields. In *GMCC 2013 - 6th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains*, Lisbon, Portugal, November 2013.

M. Bannert and P. Stamp. Cross-pollination of maize at long distance. *European Journal of Agronomy*, 27 :44–51, 2007.

O.E. Barndorff-Nielsen. Sand, wind and statistics : Some recent investigations. *Acta Mechanica*, 64(1-2) :1–18, 1986. doi : 10.1007/BF01180094. URL http://dx.doi.org/10.1007/BF01180094.

O.E. Barndorff-Nielsen. Normal inverse gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics*, 24(1) :1–13, 1997. doi : 10.1111/1467-9469.00045. URL http://dx.doi.org/10.1111/1467-9469.00045.

A.J. Bateman. Contamination in seed crops I : Insect pollination. *Journal of Genetics*, 48 :257–275, 1947a.

A.J. Bateman. Contamination in seed crops II : Wind pollination. *Heredity*, 1 :235–246, 1947b.

A.J. Bateman. Contamination in seed crops III : Relation with isolation distances. *Heredity*, 1 :303–336, 1947c.

H. J. Beckie and L. M. Hall. Simple to complex : Modelling crop pollen-mediated gene flow. *Plant Science*, 175(5) :615–628, 2008.

A. Bensadoun, H. Monod, F. Angevin, D. Makowski, and A. Messéan. Modeling of gene flow by a bayesian approach : A new perspective for decision support. In *GMCC 2013 - 6th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains*, Lisbon, Portugal, November 2013.

J. Besag, J. York, and A. Mollié. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1) :1–59, 1991.

J.J. Boreux, E. Parent, and J Bernier. *Pratique du calcul bayésien*. Springer, 2009.

A. Bouvier, K. Kiêu, K. Adamczyk, and H. Monod. Computation of the integrated flow of particles between polygons. *Environmental Modelling & Software*, 24 :843–849, 2009.

Y. Brunet, S. Dupont, M. De Luca, J.P. Pinty, J. Escobar, S. Delage, P. Tulet, C. Lac, and X. Foueillassar. Mesoscale dispersal of pollen and implications for gene flow. In *GMCC 2009 - 4th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains*, Melbourne, Australia, November 2009.

P.F. Byrne and S. Fromherz. Can GM and non-GM crops coexist ? Setting a precedent in Boulder County, Colorado, USA. *Food Agriculture and Environment*, 1 :158–261, 2003.

B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis*. Texts in Statistical Science. Chapman and Hall/CRC, Third edition, 2009.

M.S. Castellazzi, J. Matthews, F. Angevin, C. Sausse, G.A. Wood, P.J. Burgess, I. Brown, K.F. Conrad, and J.N. Perry. Simulation scenarios of spatio-temporal arrangement of crops at the landscape scale. *Environmental Modelling & Software*, 25(12) :1881 – 1889, 2010. ISSN 1364-8152. doi : http://dx.doi.org/10.1016/j.envsoft.2010.04.006. URL http://www.sciencedirect.com/science/article/pii/S136481521000099X.

CE, 2001a. Directive 2001/18/CE du Parlement Européen et du Conseil du 12 mars 2001 relative à la dissémination volontaire d'organismes génétiquement modifiés dans l'environnement et abrogeant la directive 90/220/CEE du Conseil. *Journal Officiel des Communautés européennes*, 17/04/2001, vol. 44, L106, 2001/18/CE, pp 1-38.

CE, 2003a. Commission recommandations of the 23 July 2003 on guidelines for the development of national strategies and best practices to ensure the co-existence of genetically modified crops with conventional and organic farming, 2003/556/EC (notified under document number C(2003) 2624) . *Official Journal of the European Union*, 29/07/2003, vol. 46, L189, pp 36-47.

CE, 2003b. Regulation (EC) No 1829/2003 of the European Parliament and the Concil of 22 September 2003 on genetically food and feed. *Official Journal of the European Union*, 18/10/2003, vol. 46, L268, pp 1-23.

CE, 2003c. Regulation (EC) No 1830/2003 of the European Parliament and the Concil of 22 September 2003 concerning the traceability and labelling of genetically modified organisms and the traceability of food and feed products produced from genetically modified organisms and amending Directive 2001/18/EC. *Official Journal of the European Union*, 18/10/2003, vol. 46, L268, pp 24-28.

CE, 2004. Commission recommandations of the 4 October 2004 on technical guidance for sampling and detection of genetically modified organisms and material produced from genetically modified organisms as or in products in the context of Regulation (EC) No 1830/2003. *Official Journal of the European Union*, 24/11/2003, vol. 47, L348, pp 18-26.

CE, 2007a. Décision de la Commision du 25 avril 2007 concernant le retrait du marché du maïs Bt176 (SYN-EV176-9) et des produits qui en sont dérivés (notifiée sous le numéro c(2007) 1804). *Journal Officiel des Communautés européennes*, 05/05/2007, vol. 50, L117, pp 14-16.

CE, 2007b. Règlement (CE) No 834/2007 du Conseil du 28 juin 2007 relatif à la production biologique et à l'étiquetage des produits biologiques et abrogeant le règlement (CEE) No 2092/91. *Journal Officiel des Communautés européennes*, 20/07/2007, vol. 50, L189, pp 1-23.

CE, 2009. Report from the Commision to the council and the European parlaiment on the coexistence of GM crops with conventional and organic farming. *European Commision*, SEK(2009) 408, 12p + appendix.

CE, 2011a. Commission recommandations (EU) No 619/2011 of the 24 June 2011 laying down the methods of sampling and analysis for the official control of feed as regards presence of genetically modified material for which an authorisation procedure is pending or the authorisation of which has expired. *Official Journal of the European Union*, 25/06/2011, L166, pp 9-15.

J. S. Clark. Why trees migrate so fast : confronting theory with dispersal biology and the paleorecord. *The American Naturalist*, 152 :204–224, 1998.

J. S. Clark. Why environmental scientists are becoming bayesians. *Ecology Letters*, 8 : 2–14, 2005.

J. S. Clark, M. Silman, R. Kern, E. Macklin, and J. HilleRisLambers. Seed dispersal near and far : patterns across temperate and tropical forests. *Ecology*, 80 :1475–1494, 1999.

W.G. Cochran. *Sampling techniques*. John Wiley & Sons, third edition, 1977.

N. Colbach. Modelling cropping system effects on crop pest dynamics : How to compromise between process analysis and decision aid. *Plant Science*, 179 :1–13, 2010. doi : 10.1016/j.plantsci.2010.04.009.

N. Colbach and Meynard J.M. Clermont-Dauphin, C. and. GeneSys : a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers I. temporal evolution of a population of rapeseed volunteers in a field. *Agriculture, Ecosystems and Environment*, 83 :235–253, 2001.

N. Colbach, C. Clermont-Dauphin, and J.M. Meynard. GeneSys : a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers II. genetic exchanges among volunteer and cropped populations in a small region. *Agriculture, Ecosystems and Environment*, 83 :255–270, 2001.

RB. Cunningham and DB. Lindenmayer. Modeling count data of rare species : Some statistical issues. *ECOLOGY*, 86(5) :1135–1142, 2005. ISSN 0012-9658. doi : {10.1890/04-0589}.

C. Damgaard and G. Kjellson. Gene flow of oilseed rape (*Brassica napus*) according to isolation distance and buffer zone. *Agriculture, Ecosystems and Environment*, 108 : 291–301, 2005.

M. Debeljak, A. Trajanov, F. Stojanova, D. Leprince, and S. Dzeroski. Using relational decision trees to model out-crossing rates in a multi-field setting. *Ecological Modelling*, 245 :75–83, 2012.

Y. Devos, D. Reheul, and A. De Schrijver. The co-existence between transgenic and non-transgenic maize in the european union : a focus on pollen flow and cross-fertilization. *Environmental and Biosafety Research*, 2005.

Y. Devos, K. Dillen, and M. Demont. How can flexibility be integrated into coexistence regulations ? a review. *Journal of the Science of Food and Agriculture*, 94(3) :381–387, 2014. ISSN 1097-0010. doi : 10.1002/jsfa.6358. URL http://dx.doi.org/10.1002/jsfa.6358.

D. Dietiker, P. Stamp, and W. Eugster. Predicting seed admixture in maize combining flowering characteristics and a lagrangian stochastic dispersion model. *Field Crops Research*, 121 :256–267, 2011. doi : 10.1016/j.fcr.2010.12.009.

S. Dupont and Y. Brunet. Simulation of turbulent flow in an urban forested park damaged by a windstorm. *Boundary-Layer Meteorology*, 120(1) :133–161, 2006. doi : 10.1007/s10546-006-9049-5. URL http://dx.doi.org/10.1007/s10546-006-9049-5.

S. Dupont, Y. Brunet, and N. Jarosz. Eulerian modelling of pollen dispersal over heterogeneous vegetation canopies. *Agricultural and Forest Meteorology*, 141(2–4) :82 – 104, 2006. ISSN 0168-1923. doi : http://dx.doi.org/10.1016/j.agrformet.2006.09.004. URL http://www.sciencedirect.com/science/article/pii/S0168192306002449.

J. Emberlin, B. Adams-Groom, and J. Tidmarsh. A report on the dispersal of maize pollen. University college of Worceter commisioned by the Soil Association. 22p., 1999.

H. Foudhil, Y. Brunet, and J. Caltagirone. A fine-scale k-epsilon model for atmospheric flow over heterogeneous landscapes. *Environmental Fluid Mechanics*, 5 :247 – 265, 2005.

M. Gardner. Mathematical Games – The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223 :120–123, 1970.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7 :457–511, 1992.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Texts in Statistical Sciences Series. Chapman and Hall/CRC, second edition edition, 2004.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6 :721–741, 1984.

T. Gneiting and AE. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477) :359–378, 2007.

P. W. Goedhart, H. van der Voet, F. Baldacchino, and S. Arpaia. A statistical simulation model for field testing of non-target organisms in environmental risk assessment of genetically modified plants. *Ecology and Evolution*, 4(8) :1267–1283, 2014. ISSN 2045-7758. doi : 10.1002/ece3.1019.

A.S. Goggi, P. Caragea, H. Lopez-Sanchez, M. Westgate, R. Arritt, and C. Clark. Statistical analysis of outcrossing between adjacent maize grain production fields. *Field Crops Research*, 99(2–3) :147 – 157, 2006. ISSN 0378-4290. doi : http://dx.doi.org/10.1016/j.fcr.2006.04.005. URL http://www.sciencedirect.com/science/article/pii/S0378429006000918.

A.S. Goggi, H. Lopez-Sanchez, P. Caragea, M. Westgate, R. Arrit, and C.A. Clark. Gene flow in maize fields with different local pollen densities. *Internationnal Journal of Biometeorology*, 51 :493–503, 2007.

D. Gouache, F. Bensadoun, A.and Brun, C. Pagé, D. Makowski, and D. Wallach. Modelling climate change impact on Septoria tritici blotch (STB) in France : Accounting for climate model and disease model uncertainty. *Agricultural and Forest Meteorology*, 170 :242–252, 2013.

T.G. Gregoire and C. Salas. Ratio estimation with measurement error with auxiliary variate. *Biometrics*, 65 :590–598, 2009.

D. I. Gustafson, I. O. Brants, M. J. Horak, K. M. Remund, E. W. Rosenbaum, and J. K. Soteres. Empirical modeling of genetically modified maize grain production practices to achieve European Union labeling thresholds. *Crop Science*, 46 : :2133–2140, 2006. doi : 10.2135/cropsci2006.01.0060.

J.A. Hanley and B.J. McNeil. The meaning and use of the area under the ROC curve. *Radiology*, 143 :29–36, 1982.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 :97–109, 1970. doi : 10.1093/biomet/57.1.97.

Haut Conseil des Biotechnologies. Avis en réponse à la saisine 100506-projet saisine HCB-coexistence sur la définition des conditions techniques relatives à la mise en culture, la récolte, le stockage et le transport des végétaux génétiquement modifiés. *Comité scientifique du HCB*, 46p, 2011.

J.A. Heinemann, A.D. Sparrow, and T. Traavik. Is confidencein the monitoring og GE food justified ? *Trends in Biotechnology*, 22 :331–336, 2004.

D. Higdon, M. Kennedy, J.C. Cavendish, Cafeo J.A., and R.D. Ryne. Combining field data and computer simulations for calibration and prediction. *Journal of Scientific Computing*, 26(2) :448–466, 2004.

D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103 :570–583, 2008.

S.H. Ibrahim and H.A. Selim. Honeybee activity on gathering pollen from corn plants *Zea mays*. *Agricultural Research Review*, 107–113, 1972.

J Ingram. The separation distance required to ensure cross-pollination is below specified limits in non-seed crops of sugar beet, maize and oilseed rape. *Plant Varieties & Seeds*, 13 :181–199, 2000.

E. Innocenti, X. Silvani, A. Muzy, and D. Hill. A software framework for fine grain parallelization of cellular models with openmp : Application to fire spread. *Environmental Modelling & Software*, 24(7) :819 – 831, 2009. ISSN 1364-8152. doi : http://dx.doi.org/10.1016/j.envsoft.2008.11.014. URL http://www.sciencedirect.com/science/article/pii/S136481520800234X.

D.S. Ireland, D.O. Wilson, M.E. Westgate, J.S. Burris, and M.J Lauer. Managing reproductive isolation in hybrid seed corn production. *Crop Science*, 46 :1445–1455, 2006.

A. Ivanovska, C. Vens, N Colbach, M Debeljak, and S. Džeroski. The feasibility of coexistence between conventional and genetically modified crops : Using machine learning to analyse the output of simulation models. *Ecological Modelling*, 215(1–3) :262 – 271, 2008. ISSN 0304-3800. doi : http://dx.doi.org/10.1016/j.ecolmodel.2008.02.031. URL http://www.sciencedirect.com/science/article/pii/S0304380008001075.

A. Ivanovska, L. Todorovski, M. Debeljak, and S. Džeroski. Modelling the outcrossing between genetically modified and conventional maize with equation discovery. *Ecological Modelling*, 220(8) :1063 – 1072, 2009. ISSN 0304-3800. doi : http://dx.doi.org/10.1016/j.ecolmodel.2009.01.035. URL http://www.sciencedirect.com/science/article/pii/S0304380009000957.

N. Jarosz, B. Loubet, and L. Huber. Modelling airborne concentration and deposition rate of maize pollen. *Atmospheric Environment*, 38 :5555–5566, 2004. URL https://hal.archives-ouvertes.fr/hal-00277200.

J.M. Jones and J.S. Brooks. Effectiveness and distance of border rows in preventing outcrossing in corn. *Oklahoma Agric. Exp. Sta. Tech. Bull.*, T-38 :1–18, 1950.

S. Kawashima, H. Nozaki, T. Hamazaki, S. Sakata, T. Hama, K. Matsuo, and A. Nagasawa. Environmental effects on long-range outcrossing rates in maize. *Agriculture, Ecosystems and Environment*, 142 :410–418, 2011.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society*, 63 :425–464, 2001.

E. Klein, C. Lavigne, X. Foueillassar, P.H. Gouyon, and C. Larédo. Corn pollen dispersal : Quasi-mechanistic models and field experiments. *Ecological Monographs*, 73 :131–150, 2003.

E. Klein, C. Lavigne, H. Picault, M. Renard, and P.H. Gouyon. Pollen dispersal of oilseed rape : estimation of the dispersal function and effects of field dimension. *Journal of Applied Ecology*, 43 :141–151, 2006a. doi : 10.1111/j.1365-2664.2005.01108.x.

K. Klein, C. Lavigne, and P.H. Gouyon. Mixing of propagules from discrete sources at long distance : comparing a dispersal tail to an exponential. *BMC Ecology*, 6 :3, 2006b. doi : 10.1186/1472-6785-6-3. URL http://www.biomedcentral.com/1472-6785/6/3.

K. Kobayashi and M.U. Salam. Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal*, 92(2) :345–352, 2000.

T. Krueger, T. Page, K. Hubacek, L. Smith, and K. Hiscock. The role of expert opinion in environmental modelling. *Environmental Modelling & Software*, 36 :4–18, 2012.

P. M. Kuhnert, Martin T. G., Mengersen K., and Possingham H. P. Assessing the impacts of grazing levels on bird density in woodland habitat : a bayesian appproach using expert opinion. *Environmetrics*, 16 :717–747, 2005a.

P.M. Kuhnert, T.G. Martin, K. Mengersen, and H.P. Possinghma. Assessing the impacts of grazing levels on bird density in wookland habitat : a Bayesian approach using expert opinion. *Environmetrics*, 16 :717–747, 2005b.

D. Lambert. Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34 :1–15, 1992.

M. Lamboni, D. Makowski, and H. Monod. Multivariate global sensitivity analysis for dynamic crop models. *Field Crops Research*, 113 :312–320, 2009.

M. Langhof, B. Hommel, A. Hüsken, J. Schiemann, P. Wehling, R. Wilhelm, and G. Rühl. Coexistence in maize : Do non-maize buffer zones reduce gene flow between maize fields. *Crop Science*, 48 :305–316, 2008.

M. Langhof, B. Hommel, A. Hüsken, C. Njontie, J. Schiemann, P. Wehling, R. Wilhelm, and G. Rühl. Coexistence in maize : isolation distance in dependance on conventional maize field depthand separate edge harvest. *Crop Science*, 50 :1496–1508, 2010.

C. Larédo and A. Grimaud. Stochastic models and statistical inference for plant pollen dispersal. *Journal de la Société Française de Statistique*, 148 :77–105, 2007.

A. Lavigne, L. Bel, E. Parent, and N. Eckert. A model for spatio-temporal clustering using multinomial probit regression : application to avalanche counts. *Environmetrics*, 23 :522–534, 2012.

C. Lavigne, E.K. Klein, P. Vallée, J. Pierre, B. Godelle, and M. Renard. A pollen-dispersal experiment with transgenic oilseed rape. estimation of the average pollen dispersal of an individual plant within a fiel. *Theoretical and Applied Genetics*, 96 :886–896, 1998.

C. Lavigne, C. Devaux, A. Deville, A. Garnier, E.K. Klein, J. Lecomte, S. Pivard, and P.H. Gouyon. Potentials and limits of modelling to predict the impact of trangenic crop in wild species. In H.C.M. De Nijs, D. Bartsch, and J. Sweet, editors, *Introgression from genetically modified plants into wild relatives*, pages 351–364. CABI publishing, 2004.

C. Lavigne, E. Klein, J-F. Mari, F. Le Ber, K. Adamczyk, H. Monod, and F. Angevin. How do genetically modified (GM) crops contribute to background levels of GM pollen in an agricultural landscape ? *Journal of Applied Ecology*, 45 :1104–1113, 2008. doi : 10.1111/j.1365-2664.2008.01504.x.

J.B. Lecomte, H.P. Benoît, M.P. Etienne, L. Bel, and E. Parent. Modeling the habitat associations and spatial distribution of benthic macroinvertebrates : A hierarchical bayesian model for zero-inflated biomass data. *Ecological Modelling*, 265(0) :74 – 84, 2013.

T Lenser and A. Constable. A nonparametric algorithm to model movement between polygon subdomains in a spatially explicit ecosystem model. *Ecological Modelling*, 206 (1–2) :79–92, 2007. doi : http://dx.doi.org/10.1016/j.ecolmodel.2007.03.021.

W.C. Lewin, J. Freyhof, V. Huckstorf, T. Mehner, and C. Wolter. When no catches matter : Coping with zeros in environmental assessments. *Ecological Indicators*, 10(3) : 572–583, 2010.

K. Lipsius, R. Wilhelm, O. Richter, K.J. Schmalstieg, and J. Schiemann. Meteorological input data requirements to predict cross-pollination of GMO Maize with Lagrangian approaches. *Environmental Biosafety Research*, 5(3) :151–168, 2006. doi : 10.1051/ebr: 2007005. URL http://dx.doi.org/10.1051/ebr:2007005.

J.H. Lonnquist and R.W. Jugenheimer. Factors affecting the succes of pollination in corn. *Journal of the American Society of Agronomy*, 35 :923–933, 1943.

C. Loos, R. Seppelt, S. Meier-Bethke, J. Schiemann, and O. Richter. Spatially explicit modelling of transgenic maize pollen dispersal and cross-pollination. *Journal of Theoretical Biology*, 225(2) :241 – 255, 2003. ISSN 0022-5193. doi : http://dx.doi.org/10.1016/S0022-5193(03)00243-1. URL http://www.sciencedirect.com/science/article/pii/S0022519303002431.

V.S. Luna, M.J. Figueroa, M.B. Baltazar, L.R. Gomez, R. Townsend, and J.B. Schoper. Maize pollen longevity and distance isolation requirements for effective pollen control. *Crop Science*, 41 :1551–1557, 2001.

R. Macarthur, M. Feinberg, and Y. Bertheau. Construction of measurement uncertainty profiles for quantitative analysis of genetically modified organismq based on interlaboratory validation data. *Journal of AOAC International*, 93 :1046–1056, 2010.

D. Makowski and H. Monod. *Analyse statistique desrisques environnementaux.Étude de cas*. Springer, 2011.

D. Makowski, J. B. Denis, L. Ruck, and A. Penaud. A Bayesian approach to assess the accuracy of a diagnostic test based on plant disease measurement. *Crop Protection*, 27 (8) :1187–1193, 2008.

A. Marceau, Guerineau L., L. Huber, F. Angevin, and H. Monod. Modelling maize pollen emission during the day and the flowering period. *Aspects of Applied Biology*, 89 :17–22, 2008.

A. Marceau, B. Loubet, B. Andrieu, B. Durand, X. Foueillassar, and L. Huber. Modelling diurnal and seasonal patterns of maize pollen emission in relation to meteorological factors. *Agricultural and Forest Meteorology*, 151 :11–21, 2011.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *PNAS*, 100 :15324–15328, 2003.

C.E. Mason and K.T. Tracewski. Diurnal foraging activity for corn pollen by honey bees. *Environmental Entomology*, 187–188, 1982.

A. Meillet, F. Angevin, A. Bensadoun, A. Cathary, G. Huby, H. Monod, and A. Messéan. Design of a decision support tool for coexistence at farm and regional level. In *GMCC 2013 - 6th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains*, Lisbon, Portugal, November 2013.

A. Messéan, F. Angevin, M. Gomez-Barbero, K. Menrad, and E. Rodriguez-Cerezo. New case studies on the coexistence of GM and non-GM crops in European agriculture. *Technical Report Series of the Joint Research Center of the European Commision*, EUR 22102 :886–896, 2006.

A. Messéan, G. Squire, F. Perry, J. Angevin, P. Gomez, M. Townend, C. Sauuse, B. Breckling, S. Langrell, S. Dzeroski, and J. Sweet. Sustainable introduction of GM crops into european agriculture : a summary report of the FP6 SIGMEA research project. *OCL*, 16 :37–51, 2009.

J. Messeguer, G. Peñas, J. Ballester, M. Bas, J. Serra, J. Salvia, M. Palaudelmàs, and E. Melé. Pollen-mediated gene flow in maize in real situation of coexistence. *Plant Biotechnology Journal*, 4 :633–645, 2006.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21 : 1087–1092, 1953. doi : 10.1063/1.1699114.

J.M. Meynard and M. Le Bail. Isolement des collectes et maitrise des dissémination au champ. Rapport du groupe 3, "Pertinence et faisabilité d'une filière sans utilistaion d'OGM". *INRA-FNSEA*, 56p, 2001.

R. B. Millar. Comparison of hierarchical bayesian models for overdispersed count data using DIC and bayes'factors. *Biometrics*, 65 :962–969, 2009.

G. Miller, Justus A., Vostrotin V., Dry D., and Bertelli L. Poisson mixture model for measurements using counting. *Radiation protection dosimetry*, 138(4) :363–375, 2010.

A. Moreno-Jiménez and R.L. Hodgart. Modelling a single type of environmental impact from an obnoxious transport activity : implementing locational analysis with GIS. *Environment and Planning*, 35(5) :931–946, 2003.

F. Mortier, O. Flores, and S. Gourlet-Fleury. Spatial bayesian modesl of tree density with zero inflation and autocorrelation. *Journal de la Société Française de Statistique*, 148 (1) :39–51, 2007.

C. Njontie, X. Foueillassar, N.K. Christov, and A. Hüsken. The impact of GM seed admixture on the non-GM harvest product in maize (*Zea mays* L.). *Euphytica*, 180 : 163–172, 2011.

T.O. Ojo, J. S. Bonner, and Page C. A. Simulation of constituent transport using a reduced 3d constituent transport model (ctm) driven by {HF} radar : Model application and error analysis. *Environmental Modelling & Software*, 22(4) :488 – 501, 2007. doi : http://dx.doi.org/10.1016/j.envsoft.2006.02.010.

M. Palaudelmàs, G. Penas, E. Melé, J. Serra, J. Salvia, M. Pla, A. Nadal, and J. Messeguer. Effect of volunteers on maize gene flow. *Transgenic Research*, 18 :583–594, 2009.

M. Palaudelmàs, E. Melé, A. Monfort, J. Serra, J. Salvia, and J. Messeguer. Assessment of the influence of field size on maize gene flow using SSR analysis. *Transgenic Research*, 21 :471–483, 2012. doi : 10.1007/s11248-011-9549-z.

J. Papaïx, H. Goyeau, P. Du Cheyron, H. Monod, and C. Lannou. Influence of cultivated landscape composition on variety resistance : an assessment based on wheat leaf rust epidemics. *New Phytologist*, 191(4) :1095–1107, 2011. doi : 10.1111/j.1469-8137.2011.03764.x. URL http://dx.doi.org/10.1111/j.1469-8137.2011.03764.x.

J. Papaïx, O. David, C. Lannou, and H. Monod. Dynamics of adaptation in spatially heterogeneous metapopulations. *PLoS ONE*, 8(2) :e54697, 2013. doi : 10.1371/journal.pone.0054697. URL http://dx.doi.org/10.1371%2Fjournal.pone.0054697.

J. Papaïx, J.J. Burdon, C. Lannou, and P.H. Thrall. Evolution of pathogen specialisation in a host metapopulation : Joint effects of host and pathogen dispersal. *PLoS Comput Biol*, 10(5) :e1003633, 2014. doi : 10.1371/journal.pcbi.1003633. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1003633.

E. Parent and J Bernier. *Le raisonnement bayésien*. Springer, 2007.

L. Paul, F. Angevin, C. Collonier, and A. Messéan. Impact of gene stacking on gene flow – the case of maize. *Transgenic Research*, 21 :243–256, 2012.

A. Philibert, M.L. Desprez-Loustau, B. Fabre, P. Frey, F. Halkett, C. Husson, B. Lung-Escarmant, B. Marçais, C. Robin, C. Vacher, and D. Makowski. Predicting invasion success of forest pathogenic fungi from species traits. *Journal of Applied Ecology*, 48 (6) :1381–1390, 2011. doi : 10.1111/j.1365-2664.2011.02039.x. URL http://dx.doi.org/10.1111/j.1365-2664.2011.02039.x.

J. Pierre, B. Vaissière, P. Vallée, and M. Renard. Mise en suspension du pollen par les abeilles et incidencede ce pollen sur la fécondation. In *Séminaire de restitution de l'AIP "OGM et environnement"* (1998-2001), Paris, pp.11-13, 2002.

M. Pla, José-LuisLa Paz, Gisela Peñas, Nora García, Montserrat Palaudelmàs, Teresa Esteve, Joaquima Messeguer, and Enric Melé. Assessment of real-time pcr based methods for quantification of pollen-mediated gene flow from gm to conventional maize in a field study. *Transgenic Research*, 15(2) :219–228, 2006. ISSN 0962-8819. doi : 10.1007/s11248-005-4945-x. URL http://dx.doi.org/10.1007/s11248-005-4945-x.

M. Plummer. *JAGS Version 3.3 user manual*, 2012. URL http://www-fis.iarc.fr/~martyn/software/jags/.

M. Plummer, N. Best, and K. Cowles, K. abd Vines. *The coda Package : Output analysis and diagnostics for MCMC*, 2009. URL http://www.R-project.org.

R. Pouillot, I. Albert, M. Cornu, and J. B. Denis. Estimation of uncertainty and variability in bacterial growth using bayesian inference. application to *listeria monocytogenes*. *International Journal of Food Microbiology*, 81 :87–104, 2003.

J.W. Purseglove. *Tropical Crops, Monocotyledons*. Longman Group Ltd, London, 1972.

R Development Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL http://www.R-project.org. ISBN 3-900051-07-0.

M. Ramin, S. Stremilov, T. Labencki, A. Gudimov, D. Boyd, and G.B. Arhonditsis. Integration of numerical modeling and Bayesian analysis for setting water quality criteria in Hamilton, Ontario, Canada. *Environmental Modelling & Software*, 26 :337–353, 2011.

R. Rasmussen and G. Hamilton. An approximate bayesian computation approach for estimating parameters of complex environmental processes in a Cellular Automata. *Environmental Modelling & Software*, 29 :1–10, 2012.

G. Ravazzani, D. Rametta, and M. Mancini. Macroscopic cellular automata for groundwater modelling : A first approach. *Environmental Modelling & Software*, 26(5) :634 – 643, 2011. ISSN 1364-8152. doi : http://dx.doi.org/10.1016/j.envsoft.2010.11.011. URL http://www.sciencedirect.com/science/article/pii/S1364815210003154.

A. Raveneau. Stratégies de séparation des filières OGM et non-OGM en amont de la chaîne logistique d'appovisionnement. Master's thesis, ENESAD, Mémoire de fin d'études, 2005.

G.S. Raynor, E.C. Ogden, and J.V. Hayes. Dispersion and deposition of corn pollen from experimental sources. *Agronomy Journal*, 64 :420–427, 1972.

B. J. Reich, E. Kalendra, C. B. Storlie, H. D. Bondell, and M. Fuentes. Variable selection for high dimensional bayesian density estimation : application to human exposure simulation. *Applied Statistics*, 61-1, 2012. doi : 00359254/12/61000.

O. Richter and R. Seppelt. Flow of genetic information through agricultural ecosystems : a generic modelling framework with application to pesticide-resistance weeds and genetically modified crops. *Ecological Modelling*, 174 :55–66, 2004. doi : 10.1016/j.ecolmodel.2003.12.046.

L. Riesgo, F.J. Areal, O. Sanvido, and E. Rodríguez-Cerezo. Distances needed to limit cross-fertilization between gm and conventional maize in europe. *Nature Biotechnology*, 28 :780–782, 2010. doi : 10.1038/nbt0810-780.

I. Riznov and E. Rodríguez-Cerezo. The european coexistence bureau : Five years' experience. *AgBioForum*, 1 :22–27, 2014.

C. Robert. *Le Choix Bayésien, Principes et Pratiques*. Statistique et Probabilités Appliquées. Springer, 2006.

P.R.H. Robson, R. Kelly, E.F. Jensen, G.D. Giddings, M. Leitch, C. Davey, A.P. Gay, G. Jenkins, H. Thomas, and I.S. Donnison. A flexible quantitative methodology for the analysis of gene-flow between conventionally bred maize populations using microsatellite markers. *Theoretical and Applied Genetics*, 122(4) :819–829, 2011. ISSN 0040-5752. doi : 10.1007/s00122-010-1489-0. URL http://dx.doi.org/10.1007/s00122-010-1489-0.

P. Roger and L. Eeckhoudt. Valeur ajoutée d'un transfert de risque et optimum de pareto. *Revue économique*, 49(2) :325–333, 1998. ISSN 0035-2764. doi : 10.2307/3502511. URL /web/revues/home/prescript/article/reco_0035-2764_1998_num_49_2_409981.

République Française, 2008. Loi No 2008-595 du 25 juin 2008 relative aux organismes génétiquement modifiés. *Journal Officiel de la République Française*, 2008-595, 11p.

République Française, 2012. Décret No 2012-128 du 30 janvier 2012 relatif à l'étiqutage des denrées alimentaires issues de filières qualifiées "sans organismes génétiquement modifiés". *Journal Officiel de la République Française*, 2012/128, 4p.

G. Rühl and M. Langhof. Coexistence in maize : effect of the genetically modified maize field depth on pollen-mediated gene flow. *Crop Science*, 51 :2186–2193, 2011.

R.K. Saiki, S. Scharf, F. Faloona, K.B. Mullis, G.T. Horn, H.A. Erlich, and N. Arnheim. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732) :1350–1354, 1985. doi : 10.1126/science.2999980. URL http://www.sciencemag.org/content/230/4732/1350.abstract.

O. Sanvido, F. Widmer, M. Winzeler, B. Streit, E. Szerencsits, and F. Bigler. Definition and feasibility of isolation distances for transgenic maize cultivation. *Transgenic Research*, 17(3) :317–335, 2008. doi : 10.1007/s11248-007-9103-1.

C. Sausse, M. Le Bail, B. Lecroart, B. Remy, and A. Messéan. How to manage the coexistence between genetically modified and conventional crops in grain and oilseed collection areas ? elaboration of scenarios using role playing games. *Land Use Policy*, 30(1) :719 – 729, 2013. ISSN 0264-8377. doi : http://dx.doi.org/10.1016/j.landusepol.2012.05.018. URL http://www.sciencedirect.com/science/article/pii/S0264837712001044.

G. Sileshi. The excess-zero problem in soil animal count data and choice of appropriate models for statistical inference. *Pedobiologia*, 52(1) :1–17, 2008.

S. Singh, G. Ash, and M. Hodda. Keeping 'one step ahead' of invasive species : using an integrated framework to screen and target species for detailed biosecurity risk assessment. *Biological Invasions*, pages 1–18, 2014. ISSN 1387-3547. doi : 10.1007/s10530-014-0776-0.

R. Slimi, S. El Yacoubi, E. Dumonteil, and S. Gourbière. A cellular automata model for chagas disease. *Applied Mathematical Modelling*, 33(2) :1072 – 1085, 2009. ISSN 0307-904X. doi : http://dx.doi.org/10.1016/j.apm.2007.12.028. URL http://www.sciencedirect.com/science/article/pii/S0307904X07003563.

D. Spiegelhalter, A. Thomas, Best N., and Lunn D. *WinBUGS version 1.4.3 User Manual.* MRC Biostatistics Unit, Cambridge, UK, 2007. URL http://www.mrc-bsu.cam.ac.uk.

David J. Spiegelhalter. Bayesian graphical modelling : a case-study in monitoring health outcomes. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 47 (1) :115–133, 1998. ISSN 1467-9876. doi : 10.1111/1467-9876.00101. URL http://dx.doi.org/10.1111/1467-9876.00101.

P.C. Struik and T. Makonnen. Effects of timing, intensity and duration of pollination on kernel set and yield in maize (*Zea mays* L.) under temperature conditions. *Netherlands Journal of Agricultural Science*, 40 :409–429, 1992.

S. Trapmann, P. Corbisier, H. Schimmel, and H. Emons. Towards future reference systems for GM analysis. *Analytical and Bioanalytical Chemistry*, 396 :1969–1975, 2010.

R. Treu and J. Emberlin. Pollen dispersal in the crops Maize (*Zea mays*), Oilseed rape (*Brassica napus ssp oleifera*), Potatoes (*Solanum tuberosum*), Sugar beet (*Beta vulgaris ssp vulgaris*) and wheat (*Triticum aestivum*), Evidence from publications ; a report for the Soil Association from the National Pollen research unit. 2000.

Y. Trifa and D. Zhang. DNA content in embryo and endospermof maize kernel (Zea mays l.) : Impact on GMO quantification. *Journal of Agricultural and Food Chemistry*, 52 : 1044–1048, 2004.

C.C.M. Van De Wiel, R.M.W. Groeneveld, O. Dolstra, E.J. Kok, I.M.J. Scholtens, J.T.N.M. Thissen, M.J.M. Smulders, and L.A.P. Lotz. Pollen-mediated gene flow in maize tested for coexistence of {GM} and non-gm crops in the netherlands : effect of isolation distances between fields. {*NJAS*} - *Wageningen Journal of Life Sciences*, 56(4) :405 – 423, 2009. ISSN 1573-5214. doi : http://dx.doi.org/10.1016/S1573-5214(09)80007-9. URL http://www.sciencedirect.com/science/article/pii/S1573521409800079.

V. Viaud, H. Monod, C. Lavigne, F. Angevin, and K. Adamczyk. Spatial sensitivity of maize gene-flow to landscape pattern : a simulation approach. *Landscape Ecology*, 23 : 1067–1079, 2008.

John Von Neumann. *Theory of Self-Reproducing Automata*. University of Illinois Press, Champaign, IL, USA, 1966.

D. Wallach, D. Makowski, and J.W. Jones. *Working with Dynamic Crop Models*. Elsevier, 2006.

J. Wang and X. Yang. Development and validation of atmospheric gene flow model for assessing environmental risks from transgenic corn crops. *International Journal of Agricultural and Biological Engineering*, 3(2) :18–30, 2010.

W.E. Weber, T. Bringezu, I. Broer, J. Eder, and F. Holz. Coexistence between GM and non-GM maize crops - tested in 2004 at the field scale. *Journal of Agronomy and Crop Science*, 193 :79–92, 2007.

R. Weekes, C. Deppe, T. Allnutt, C. Boffey, D. Morgan, S. Morgan, M. Bilton, R. Daniels, and C. Henry. Crop-to-crop gene flow using farm scale sites of oilseed rape (brassica napus) in the uk. *Transgenic Research*, 14(5) :749–759, 2005. ISSN 0962-8819. doi : 10.1007/s11248-005-0943-2. URL http://dx.doi.org/10.1007/s11248-005-0943-2.

D. Zhang, A. Corlet, and S. Fouilloux. Impact of genetic structures on haploid genome-based quantification of genetically modified DNA : theoretical considerations, experimental data in MON 810 maize kernels (Zea mays l.) and some practical applications. *Transgenic Research*, 17 :393–402, 2008.

J. Šuštar Vozlič, K. Rostohar, A. Blejec, P. Kozjak, Z. Čergan, and V. Meglič. Development of sampling approaches for the determination of the presence of genetically modified organisms at the field level. *Analytical and Bioanalytical Chemistry*, 396(6) :2031–2041, 2010. ISSN 1618-2642. doi : 10.1007/s00216-009-3406-4. URL http://dx.doi.org/10.1007/s00216-009-3406-4.

Annexes

Annexes 1

Modelling dispersal : a Bayesian Approach Decision Support. Bensadoun, A., Monod, H., Makowski, D., Messéan, A. Soumis à *Environmental Modelling & Software*, 2014.

A Bayesian approach to model dispersal for decision support

Arnaud Bensadoun[a,c,*], Hervé Monod[a], David Makowski[b], Antoine Messéan[c]

[a] *INRA, UR 341 MIAJ, 78352 Jouy-en-Josas, France*
[b] *INRA, UMR 211 Agronomie, 78850 Thiverval-Grignon, France*
[c] *INRA, UAR 1240 Eco-Innov, 78850 Thiverval-Grignon, France*

Abstract

In agricultural and environmental sciences dispersal models are often used for risk assessment to predict the risk associated with a given configuration and also to test scenarios that are likely to minimise those risks. Like any biological process, dispersal is subject to biological, climatic and environmental variability and its prediction relies on models and parameter values which can only approximate the real processes. In this paper, we present a Bayesian method to model dispersal using spatial configuration and climatic data (distances between emettors/receptor and main wind direction) while accounting for uncertainty, with an application to the prediction of adventitious presence rate of genetically modified (GM) in a nonGM field. This method includes the design of candidate models, their calibration, selection and evaluation on an independent dataset. A group of models was identified that is sufficiently robust to be used for prediction purpose. The group of models allows to include local information and it reflects reliably enough the observed variability in the data so that probabilistic model predictions can be performed and used to quantify risk under different scenarios or derive optimal sampling schemes.

Keywords: Dispersal, variability, uncertainty, Bayesian inference, MCMC, decision support, risk assessment, sampling, zero-exess data

*Corresponding author
Email address: arnaud.bensadoun@jouy.inra.fr (Arnaud Bensadoun)

Highlights

- A Bayesian approach is proposed to model dispersal and to make probabilistic predictions which account for uncertainty.
- The proposed approach is applied to assess the risk of cross-pollination due to pollen dispersal between genetically (GM) modified and conventional (non-GM) maize fields.
- 16 statistical gene flow models were designed, calibrated and compared within the Bayesian framework.
- Models with a zero-inflated Poisson distribution and with exponential decay turn out to provide the most reliable predictions.
- The Bayesian framework allows to set up context-specific isolation distances by providing accurate probabilistic predictions.
- Thanks to a precise prediction of the intra-field variability, our models allow to design context-specific optimal stratified sampling schemes.

1. Introduction

1.1. Dispersal

Dispersal refers to the spreading of organisms or particles from one place to another. In agroecological systems, this process is of paramount importance as it can drive several spatial processes associated with adverse effects such as an epidemic spread (see Papaïx et al., 2013, 2014, for instance) or the colonization of a territory by an invasive species, see examples given by Philibert et al. (2011); Singh et al. (2014). In other fields, pollutant transport (Ojo et al., 2007), movement of biomass in oceans (Lenser & Constable, 2007) as well as dispersion of aircraft noise around airport (Moreno-Jiménez & Hodgart, 2003) also involve dispersal as a key driver.

The general question addressed in those situations is to control and possibly pilot a system in such a way that the probability of exceeding a certain threshold

or level of abundance is minimised or at least made acceptable. Models that include dispersal are therefore needed for the risk assessment associated with a given configuration and eventually to test scenarios that are likely to minimise the risk. Another purpose of dispersal models, provided they are able to properly transcribe the response variability within the surface or volume of interest, is to derive optimal and cost-efficient sampling schemes (Messeguer et al., 2006; Bancal et al., 2013). Developing dispersal models adequate for such objectives is necessary but it requires to tackle a number of modelling issues.

In agricultural and environmental sciences, spatially explicit models have been developed to estimate outcomes of dispersal processes (e.g. Colbach et al., 2001; Angevin et al., 2008). To describe the decrease of the expected result when distance to the emission sources increases, various dispersal functions or kernels have been defined (Lavigne et al., 1998; Clark et al., 1999; Klein et al., 2003). Depending on how they are selected and calibrated, those dispersions can be extrapolated more or less reliably to situations of interest that are often substantially different from the situations used for calibration.

Other difficulties that need to be addressed when modelling dispersal processes arise from their dependence on multiple environmental factors and from their stochasticity, even conditionally to the knowledge of specific factors like spatial configuration and climate (Larédo & Grimaud, 2007). A particular feature of data gathered in dispersal experiments is the great variability they exhibit, which is directly related to the stochastic nature of the dispersal process and possibly to additional variability related to the observation process. To take this into account, one has to make particular hypotheses about the distribution of the observations. This distribution is often characterised by an excess of zeros with respect to classical assumptions about counts (Kuhnert et al., 2005; Lecomte et al., 2013; Goedhart et al., 2014), and by a complex relationship between the variance and the mean of the observed response.

Another point to be raised is the difference between the scale at which the process takes place and the scale at which decisions are to be made. Roughly said, the dispersal is a *point-to-point* process whereas decisions are taken at a

surface or volume level. Thus pointwise responses need to be integrated over the support of interest and this raises the question of the most appropriate scales for observation, prediction and decision. Typical examples of this situation are found in epidemiology, where the spread of a disease is often due to individual contacts whereas control strategies are only relevant at the population scale (Papaïx et al., 2013). Other examples are to be found in the context of pollen mediated gene flow, where cross-fertilisation occurs at the plant scale while appropriate decisions are only applicable at the field or farm scale (Angevin et al., 2008).

1.2. Bayesian approach

Dispersal is subject to biological, climatic and environmental variability and its prediction relies on models and parameter values which can only approximate the real processes (Klein et al., 2003; Larédo & Grimaud, 2007; Gouache et al., 2013). Therefore, in the context of risk assessment, it is necessary that model predictions account not only for factors affecting dispersal that are known but also for the uncertainty in the parameters and for the unpredictable sources of variability that could arise, for instance, from a poor knowledge of factors (Clark, 2005; Makowski et al., 2008). Similarly, when models are used to devise a sampling design, it is essential to have an accurate assessment of the correlation between predicted and actual responses but also of the correlation between the predicted and observed variabilities (Cochran, 1977; Gregoire & Salas, 2009; Bancal et al., 2013).

Bayesian methods provide the scientist with the ability to formally incorporate expert knowledge into the model via prior distributions (Clark, 2005; Carlin & Louis, 2009; Krueger et al., 2012), and they allow quantitative assessment of uncertainty and its integration through probabilistic predictions (see Ramin et al., 2011, for instance). In the context of dispersal they have been considered for oilseed rape pollen by Damgaard & Kjellson (2005) and for wheat leaf rust epidemics by Papaïx et al. (2011), but with no specific focus on modelling accurately the variability and/or uncertainty.

Another advantage of Bayesian methods is that they give much flexibility to incorporate refined modelling of both the deterministic and the stochastic components of the biophysical and observation processes. So a promising direction for model-based decision support is to develop models that incorporate context-specific information as well as uncertainties and to use Bayesian inference on these models for prediction, risk assessment (Bensadoun et al., 2013) and sampling design (Bancal et al., 2013).

1.3. Objectives

Our main research goal is to design model-based and Bayesian approaches and to propose a general procedure to *i)* model dispersal from donor to receptor; *ii)* estimate such models using a Bayesian approach; *iii)* select the best model or the best combination of models depending on the prediction purpose.

The major issue here is to develop models and inference that allow to make reliable probabilistic predictions, to be used for decision support or for devising optimal sampling schemes. Such models must integrate input variables adapted to the available information and provide accurate description of the variability at different scales.

In this paper, we apply the proposed approach to assess the risk of cross-pollination due to pollen dispersal between genetically modified (GM) and conventional (non-GM) maize fields. We reconsider data from a maize pollen dispersal experiment presented in Klein et al. (2003) whose characteristics are, to the best of our knowledge, representative of typical data gathered in dispersal experiments. We assume the data for model selection and parameter estimation consist of such representative past experiments and the available information for the new site where prediction is needed includes its spatial configuration and the main wind direction.

2. Models

In the following, we focus on a target area in the site of interest and assume that dispersal occurs from a number of surrounding sources, possibly including

the target area itself. Observations are made on the result of the dispersal process in the target area.

In the dispersal models, the observed value y_s at location s in the target area is considered to be the realisation of a random variable Y_s depending on $\lambda_s^{'}$, where $\lambda_s^{'}$ denotes the expected outcome of the dispersal process at location s. We present now a series of alternative choices for Y_s and $\lambda_s^{'}$, which we consider to be key components in the construction of dispersal models.

2.1. Observation Model

In our experience on agricultural and environmental applications, observations can rarely be assumed to follow the most standard probability distributions. In particular, attention must be paid to overdispersion and to an excess of zero values compared to default hypotheses.

2.1.1. Poisson or ZIP distribution

To be coherent with the case study, we consider the case when the observation Y_s is a counting variable. The Poisson distribution is convenient to model counts of rare events (Besag et al., 1991). Thus, the default assumption on the distribution of Y_s is likely to be the Poisson distribution:

$$Y_s|\lambda_s^{'} \sim \mathcal{P}(\lambda_s^{'}). \tag{1}$$

However, outcomes of a dispersal process over spaces (for instance samples of marine species in Lecomte et al. (2013) or counts of genetically modified grains in a conventional maize cob in Goedhart et al. (2014)) exhibit a high variability, and especially an excess of zeros (see Cunningham & Lindenmayer, 2005; Sileshi, 2008; Lewin et al., 2010, for instance). The Zero Inflated Poisson distribution (ZIP) is a good candidate to cope with that excess (Kuhnert et al., 2005; Goedhart et al., 2014) as it consists of a mixture of a Poisson and a Dirac distribution in zero (see Lambert, 1992, for full description). Thus an alternative assumption on the distribution of Y_s is

$$Y_s|\lambda_s^{'}, q_s \sim \mathrm{ZI}\mathcal{P}(1-q_s, \lambda_s^{'}). \tag{2}$$

According to the ZIP distribution, the observation is either constrained to zero with probability p_s, or it follows a Poisson distribution with probability $q_s = 1 - p_s$. To implement the ZIP mixture we define a hidden variable Z distributed as a Bernoulli variable :

$$Z_s \sim \mathcal{B}\text{ern}(q_s), \tag{3}$$

with

$$\begin{cases} Y_s = 0 & \text{if } Z_s = 0, \\ Y_s \sim \mathcal{P}(\lambda_s^{'}) & \text{if } Z_s = 1. \end{cases}$$

In addition, the excess of zeros may depend on the distance from s to the source. Thus, denoting by d_s the shortest distance from the source to location s, a logit link may be introduced between d_s and the probability q_s of observing a Poisson process in location s :

$$\text{logit}(q_s) = \beta_1(\beta_2 - d_s), \tag{4}$$

where β_2 is the abscissa of the inflection point of the logistic curve and β_1 is proportional to the slope of the tangent to the logistic curve at the inflection point. Thus the ZIP model has two additional parameters β_1 and β_2 with respect to the Poisson model.

2.1.2. Fixed or Random expected outcome

Very often, the expected outcome $\lambda_s^{'}$ is given by the fixed output λ_s of a deterministic dispersal model. We denote this by the equality :

$$\lambda_s^{'} = \lambda_s. \tag{5}$$

However, inter-site variability can be strong even for sites located very close to each other. In order to take into account such extra and pointwise inter-location variability (or nugget effect), a random lognormal version of the expected outcome can be used:

$$\ln(\lambda_s^{'}) \sim \mathcal{N}(\ln(\lambda_s), \sigma^2). \tag{6}$$

This version has one additional parameter σ^2.

2.2. Expectation model on dispersal

Dispersal *per se* is integrated in the model equations which define λ_s. For simplicity, we consider for now that space is discretised into a finite number of source locations, all of which produce the same quantity Q of a single type of particles. In practice such assumptions are case-specific; see Section 4.3 for their adaptation to the case study.

2.2.1. Individual or Global Dispersal framework

An individual dispersal function $\gamma(s, s')$ (or kernel) gives the proportion of particles emitted at location s' that fall on a site at location s (see Lavigne et al., 1998; Klein et al., 2003, for details). It is mathematically defined as a two-dimensional probability density function (Klein et al., 2006a).

In the individual dispersal framework advocated for example in Lavigne et al. (1998), the calculus of particles arriving at location s requires to compute and integrate the individual dispersal function for all distances between s and any emitting location s'. Thus, the expected count at location s is given by

$$\lambda_s = Q \sum_{s'} \gamma(s, s'), \tag{7}$$

where the sum is over all source locations s'.

When using this framework with a very fine discretisation within a MCMC algorithm (see Section 3.2), even with approximations, the computing time is a major issue, firstly for estimation and secondly for prediction and decision making purposes. In practice, this argues for considering a much simpler way to take sources into account. The global dispersal framework is an example of simplification which was used, for example, by Damgaard & Kjellson (2005) to model oilseed rape gene flow in a *field-to-field* setting (i.e only one source field and only one receptor field). It amounts to replacing equation (12) by e.g.

$$\lambda_s = Q\, \gamma(s, s''). \tag{8}$$

where s'' is the source location closest to s. If several source areas are to be

taken into account, it may become

$$\lambda_s = Q \sum_{s''} \gamma(s, s''), \tag{9}$$

where the sum is over the locations s'' closest to s within each area.

2.2.2. Dispersal functions

Many alternative dispersal functions are available in the litterature. In practice, it is useful to compare several of them. In particular, an important characteristic to consider is the dispersal function behaviour over short and long distances (see Klein et al., 2006b, for instance), either allowing for fat tails (e.g. the 2Dt function below) or for light tails (e.g. the compound exponential below).

Furthermore, dispersal is sensitive to meteorological factors which should be taken into account, as far as possible. In particular wind was identified in Klein et al. (2003) as a key factor of dispersal phenomena in agricultural science (e.g. spores, pollen, ...) and Bensadoun et al. (2013) showed that incorporating wind effect into a dispersal model can substantially reduce prediction uncertainty. Thus we introduce anisotropy due to wind effects in the dispersal functions. We consider a context where predictions for the site of interest must be given long before dispersal has actually occurred, so that only past measures of the average wind direction and speed are available.

In the expressions below, we denote by (x, y) and (x', y') the coordinates of s and s' respectively (in meters), and by Δ the angle in radians between the vector (s', s) and the main wind direction. The main wind direction ω_0 is assumed to be known for each experiment and the average wind speed is assumed to be homogeneous across the sites under study.

Bivariate Student. The Bivariate Student (2Dt) dispersal function defined in Clark et al. (1999) and Lavigne et al. (2008) for its anisotropic version is given by

$$\gamma_{\mathrm{2Dt}}(s, s') = \frac{b-1}{\pi a^2} \left(1 + \frac{d(s, s')^2}{a^2}\right)^{-b} \exp\big(\kappa \cos(\Delta)\big), \tag{10}$$

where a, b and κ are three parameters related to the median dispersal distance, the shape and the anisotropy respectively.

Compound Exponential. The second dispersion function is a compound exponentially decreasing function similar to the one used in Damgaard & Kjellson (2005). This dispersion has the advantage to be very simple but it doesn't take wind effect into account. To overcome this limitation we modify this function by introducing the same exponential anisotropy factor as for the 2Dt:

$$\gamma_{\text{Exp}}(s,s') = \begin{cases} K_e \exp\big(-a_1 d(s,s')\big) \exp\big(\kappa\cos(\Delta)\big), & \text{if } d(s,s') \leq D, \\ K_e \exp\big(-a_1 D - a_2(d(s,s') - D)\big) \exp\big(\kappa\cos(\Delta)\big), & \text{if } d(s,s') \geq D \end{cases} \tag{11}$$

where K_e, a_1, a_2, D, κ are the five parameters to estimate.

3. Inference and model comparison

3.1. Factorial set of models

In order to account for known and unknown sources of variability as well as possible, two alternative choices have just been defined in the previous section for four key model components. Two of these components are related to the variability of the observations Y_s : the distribution of counts resulting from a dispersal process (Poisson or Zero-Inflated Poisson), and the individual variability (Fixed or Random local $\lambda_s^{'}$ values). The two other components are related to the fixed part of the model and more precisely to the dispersal : the way sources are taken into account (Global or Individual) and the dispersal function *per se* (2Dt or Compound exponential).

Without prior information on which model combinations will perform better, we opt for a systematic evaluation of the 16 models resulting from all possible combinations of the four model components. The full factorial set of models is listed in Table 1.

Table 1: List of the factorial set of models used in the case study

Framework	Observation model	Dispersal function	Distribution $\lambda^{'}$	ModelName
Global	$\mathcal{P}(\lambda')$	Exponential	Fixed to λ	GPExpoA
Global	$\mathcal{P}(\lambda')$	Exponential	$\mathcal{N}(\ln(\lambda_s), \sigma^2)$	GPExpoB
Global	$\mathcal{P}(\lambda')$	2Dt	Fixed to λ	GP2DtA
Global	$\mathcal{P}(\lambda')$	2Dt	$\mathcal{N}(\ln(\lambda_s), \sigma^2)$	GP2DtB
Global	$ZI\mathcal{P}(1-q, \lambda')$	Exponential	Fixed to λ	GZExpoA
Global	$ZI\mathcal{P}(1-q, \lambda')$	Exponential	$\mathcal{N}(\ln(\lambda_s), \sigma^2)$	GZExpoB
Global	$ZI\mathcal{P}(1-q, \lambda')$	2Dt	Fixed to λ	GZ2DtA
Global	$ZI\mathcal{P}(1-q, \lambda')$	2Dt	$\mathcal{N}(\ln(\lambda_s), \sigma^2)$	GZ2DtB
Individual	$\mathcal{P}(\lambda')$	Exponential	Fixed to λ	IPExpoA
Individual	$\mathcal{P}(\lambda')$	Exponential	$\mathcal{N}(\ln(\lambda_s), \sigma^2)$	IPExpoB
Individual	$\mathcal{P}(\lambda')$	2Dt	Fixed to λ	IP2DtA
Individual	$\mathcal{P}(\lambda')$	2Dt	$\mathcal{N}(\ln(\lambda_s), \sigma^2)$	IP2DtB
Individual	$ZI\mathcal{P}(1-q, \lambda')$	Exponential	Fixed to λ	IZExpoA
Individual	$ZI\mathcal{P}(1-q, \lambda')$	Exponential	$\mathcal{N}(\ln(\lambda_s), \sigma^2)$	IZExpoB
Individual	$ZI\mathcal{P}(1-q, \lambda')$	2Dt	Fixed to λ	IZ2DtA
Individual	$ZI\mathcal{P}(1-q, \lambda')$	2Dt	$\mathcal{N}(\ln(\lambda_s), \sigma^2)$	IZ2DtB

3.2. Bayesian inference

The Bayesian approach allows for probabilistic predictions which are easily derived from the posterior distribution of the model parameters. This methodology is particularly relevant in our case given the variability of the observations and our interest in assessing the prediction uncertainty.

In the Bayesian approach, model parameters are treated as random variables. The fundamental equation is $P(\theta|Y) \propto P(\theta) \times P(Y|\theta)$. Here θ is the vector of parameters in the model, Y is the vector of observations and $P(\cdot|\cdot)$ denotes conditional probability. The above equation states that the posterior distribution $P(\theta|Y)$, which specifies our knowledge about θ after using the data, is proportional to the product of the prior distribution $P(\theta)$, which represents our knowledge of the parameters before using the data, and the likelihood function $P(Y|\theta)$, which specifies the probability of Y given the parameter vector θ. Note that, in the case where the likelihood is not numerically computable within a reasonable time or not analytically available, the Bayesian approach can still be used in a likelihood-free setting (Marjoram et al., 2003; Rasmussen & Hamilton, 2012)

Bayesian inference is usually achieved using Markov Chain Monte-Carlo (MCMC), for example by using the software JAGS (Plummer, 2012). Simulated data produced by JAGS can then be processed using the CODA statistical R package (Plummer et al., 2009). The main output is a set of hundreds or thousands of parameter vectors which constitute a representative sample of the posterior distribution. Running MCMC algorithms requires to set a few tuning parameters, as detailed in Section 4.5 for the case study.

3.3. Model fit and selection

Classical criteria on the mean response are available to check the global fit of models (Section 3.3.1). But more refined criteria must be used in order to *i)* take full advantage of the probabilistic nature of models predictions (scoring rules) *ii)* assess the quality of the decision resulting from model predictions (ROC Analysis). They are presented in Sections 3.3.2 and 3.3.3.

3.3.1. Classical statistical criteria

Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ be the vector of all the observed data and $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$ be the vector of predicted mean responses. Root mean squared error (RMSE), coefficient of determination (R^2) and correlation (r) are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

$$r = \frac{\sigma_{\mathbf{y}\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}},$$

with $\sigma_{\mathbf{y}}$, $\sigma_{\hat{\mathbf{y}}}$ and $\sigma_{\mathbf{y}\hat{\mathbf{y}}}$ the empirical standard deviations and covariance of $\mathbf{y}$ and $\hat{\mathbf{y}}$.

3.3.2. Scoring rules

Scoring rules have been defined to assess the quality of a probabilistic forecast (Gneiting & Raftery, 2007; Lavigne et al., 2012). Since the forecast is probabilistic, it can be represented by its cumulative distribution function F_s for an observation y_s. The continuous ranked probability score at location s is defined by

$$\mathrm{CRPS}(F_s, y_s) = -\int_{-\infty}^{+\infty}(F_s(x) - \mathbb{1}_{x \geqslant y_s})^2 \mathrm{dx}$$

where $\mathbb{1}_{x \geqslant y_s}$ takes value 1 if $x \geqslant y_s$ and 0 otherwise. Solutions to this integral can be hard to compute. Fortunately, the CRPS can be expressed in a readily computable expression as

$$\mathrm{CRPS}(F, y_s) = \frac{1}{2}\mathbb{E}_F|Y - Y'| - \mathbb{E}_F|Y - y_s|,$$

where Y and Y' are independent variables distributed according to F_s. In a spatial setting, a global CRPS criterion can be calculated by averaging $\mathrm{CRPS}(F, y_s)$ over all locations s.

3.3.3. ROC Analysis

For decision making, the most important feature of a dispersal model is not necessarily its ability to predict the actual outcome of the dispersal with the lowest possible error but rather to predict accurately whether or not the result is above or below a given threshold. The problem is therefore a classification problem.

Receiver operating characteristic (ROC) curves are often used as a means to evaluate diagnostic tests for decision making (Hanley & McNeil, 1982). ROC analysis is a procedure derived from statistical decision theory, that was developed in the context of electronic signal detection. It has become widely used in agronomic applications to asses the accuracy of a classification procedure based on a quantitative diagnostic or model (see Makowski et al., 2008, for instance).

The ROC curve represents a plot of sensitivity values as a function of (1−specificity) values, where sensitivity is the rate of true positives and 1−specificity is the rate of false positives. The area under the ROC curve (AUC) is a popular index of the overall performance of a test. This synthetic index is equal to one if the predicted classification is perfect (i.e. no differences between the real observed classification and the classification obtained with model predictions) and equal to 0.5 if the classification is not better than a random classification.

3.3.4. Global scale evaluation

The five criteria (r, RMSE, R^2, CRPS and AUC) can be calculated on the training dataset for each model, using the parameter posterior distribution estimated from the same dataset. This measures the fitting quality. The same criteria can also be calculated on the validation dataset. This provides an estimation of the quality for local prediction.

In practical applications it may be more interesting to estimate the prediction quality at more global scales. This can be done by applying the quality criteria to observed and predicted averages over areas of interest, as presented in Section 4.6.

4. Case study on gene flow

In this section, a case study on maize pollen flow in the context of coexistence between genetically modified (GM) and non-GM crops is presented. In these experiments, dispersal of a dominant phenotypic marker (blue color grains) from a patch of plants located at the center of a field of unmarked plants (yellow grains) was monitored. Although these data are restricted to contiguous plots with surrounding receptors, they provide a very good description of the gene flow processes mainly thanks to a large sampling effort. In addition, we evaluate the prediction accuracy on an independent set of data with the same characteristics (dominant phenotypic marker, contiguous plots with surrounding receptors) collated in Catalonia (Spain).

4.1. Experimentation

The training dataset comes from an experiment that was described in Klein et al. (2003) and reconsidered in Larédo & Grimaud (2007). Experiments were performed during Summer 1998 and 1999 near Montargis (France). A maize field measuring 120×120 m was sown in a production design: 160 rows 0.8 m apart each containing 800 plants 0.15 m apart. A central plot measuring 20×20 m was sown with plants producing blue coloured seed and the rest of the field contained yellow seed maize. The blue maize was a variety isogenic to the yellow one and homozygous for the "blue" allele. The blue colour is coded by the anthocyanin complex, which behaves as a monogenic dominant marker. All emitting plants were homozygous at the loci coding for the seed colour. Checks in the field and on control crosses did not reveal any systematic difference in pollen production and pollen efficiency between plants producing blue or yellow seeds.

In 1998 and 1999, pollen dispersal began on July 18 and on July 17 respectively. Both blue and yellow plants flowered almost synchronously: blue male maize started blooming on July 19 and July 18, respectively while yellow maize started to bloom on July 20 and July 17. Dispersal lasted 14 days and ended on August 1 (July 30 in 1999). A total of $N_1 = 2937$ cobs in 1998 and $N_2 = 4430$

cobs in 1999 were sampled on a rectangular grid. The grid consisted of 101 rows (every row for the 72 rows centered on the central plot and every third row elsewhere) and 31 plants on each row (one every 4 meters). The number y_s of blue grains on the cob sampled at location s was then determined (Figure 1). The total number K of seeds per cob was considered constant and estimated by counting the total number of seeds on 34 randomly chosen cobs (mean 394 and standard deviation 65).

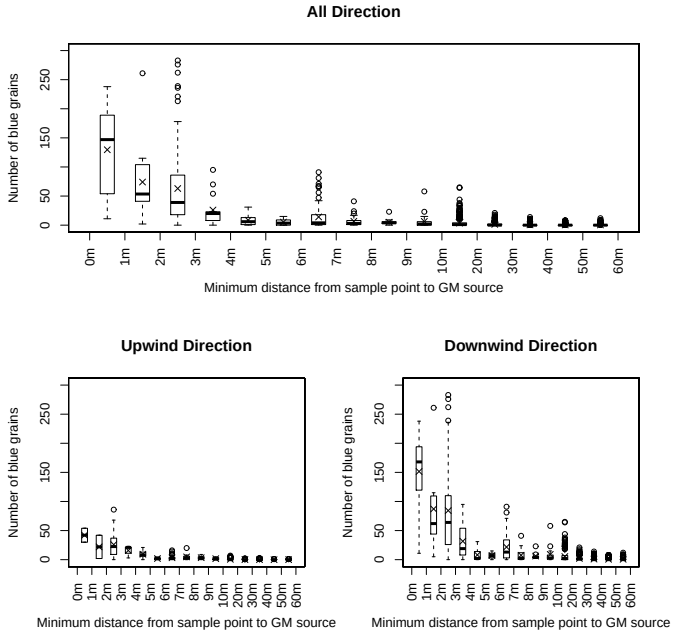


Figure 1: Boxplot of the number of blue grains on each sampled cob as a function of the distance from the sampling point to the closest GM pollen source. Crosses correspond to the average number of blue grains by distance interval.

We used a second and independent dataset called Mas Cebria for cross-validation. This dataset comes from an experiment performed in 2004 at Pla de Foixà, Girona (Spain) and is fully described in Palaudelmàs et al. (2012). In this trial, four different Bt yellow commercial hybrids derived from event Mon810 (Monsanto Co) were sown in the centre of a field, along an axis following the dominant wind direction and forming rectangles of increasing areas. Nontransgenic white kernel hybrid maize (Pioneer Hi Bred) was sown in the surrounding area to fill a total area of approximately 27 ha. The assessment of gene flow was carried out by counting the number of yellow kernels in white cobs. Figure 2 shows the sampling plan for all defined datasets.

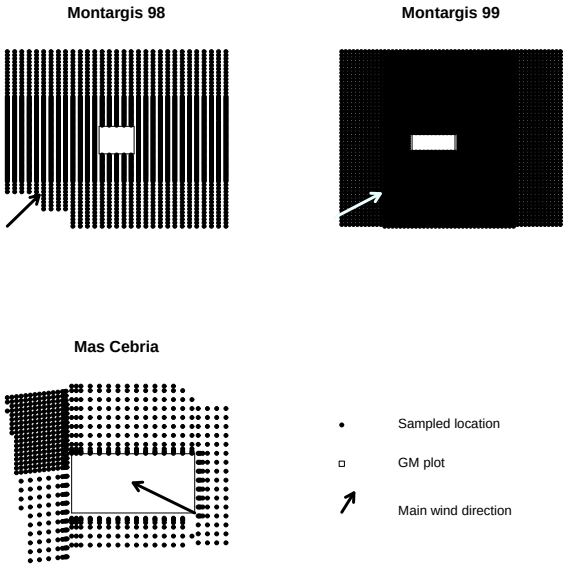


Figure 2: Sampling plans used in the three experiments.

The Mas Cebria dataset allows an unbiased quantification of predictive accuracy. It is of particular interest since it was obtained under very different conditions compared to Montargis in terms of envirionmental conditions, scale, sampling effort and sampling scheme (see Palaudelmàs et al., 2012, for details). Another great interest of this dataset is the fact that it has been used in other studies (see Allnutt et al., 2008, for instance) for the same validation purpose.

4.2. Meteorological Data

The Meteo France station nearest to the experiment in Montargis was 70 km West of the field (Orléans) in 1998. Wind direction and intensity were collected 10 m above ground at 3 hour intervals by Meteo France. We then calculated the distribution of wind direction over the pollination period from wind data between 8 a.m. and 7 p.m., when pollination occurs, and derived the main wind direction. In 1999, a meteorological station was placed inside the maize field. A comparison between data from the Orléans station and the local data was performed and showed little difference over the 15-day dispersal period. This indicated that data from Meteo France in 1998 could be used with confidence.

4.3. Model adaptation

The equations of the individual and global dispersal frameworks must be adapted to the case study. Indeed in this special case there are two types of particle sources A (e.g. GM) and B (e.g. non-GM) and the expected observed outcome λ_s is equal to $K\mu_s$, where K is the number of seeds per cob and μ_s is the expected rate of A pollen arriving on location s. Thus the expected outcome at location s is given by

$$\lambda_s = K\frac{\sum_{s'\in A}\gamma(s,s')}{\sum_{s'\in A}\gamma(s,s')+\sum_{s'\in B}\gamma(s,s')}. \quad (12)$$

in the individual framework. For the global dispersion framework, as in Damgaard & Kjellson (2005), we defined the expected outcome simply by

$$\lambda_s = K\sum_{s''}\gamma(s,s''), \quad (13)$$

where s'' is the A location closest to s.

Note that the individual framework appears particularly relevant in the case when recepting sites are also emitting sites, for instance the case of pollen mediated gene flow when emitting sites include plants that are also recepting sites (Angevin et al., 2008). However, as mentioned before, it is time-consuming, especially when the dispersal calculations have to be repeated thousands of time within the iterations of an MCMC algorithm. So in order to perform Bayesian estimation of the parameters of this function, which needs intensive call to that function, we derived an approximation method for the integration stage (details can be found in Appendix) similarly to Bouvier et al. (2009).

4.4. Prior Distributions

In the Bayesian framework, prior distributions must be assigned to each parameter in order to perform estimation of the posterior distribution. Priors for the parameters of the observation models, namely β_1, β_2 and σ^2 are listed in Table 2. No information was available for the dispersal parameters except for bounds on their possible values, so we chose fairly non-informative prior distributions. Note in particular that a_2 in the Exp function was constrained to be lower than a_1 to reflect the quick decrease of cross pollination in the first meters and slower decrease afterwards (as it is often observed in datasets). The parameter K_e in the Exp function, which reflects the cross pollination rate at distance 0, was constrained to lie between 0 and 1. We chose to consider uniform distributions within such bounds for all parameters of the defined dispersal functions. The prior distributions are summarized in Table 3.

Table 2: Prior distributions for the ZI$\mathcal{P}$ and random expectation models

Parameter	Distribution
β_1	$\mathcal{U}(0, 10)$
β_2	$\mathcal{U}(-150, 150)$
σ^2	$InvGamma(0.001, 0.001)$

Table 3: Prior distributions used for parameters of dispersal functions. DF: dispersal function, SD: standard deviation, CV: coefficient of variation (=SD/Mean)

IDF	Parameter	Lower bound	Upper bound	Mean	SD	CV
2Dt	a	0	20	10	100/3	10/3
2Dt	b	1	2	1.5	0.29	0.19
2Dt	κ	0	3	1.5	0.75	0.5
Exponential	Ke	0	1	0.5	0.29	0.58
Exponential	a_1	0	2	1	0.58	0.58
Exponential	a_2	0	a_1	$a_1/2$	$a_1/\sqrt{12}$	0.58
Exponential	D	1	10	5.5	2.6	0.47
Exponential	κ	0	3	1.5	0.75	0.5

4.5. Bayesian inference details

The burn-in was set to 2×10^4 iterations, the convergence of the MCMC algorithm was checked by visually analysing three independent MCMC chains using three different initial values for parameters. Gelman and Rubin convergence criterion (Gelman & Rubin, 1992) was also calculated and checked. It indicated that we could assume convergence of the Markov Chain to the posterior distribution with the 150,000 iterations following the burn-in period. We then thinned the chain by using only one value out of every 25 in the Markov chain. Thinning reduces autocorrelations in the chain and reduces also computing time when using the posterior distribution to make predictions. Overall, that left a sample of 6000 parameter vectors to represent the posterior distribution $P(\theta|Y)$.

4.6. Measure of quality of prediction at a global scale

In order to perceive models behaviour at field scale in a *multi-field* setting, 18 pseudo-plots (with about thirty sampled plants each) were defined on the validation dataset. Predictions as well as observations were averaged over those pseudo-plots. This provided two vectors of length 18 that were used to calculate the criteria defined above (r, RMSE, R^2, CRPS and AUC) at pseudo-plot scale. To keep in mind the general objective to take variability into account, a sixth criterion was defined as the linear correlation between observed and predicted standard deviations at pseudo-plot scale. Let $\mathbf{z} = (z_1, z_2, \ldots, z_{18})$ be the vector of the observed pseudo-plots (averages over raw data) and $\hat{\mathbf{z}} = \{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{18}\}$ be the vector of predicted mean responses (averages over raw precdictions):

$$r_\sigma = \frac{\sigma_{\sigma_{\mathbf{z}\hat{\mathbf{z}}}}}{\sigma_{\sigma_{\mathbf{z}}}\sigma_{\sigma_{\hat{\mathbf{z}}}}}.$$

5. Results

5.1. Performance of the Bayesian Estimation Algorithmn

The Bayesian estimation algorithms performed well. The Gelman and Rubin convergence statistics (Gelman & Rubin, 1992) calculated for each model were

always below 1.1. This means that we can assume convergence of the estimation algorithm to parameters posterior distribution with confidence. Computation time for full model estimation ranged from three hours (Global framework) to four days (Individual framework) on a computer grid (3Ghz, 64Go Ram). For all the parameters, the posterior distribution was much narrower than the prior. This indicates that the data allowed us to considerably reduce our uncertainty about those parameters. Posterior distribution of parameters and joint distribution of the main parameters can be found in Supplementary Information.

5.2. Model evaluation and selection

Tables 4, 5 and 6 present the criteria values for all 16 models, calculated with the training, validation and pseudo-plot validation datasets respectively. Models with high r, R^2, AUC values have low RMSE and CRPS values, that is to say there is no contradiction between criteria. In addition, good models for adjustment are also good for prediction on an independent dataset. The criteria values for other combinations of training and validation datasets can be found in Supplementary Information.

A careful examination of both tables 5 and 6 shows that two models stand out from the others, namely *GZExpoA* and *GZExpoB*. The *IPExpoB*, *GPExpoB* and *IP2DtA* models follow but do not maximise (or minimise) the same criteria. For instance, the *IP2DtA* model maximises the correlation between observations and predictions for an independent dataset (even when prediction are averaged over peudo-plot) and this is the most important criterion for the output of a gene-flow model to maximise in order to design cost-effective sampling strategies. So the *IP2DtA* model can be a good candidate for sampling purposes whereas it would be a very poor choice to predict the real cross pollination rate. Indeed for such a purpose, it is more interesting to look at the RMSE criterion which for this model is more than twice the RMSE of the best model. The same type of reasoning can be made about the *IZExpoB* model which could be used for both purposes (sampling and prediction), since it exhibits high correlation and very low RMSE but only if calibrated in the area of interest; indeed, its good

performances can only be found with the training dataset.

From all these observations, we conclude that the exponential dispersal function, despite its simplicity, is well adapted to prediction purposes. In the same way, the ZIP model seemed to be better than the Poisson model. The individual dispersal framework, despite its better description of the system, did not provide more accurate predictions.

Overall, the model components that most influenced the responses were the dispersal framework, the dispersal function and the probability distribution of the observations (Poisson or zero-inflated). The fourth factor (i.e fixed or random expectations of the cross pollination rate) did not play a major role regarding the statistical criteria based on the mean response only. However the choice of such a model component may have a strong influence on particular features of interest when making probabilistic predictions. In the following, we concentrate the presentation of the results on the *GZExpoA* and *GZExpoB* models, in particular to investigate the relatively subtle role of the difference between those two models.

Table 4: Criteria values for all models calculated with the training dataset (Montargis 98-99). CRPS: continuous ranked probability score, r: correlation, RMSE: root mean squared error, R^2 coefficient of determination, AUC: area under curve.

CodeName	CRPS	r	RMSE	R^2	AUC
GPExpoA	-1.955	0.717	10.39	0.513	0.9079
GPExpoB	-2.36	0.377	15.01	-0.016	0.9065
GP2DtA	-2.113	0.608	12.12	0.338	0.9087
GP2DtB	-1.761	0.625	11.67	0.386	0.9097
GZExpoA	-1.704	0.72	10.64	0.49	0.9065
GZExpoB	-1.637	0.724	10.66	0.487	0.9077
GZ2DtA	-1.945	0.605	12.54	0.291	0.9024
GZ2DtB	-2	0.489	13.35	0.197	0.9036
IPExpoA	-1.959	0.716	10.44	0.509	0.8909
IPExpoB	-2.344	0.393	14.87	0.003	0.8974
IP2DtA	-1.995	0.712	10.48	0.505	0.8929
IP2DtB	-1.639	0.713	10.57	0.497	0.8937
IZExpoA	-1.742	0.712	11.16	0.439	0.8946
IZExpoB	-1.628	0.719	10.59	0.495	0.8908
IZ2DtA	-1.785	0.699	11.19	0.436	0.8985
IZ2DtB	-1.782	0.664	11.99	0.352	0.8952

5.3. Comparison of observed and predicted responses

5.3.1. Overall predicted variability

The ability of the *GZExpoA* and the *GZExpoB* models to reproduce the overall variability observed in the data can be assessed on Figures 3 and 4. They show boxplots of observed and predicted cross pollination rates for different classes of distances and different orientations. The decomposition of observations and predictions in two directions (upwind and downwind) was made to identify the *real* variability of the response and not the artefact that is due to the anisotropic nature of process when looking at all directions. The overall variability observed in the data is much better reproduced by the predictions of the *GZExpoB* model. This behaviour being always recovered when switching the distribution of the expected cross pollination rate from fixed to normal, all else being equal. This indicates that giving a normal distribution to λ' allows to add the right amount of extra variability to the prediction.

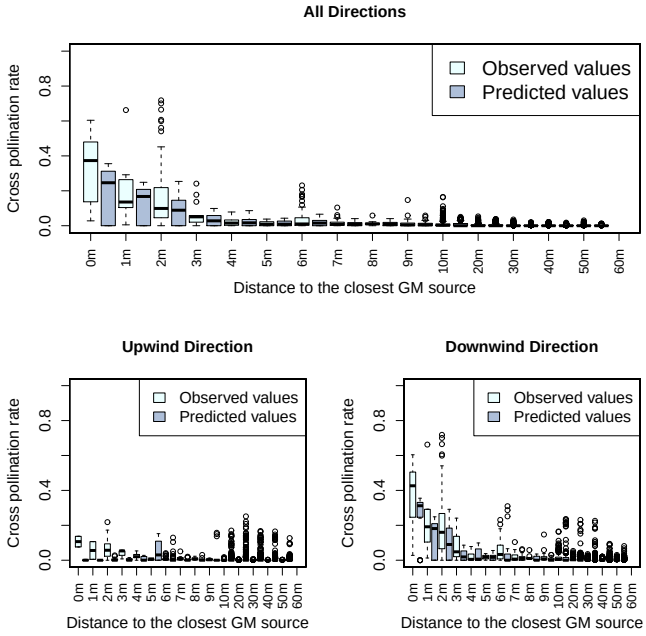


Figure 3: Boxplots of observed and predicted cross pollination rates (*GZExpoA* model) as a function of distance to the closest GM source.

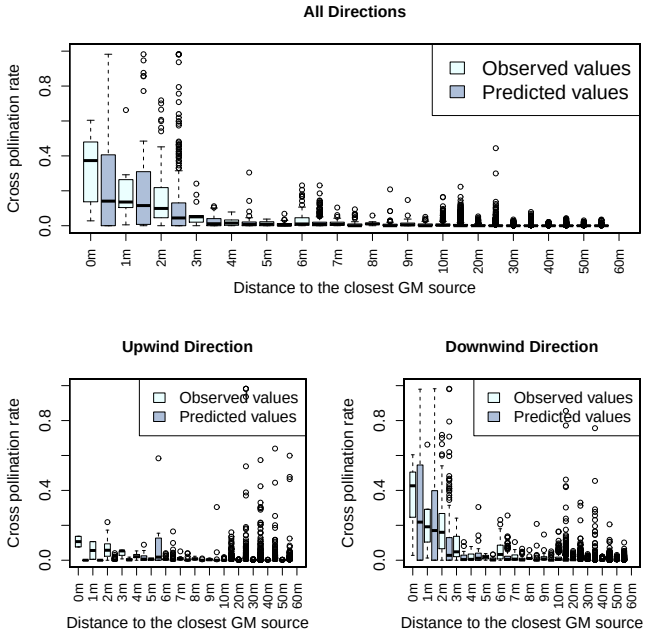


Figure 4: Boxplots of observed and predicted cross pollination rates (*GZExpoB* model) as a function of distance to the closest GM source.

5.3.2. Predicted accuracy with respect to distance

Figure 5 shows boxplots of predicted mean and standard deviation for different classes of distances. Both models tend to underestimate the observed mean in the first meters but only the model with a fixed expectation underestimates the standard deviation. Conversely, models with random expectation overestimate the standard deviation in the very first meters. This remark allows to conclude that the choice of a model for the expected cross pollination rate has a great impact on the predicted mean and standard deviation. The consequences may differ depending on the prediction purpose. For example if one wants to sample plants in a given field to have a more reliable estimate of cross-fertilzation, a random expectation model should be used to give a better description of the intra-field variability. Whereas if one wants to have an accurate prediction without any supplementary field sample, the fixed expectation model would be sufficient.

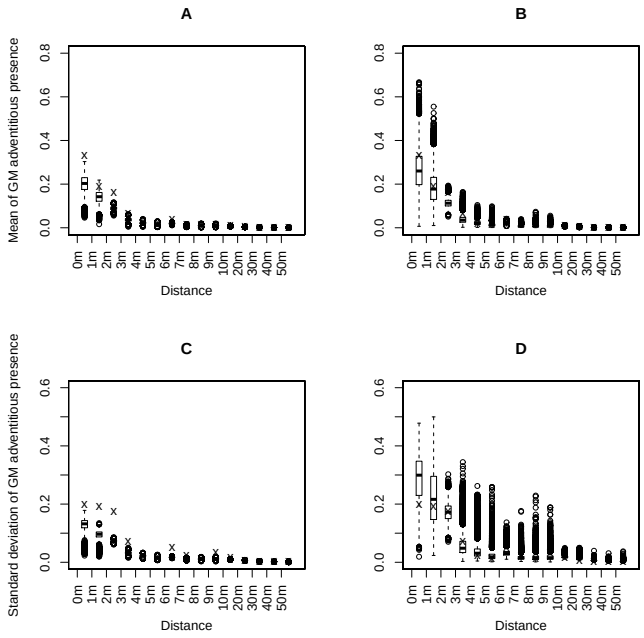


Figure 5: Boxplot of mean (A, B) and standard deviation (C, D) of the predicted cross pollination rates as a function of distance to the closest GM source. Crosses correspond to the average (A, B) or standard deviation(C, D) of observed number of blue grains by distance interval. Plots A and C (respectively B and C) represents summary statistics of predictions derived from the *GZExpoA* model (respectively *GZExpoB* model).

5.4. Model's predictive ability

Cross-validation was performed on all the defined models by predicting the Mas Cebria dataset using the posterior parameter distribution evaluated on the Montargis datasets.

5.4.1. At plant scale

The statistical behaviours of the best models for validation, namely *GPExpoA*, *GP2DtB GZExpoA* and *GZExpoB* were similar (see table 5). The RMSE at plant scale, which includes a high variability between plants, amounts to around plus or minus twenty grains carrying the transgene which corresponds to a cross-pollination rate of about 5%. This value is twice higher than the performance of the same models on the training dataset. In terms of correlation, the best models reached 0.744 and were always around 0.7. This is promising since another study (see Allnutt et al., 2008, for details) with relative close purpose and the same dataset for validation (i.e. Mas Cebria) hardly reached 0.6.

Table 5: Criteria values for all models calculated with the validation dataset (Mas Cebria). CRPS: continuous ranked probability score, r: correlation, RMSE: root mean squared error, R^2 coefficient of determination, AUC: area under curve.

CodeName	CRPS	r	RMSE	R^2	AUC
GPExpoA	-7.157	0.744	21.27	0.434	0.9077
GPExpoB	-9.5	0.487	24.75	0.234	0.8533
GP2DtA	-8.462	0.675	24.47	0.251	0.9201
GP2DtB	-7.738	0.684	22.08	0.39	0.9187
GZExpoA	-6.632	0.753	19.41	0.529	0.9073
GZExpoB	-7.178	0.739	19.86	0.507	0.9091
GZ2DtA	-8.662	0.675	25.74	0.172	0.91
GZ2DtB	-8.434	0.666	25.75	0.171	0.9131
IPExpoA	-39.81	0.662	68.85	-4.926	0.8796
IPExpoB	-8.991	0.556	23.54	0.307	0.8936
IP2DtA	-27.12	0.726	48.03	-1.884	0.8766
IP2DtB	-9.474	0.687	40.62	-1.063	0.8742
IZExpoA	-24.701	0.657	71.03	-5.307	0.9012
IZExpoB	-16.263	0.618	71.44	-5.38	0.9002
IZ2DtA	-20.215	0.682	56.61	-3.006	0.9015
IZ2DtB	-15.021	0.602	66.94	-4.602	0.8998

5.4.2. At field scale

In order to determine if models' good performances at plant scale were consistent with performance that the same models could have at field scale, a grouping of close plants was made to emulate several receptor fields (18 pseudo-plots in all). The statistical criteria were then calculated between the means of the observed groups of plants and the means of the predicted groups of plants (Table 6). Note that at this scale, prediction uncertainty is necessarily reduced as we average over several predictions (similarly to Gouache et al., 2013), which may modify the ranking of the models. The statistical performance of the best models for validation on pseudo-plot, namely *IPExpoB*, *GPExpoB* *GZExpoA* and *GZExpoB* were generally similar with correlation around 0.85 and RMSE ranging from 6 to 9 grains carrying the transgene which corresponds to a cross-pollination rate ranging from 1.5% to 2.5%. Note that *GZExpoA* and *GZExpoB* were among the best models also at plant scale which makes them the most reliable models overall.

Table 6: Criteria values for all models calculated with the validation dataset (Mas Cebria) on pseudo-plots. CRPS: continuous ranked probability score, r_μ, r_σ: correlation coefficents for the mean and standard deviation, RMSE: root mean squared error, R^2 coefficient of determination, AUC: area under curve.

CodeName	$CRPS$	r_μ	r_σ	$RMSE$	R^2	AUC
GPExpoA	-6.341	0.854	0.7873	9.4	0.511	0.9625
GPExpoB	-4.61	0.84	0.7545	7.4	0.697	0.7375
GP2DtA	-8.457	0.838	0.7943	12.1	0.19	0.95
GP2DtB	-6.149	0.849	0.8135	9.37	0.514	0.9625
GZExpoA	-5.859	0.86	0.8052	8.69	0.582	0.9625
GZExpoB	-5.853	0.853	0.8043	8.88	0.564	0.95
GZ2DtA	-9.202	0.835	0.7882	13.49	-0.008	0.95
GZ2DtB	-7.62	0.886	0.8807	12.41	0.147	0.9625
IPExpoA	-43.85	0.898	0.8984	59.26	-18.432	0.9625
IPExpoB	-3.379	0.879	0.919	6.65	0.755	0.95
IP2DtA	-30.458	0.898	0.9168	43.3	-9.378	0.95
IP2DtB	-20.211	0.892	0.924	36.38	-6.325	0.95
IZExpoA	-38.628	0.898	0.8934	62.81	-20.835	0.95
IZExpoB	-36.984	0.888	0.8935	63.51	-21.323	0.95
IZ2DtA	-29.751	0.902	0.9136	49.8	-12.723	0.95
IZ2DtB	-33.634	0.884	0.8945	58.93	-18.219	0.95

6. Discussion

We have presented a Bayesian method to model dispersal using spatial configuration and climatic data (distances between emettors/receptor and main wind direction) while accounting for uncertainty, with an application to the prediction of adventitious presence rate of GM in a nonGM field. To the best of our knowledge, the characteristics of data that were used in our case study are representative of typical data encountered in other dispersal experiments or ecological monitoring. As a consequence, this approach can apply to multiple situations in agronomy (pollen mediated gene flow), epidemiology (disease spread via spores), ecology (rare species monitoring), and in any dispersal process of particle over space.

The proposed method includes the design of candidate models, their calibration, selection and evaluation on an independent dataset. These ingredients are well known but they are often presented in a well defined context with many datasets and much homogeneity between the test and validation data, for example. Here a group of models was identified that is sufficiently parsimonious and robust to be used for prediction purposes in less ideal situations. The group of models allows to include local information and it reflects reliably enough the observed variability in the data so that probabilistic model predictions can be performed and used to quantify risk under different scenarios or derive optimal sampling schemes. Of course the group of models must be adapted to each particular application. However, we believe that the model components presented in Section 2 represent the basic choices that must be made for the dispersal model in many agro-environmental applications.

As monitoring for coexistence may present a significant cost for non-GM farmers (Messéan et al., 2009), the selected models in our case study could be used to predict the cross pollination rate in various situations. *i)* before sowing (with assumption about main wind direction) to identify those situations with high risks of exceeding the legal threshold and, consequently, help farmers make decision about where to allocate GM maize fields. *ii)* before flowering, with a

reduced prediction uncertainty as the main wind direction is less uncertain at that period. *iii)* before harvesting either to make coherent batches (with similar levels of adventitious presence), or, if prediction uncertainty remains too high, to decide where the sampling should be done.

Beyond the special case of this paper, a dispersal model can be used for different objectives such as: *i)* to predict the mean value of a target area; *ii)* to establish a ranking of target areas with respect to a given threshold; *iii)* to decide where, when and how to sample in the target area in order to get more reliable estimates. Thus it is useful to assess and compare models with respect to several criteria. Based on the prediction purpose, one would typically not look at the same criterion to decide what model should be used: the best model to predict the mean value of a target area is the one that minimises the RMSE. To establish a ranking of target areas, one would look for the model to maximises the AUC. Finally, the best model to determine an optimal sampling scheme is the model that maximises correlation between observed and predicted values. *iv)* Given that special emphasis was put on reproducing the variability of the cross pollination rate, our models can be used not only to predict the mean cross pollination rate of a field but also to identify areas of greatest variability in the field in order to determine optimal sampling scheme (see Bancal et al., 2013, for examples of application).

The relative loss of accuracy on the predicted mean when using a random expectation model is offset by a gain on the accuracy of the predicted variability. This information is beneficial for decision making, depending on the purpose of the decision (e.g. before sowing, to choose where to sow GM maize in order to keep cross pollination rate below the legal threshold or before harvest, when GM plots are sown, either to define coherent batches, or to design sampling strategies). Indeed, depending on the objective of the end-user, the best model may not be the same. If one wants an accurate prediction of the mean cross pollination rate in a field, the *GZExpoA* model would be sufficient although it would not be a good choice if the prediction is close to 0.9% (which is the legal threshold in European Union). In this case, the most likely strategy for

a decision maker is to sample plants in the field to get more reliable estimates. And the best way to know where to sample in the field is to have a good approximation for the variability of the cross pollination rate within the field, so in this situation one would prefer the *GZExpoB* model.

Moreover, we observed a poor performance of Individual Dispersal Framework's models in their predictive ability when used on the validation dataset. While the correlation between observed and predicted values are similar between both types of models, these individual models overestimated the observed values and, consequently, their variability whereas the global framework models tended to underestimate the observed values. This shows *i)* that bias can arise even when using more complex models; *ii)* the importance of a proper validation on an independent dataset for each context and the definition of appropriate criteria depending on the purpose of the prediction. Indeed, the model that gives the best adjustement to the training dataset is not necessarily the best model to give recommandations in an independent situation.

It should be noted here that the size of the donor field was much greater in the validation dataset than in the training dataset (80 times larger). As the global framework considers areas as a single point, differences in sizes may explain the poor behaviour. When looking at (supplementary material), the outcomes of cross-validation of Montargis 98 on Montargis 99 (when sizes were similar) there was less discrepancy between the two frameworks. This indicates that the difference of the source sizes between the training and validation datasets have a great impact on the predictive ability and advocates for using a representative set of training datasets exhibiting different landscape patterns (which is one of the next steps of our research project).

Another point that needs to be aknowledged, is the number of observations needed to reach a given accuracy. In this paper, we have studied the case when few datasets area available but the amount of data within each dataset is very large. Nevertheless, an attractive lead would be to compare our calibrated models with the same models calibrated with a lot of datasets even if the amount of data in each is relatively low compare to those that were used here.

7. Conclusion and further work

In this paper we demonstrate the feasibility of using a Bayesian statistical framework to estimate parameters of a dispersal model and to make probabilistic predictions. This methodology can be helpful, for instance, to better inform coexistence by providing accurate predictive values of adventitious presence of GM material in conventional maize fields. The methodology may also apply to a broad range of issues in which dispersal is involved. In addition, we were able to account for the rather high variability of observations at the individual level, which makes it possible to design optimal stratified sampling schemes and leads to more accurate decisions.

This work is being pursued by enriching the range of gene flow models that could be used within this Bayesian framework but also with a more representative set of training datasets. Indeed a promising approach would be the use of experimental data exhibiting very contrasted landscape patterns even if sampling effort is less intensive and of a hierarchical Bayesian analysis so that model predictions can better adapt to the specificities of each situation.

Acknowledgements

This study was partially funded by the European Union project PRICE (PRactical Implementation of Coexistence in Europe), contract number 289157.

References

Allnutt, T., Dwyer, M., McMillan, J., Henry, C., & Langrell, S. (2008). Sampling and modeling for the quantification of adventitious genetically modified presence in maize. *Journal of agricultural and food chemistry*, *56*, 3232–3237.

Angevin, F., Klein, E. K., Choimet, C., Gauffreteau, A., Lavigne, C., Messéan, A., & Meynard, J. M. (2008). Modelling impacts of cropping systems and climate on maize cross-pollination in agricultural landscapes: The MAPOD model. *European Journal of Agronomy*, *28*, 471–484.

Bancal, R., Makowski, D., & Bensadoun, A. (2013). Comparison of sampling strategies to evaluate the rate of transgenic adventitious presence in agricultural fields. In *GMCC 2013 - 6th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains*. Lisbon, Portugal.

Bensadoun, A., Monod, H., Angevin, F., Makowski, D., & Messéan, A. (2013). Modeling of gene flow by a bayesian approach: A new perspective for decision support. In *GMCC 2013 - 6th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains*. Lisbon, Portugal.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43(1)*, 1–59.

Bouvier, A., Kiêu, K., Adamczyk, K., & Monod, H. (2009). Computation of the integrated flow of particles between polygons. *Environmental Modelling & Software*, *24*, 843–849.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian Methods for Data Analysis*. Texts in Statistical Science (Third ed.). Chapman and Hall/CRC.

Clark, J. S. (2005). Why environmental scientists are becoming bayesians. *Ecology Letters*, *8*, 2–14.

Clark, J. S., Silman, M., Kern, R., Macklin, E., & HilleRisLambers, J. (1999). Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology*, *80*, 1475–1494.

Cochran, W. (1977). *Sampling techniques*. John Wiley & Sons, third edition.

Colbach, N., Clermont-Dauphin, C., & Meynard, J. (2001). GeneSys: a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers II. genetic exchanges among volunteer and

cropped populations in a small region. *Agriculture, Ecosystems and Environment*, *83*, 255–270.

Cunningham, R., & Lindenmayer, D. (2005). Modeling count data of rare species: Some statistical issues. *ECOLOGY*, *86*, 1135–1142. doi:{10.1890/04-0589}.

Damgaard, C., & Kjellson, G. (2005). Gene flow of oilseed rape (*Brassica napus*) according to isolation distance and buffer zone. *Agriculture, Ecosystems and Environment*, *108*, 291–301.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gneiting, T., & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102(477)*, 359–378.

Goedhart, P. W., van der Voet, H., Baldacchino, F., & Arpaia, S. (2014). A statistical simulation model for field testing of non-target organisms in environmental risk assessment of genetically modified plants. *Ecology and Evolution*, *4*, 1267–1283. doi:10.1002/ece3.1019.

Gouache, D., Bensadoun, A., Brun, F., Pagé, C., Makowski, D., & Wallach, D. (2013). Modelling climate change impact on Septoria tritici blotch (STB) in France: Accounting for climate model and disease model uncertainty. *Agricultural and Forest Meteorology*, *170*, 242–252.

Gregoire, T., & Salas, C. (2009). Ratio estimation with measurement error with auxiliary variate. *Biometrics*, *65*, 590–598.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under the ROC curve. *Radiology*, *143*, 29–36.

Klein, E., Lavigne, C., Foueillassar, X., Gouyon, P., & Larédo, C. (2003). Corn pollen dispersal: Quasi-mechanistic models and field experiments. *Ecological Monographs*, *73*, 131–150.

Klein, E., Lavigne, C., Picault, H., Renard, M., & Gouyon, P. (2006a). Pollen dispersal of oilseed rape: estimation of the dispersal function and effects of field dimension. *Journal of Applied Ecology*, *43*, 141–151. doi:10.1111/j.1365-2664.2005.01108.x.

Klein, K., Lavigne, C., & Gouyon, P. (2006b). Mixing of propagules from discrete sources at long distance: comparing a dispersal tail to an exponential. *BMC Ecology*, *6:3*. URL: http://www.biomedcentral.com/1472-6785/6/3. doi:10.1186/1472-6785-6-3.

Krueger, T., Page, T., Hubacek, K., Smith, L., & Hiscock, K. (2012). The role of expert opinion in environmental modelling. *Environmental Modelling & Software*, *36*, 4–18.

Kuhnert, P., Martin, T., Mengersen, K., & Possinghma, H. (2005). Assessing the impacts of grazing levels on bird density in wookland habitat: a Bayesian approach using expert opinion. *Environmetrics*, *16*, 717–747.

Lambert, D. (1992). Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1–15.

Larédo, C., & Grimaud, A. (2007). Stochastic models and statistical inference for plant pollen dispersal. *Journal de la Société Française de Statistique*, *148*, 77–105.

Lavigne, A., Bel, L., Parent, E., & Eckert, N. (2012). A model for spatio-temporal clustering using multinomial probit regression: application to avalanche counts. *Environmetrics*, *23*, 522–534.

Lavigne, C., Klein, E., Mari, J.-F., Le Ber, F., Adamczyk, K., Monod, H., & Angevin, F. (2008). How do genetically modified (GM) crops contribute to background levels of GM pollen in an agricultural landscape? *Journal of Applied Ecology*, *45*, 1104–1113. doi:10.1111/j.1365-2664.2008.01504.x.

Lavigne, C., Klein, E., Vallée, P., Pierre, J., Godelle, B., & Renard, M. (1998). A pollen-dispersal experiment with transgenic oilseed rape. estimation of the

average pollen dispersal of an individual plant within a fiel. *Theoretical and Applied Genetics*, *96*, 886–896.

Lecomte, J., Benoît, H., Etienne, M., Bel, L., & Parent, E. (2013). Modeling the habitat associations and spatial distribution of benthic macroinvertebrates: A hierarchical bayesian model for zero-inflated biomass data. *Ecological Modelling*, *265*, 74 – 84. URL: http://www.sciencedirect.com/science/article/pii/S0304380013003013. doi:http://dx.doi.org/10.1016/j.ecolmodel.2013.06.017.

Lenser, T., & Constable, A. (2007). A nonparametric algorithm to model movement between polygon subdomains in a spatially explicit ecosystem model. *Ecological Modelling*, *206*, 79–92. doi:http://dx.doi.org/10.1016/j.ecolmodel.2007.03.021.

Lewin, W., Freyhof, J., Huckstorf, V., Mehner, T., & Wolter, C. (2010). When no catches matter: Coping with zeros in environmental assessments. *ECOLOGICAL INDICATORS*, *10*, 572–583.

Makowski, D., Denis, J. B., Ruck, L., & Penaud, A. (2008). A Bayesian approach to assess the accuracy of a diagnostic test based on plant disease measurement. *Crop Protection*, *27*, 1187–1193.

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *PNAS*, *100*, 15324–15328.

Messéan, A., Squire, G., Perry, F., J. Angevin, Gomez, P., M. Townend, Sauuse, C., Breckling, B., Langrell, S., Dzeroski, S., & Sweet, J. (2009). Sustainable introduction of GM crops into european agriculture: a summary report of the FP6 SIGMEA research project. *OCL*, *16*, 37–51.

Messeguer, J., Peñas, G., Ballester, J., Bas, M., Serra, J., Salvia, J., Palaudelmàs, M., & Melé, E. (2006). Pollen-mediated gene flow in maize in real situation of coexistence. *Plant Biotechnology Journal*, *4*, 633–645.

Moreno-Jiménez, A., & Hodgart, R. (2003). Modelling a single type of environmental impact from an obnoxious transport activity: implementing locational analysis with GIS. *Environment and Planning*, *35*, 931–946.

Ojo, T., Bonner, J. S., & A., P. C. (2007). Simulation of constituent transport using a reduced 3d constituent transport model (ctm) driven by {HF} radar: Model application and error analysis. *Environmental Modelling & Software*, *22*, 488 – 501. doi:http://dx.doi.org/10.1016/j.envsoft.2006.02.010.

Palaudelmàs, M., Melé, E., Monfort, A., Serra, J., Salvia, J., & Messeguer, J. (2012). Assessment of the influence of field size on maize gene flow using SSR analysis. *Transgenic Research*, *21*, 471–483. doi:10.1007/s11248-011-9549-z.

Papaïx, J., Burdon, J., Lannou, C., & Thrall, P. (2014). Evolution of pathogen specialisation in a host metapopulation: Joint effects of host and pathogen dispersal. *PLoS Comput Biol*, *10*, e1003633. URL: http://dx.doi.org/10.1371%2Fjournal.pcbi.1003633. doi:10.1371/journal.pcbi.1003633.

Papaïx, J., David, O., Lannou, C., & Monod, H. (2013). Dynamics of adaptation in spatially heterogeneous metapopulations. *PLoS ONE*, *8*, e54697. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0054697. doi:10.1371/journal.pone.0054697.

Papaïx, J., Goyeau, H., Du Cheyron, P., Monod, H., & Lannou, C. (2011). Influence of cultivated landscape composition on variety resistance: an assessment based on wheat leaf rust epidemics. *New Phytologist*, *191*, 1095–1107. URL: http://dx.doi.org/10.1111/j.1469-8137.2011.03764.x. doi:10.1111/j.1469-8137.2011.03764.x.

Philibert, A., Desprez-Loustau, M., Fabre, B., Frey, P., Halkett, F., Husson, C., Lung-Escarmant, B., Marçais, B., Robin, C., Vacher, C., & Makowski, D. (2011). Predicting invasion success of forest pathogenic

fungi from species traits. *Journal of Applied Ecology*, *48*, 1381–1390. URL: http://dx.doi.org/10.1111/j.1365-2664.2011.02039.x. doi:10.1111/j.1365-2664.2011.02039.x.

Plummer, M. (2012). *JAGS Version 3.3 user manual*. URL: http://www-fis.iarc.fr/~martyn/software/jags/.

Plummer, M., Best, N., & Cowles, K., K. abd Vines (2009). *The coda Package: Output analysis and diagnostics for MCMC*. URL: http://www.R-project.org.

Ramin, M., Stremilov, S., Labencki, T., Gudimov, A., Boyd, D., & Arhonditsis, G. (2011). Integration of numerical modeling and Bayesian analysis for setting water quality criteria in Hamilton, Ontario, Canada. *Environmental Modelling & Software*, *26*, 337–353.

Rasmussen, R., & Hamilton, G. (2012). An approximate bayesian computation approach for estimating parameters of complex environmental processes in a Cellular Automata. *Environmental Modelling & Software*, *29*, 1–10.

Sileshi, G. (2008). The excess-zero problem in soil animal count data and choice of appropriate models for statistical inference. *PEDOBIOLOGIA*, *52*, 1–17.

Singh, S., Ash, G., & Hodda, M. (2014). Keeping 'one step ahead' of invasive species: using an integrated framework to screen and target species for detailed biosecurity risk assessment. *Biological Invasions*, (pp. 1–18). doi:10.1007/s10530-014-0776-0.

Appendix

Approximation method for the integration step of the individual framework

The individual dispersal framework is based on the calculation of the rate between GM and nonGM pollen arrivng at location s. This rate is defined as

$$\lambda_s = K \frac{\sum_{s' \in A} \gamma(s, s')}{\sum_{s' \in A} \gamma(s, s') + \sum_{s' \in B} \gamma(s, s')}, \tag{14}$$

where s is the location of a given receptor plant and s' are the locations of emettor plants. Bayesian estimation of the models parameters by MCMC requires intensive calls to the λ_s function, so that an approximation method is necessary to calculate λ_s fastly.

The main idea of the approximation relies on the fact that dispersal functions are very sharp. Consequently the influence of a pollen emettor on the expected cross pollination rate decreases very quickly with distance in such a way that only very close GM emettors significantly contribute to the actual cross pollination rate of a receptor. In other words, the closer the emettor, the more it contributes to pollination rate and the more accurate should be the calculus. So for a given receptor, the calculus of λ_s should take into account $\gamma(s, s')$ for almost all emettor plants s' in the very first meters around s. After those first meters, the emettor grid can be degraded by making the calculus of $\gamma(s, s')$ only for a fraction of emettor plants. Finally, from a certain distance each remaining GM source can be summarized by a single point. This general principle results in a unique grid with sparsity increasing with distance to the origin. This grid can be used for each receptor plant in combination with information on the GM or nonGM status of each neighbour plant.

After a simulation study we chose three radiuses ($R_1 = 3m$; $R_2 = 10m$; $R_3 = 20m$) and a fraction of emettors to be taken into account inside each radius ($F_1 = \frac{1}{4}$; $F_2 = \frac{1}{50}$; $F_3 = \frac{1}{500}$). Those parameters (radius and fraction) were optimised according to a factorial design (unshown results) and they allowed to define a unique sparse grid that was used for each receptor.

Annexes 2

Modeling of Gene Flow by a Bayesian Approach : A New Perspective for Decision Support. Bensadoun, A., Monod, H., Angevin, F., Makowski, D., Messéan, A. *AgBioForum*, 2014, 17(2) : 213–220.
The original publication is available at www.agbioforum.org

MODELING OF GENE FLOW BY A BAYESIAN APPROACH: A NEW PERSPECTIVE FOR DECISION SUPPORT

Arnaud Bensadoun[a,*], Hervé Monod[a], Frédérique Angevin[b], David Makowski[c], Antoine Messéan[b]

[a] INRA, UR 341 MIAJ, 78352 Jouy-en-Josas, France.

[b] INRA, UAR 1240 Eco-Innov, 78850 Thiverval-Grignon, France

[c] INRA, UMR 211 Agronomie, 78850 Thiverval-Grignon, France

*E-mail: *arnaud.bensadoun@jouy.inra.fr*, Tel: +33 1 34 65 28 51

Abstract

In the European debate about GMOs, the coexistence between GM and non-GM crops is a major stake. The regulatory coexistence measures currently considered by member states mostly rely on fixed separation distances at a national scale. Several spatially explicit modelling approaches have been studied to help determine these separation distances. However the formalism used in those models and the availibility of relevant and independant data for calibration and validation make the uncertainty analysis of those models almost impossible. The study presented here aims at developing an alternative model-based approach with emphasis on uncertainty to better adapt coexistence rules to any specific situation. The research work focuses on the use of bayesian methods to design a collection of statistical models at the scale of an agricultural landscape. Those models yield cross-pollination rate in non-GM fields and are flexible enough to adapt to the available in situ information.Thanks to the bayesian approach, estimates are computed as distributions whose dispersion depends on the amount and quality of available data; the more abundant and accurate the data, the narrower the distribution. In addition to model construction, we propose a coherent approach to select the best model for a given situation. The selection does not only rely on goodness of fit but also on the quality of the resulting decision for a given threshold. Models are already compatible with the decision support tool of the EU project PRICE.

Keywords: Bayesian methods, Coexistence, decision support, gene flow, pollen dispersal.

1 Introduction

The major crop in Europe and the second most widely-cultivated GM crops in the world after soybean is maize. Maize is one of the only GM crops commercially grown in Europe (with potato). Since maize is a cross pollinated crop relying on wind for the dispersal of its pollen, pollen flow between neighbouring maize fields is one of the major potential on-farm sources of adventitious mixing between GM and non-GM material (Devos et al., 2005). The cross-fertilization between GM and non-GM crops has been widely studied through measurements of pollen concentration and measurements of levels of cross-fertilization. Experimental data on gene flow for maize were collated and synthesized within the SIGMEA European research project (Messéan et al., 2009).

As stated before, GM maize is commercially grown in Europe (except in some countries, as in France), thus coexistence situations may occur. Coexistence refers to the ability of farmers and consumers to make a practical choice between conventional and genetically modified (GM) products, based on compliance with the legal obligation for labeling and/or purity standards (European Commission, 2003a). In Europe, up to 0.9% of GM material in non-GM food and feed is authorized, provided these traces of GMO are adventitious or otherwise technically unavoidable (European Commission 2003b). Above this threshold, in order to allow consumers to make a practical choice about the product, it must be labeled as consisting of, containing or being produced from a GMO.

In order to meet regulatory requirements, accurate prediction of maize gene flow is thus needed to assess risk of commingling between GM and non-GM crops. Moreover, tools are needed to help stakeholders of the maize supply chain to manage coexistence between GM and non-GM maize. In this context, considerable efforts have been made to model maize pollen dispersal with different modelling approaches. Spatially explicit and quasi-mechanistic models were defined (Colbach et al., 2001, Klein et al., 2003, Angevin et al., 2008) and tested in order to determine a legal separation distances between GM and non-GM maize field. However, the formalism used in those models and the availibility of relevant and independant data for calibration and validation make the uncertainty analysis of those models very difficult and computation time makes it totally impossible within a reasonable lapse of time.

This study aims at developing an alternative model-based approach to better adapt coex-

istence rules to the specificity of each situation. The research work focuses on the use of Bayesian methods to design a collection of statistical gene flow models. Those models yield cross-pollination rate in non-GM fields and have the particularity to be i) stochastic, so that a prediction is not represented by a single value but rather by a range of possible values with associated probability distribution, ii) adaptable to the level of *in situ* information. So that the model can be used with only spatial information (position of GM and non-GM fields), or with more information such as climatic variables if available in order to obtain more accurate distributions.

2 Material and Methods

2.1 Data

2.1.1 Experimentation

The training dataset used here comes from an experiment that was described in Klein et al. (2003) and reconsidered in Larédo and Grimaud (2007). The experimentation was performed during Summer 1998 near Montargis (France). A maize field measuring 120×120 m was sown in a production design: 160 rows 0.8 m apart each containing 800 plants 0.15 apart. A central plot measuring 20×20 m was sown with plants producing blue coloured seed and the rest of the field contained yellow seed maize. The blue maize was a variety close the yellow one, homozygous for the "blue" allele. The blue colour is coded by the anthocyanin complex, which behaves as a monogenic dominant marker. All plants were homozygous at the loci coding for the seed colour. Checks in the field and on control crosses did not reveal any systematic difference in pollen production and pollen efficiency between plants producing blue or yellow seeds.

The pollen dispersal began on July 18. Both blue and yellow plants flowered almost synchronously: blue maize began blooming on July 19 (male) and July 20 (female). Dispersal lasted 14 days and ended on August 1. The cobs were harvested and analysed on 16 October. A total of $N = 2937$ cobs were sampled on a rectangular grid. An amount of 101 rows was sampled (every row for the 72 rows centered on the central plot and every third row elsewhere) and 31 cobs on each row (every 4 meters). Sixty-four cobs could not be sampled in the West corner of the field. The number of blue grains (y_s) on each sampled cob was then determined (Figure 1). The total number K of seeds per cob was considered constant and estimated by counting the

total number of seeds on 34 randomly chosen cobs (mean 394 and standard deviation 65).

We also used an independant dataset for validation. This dataset comes from an experimentation performed in 1999 in the same place (Montargis, France) and with the same settings than the previous one (i.e. a central plot with plants producing blue coloured seed and the rest of the field containing yellow maize).

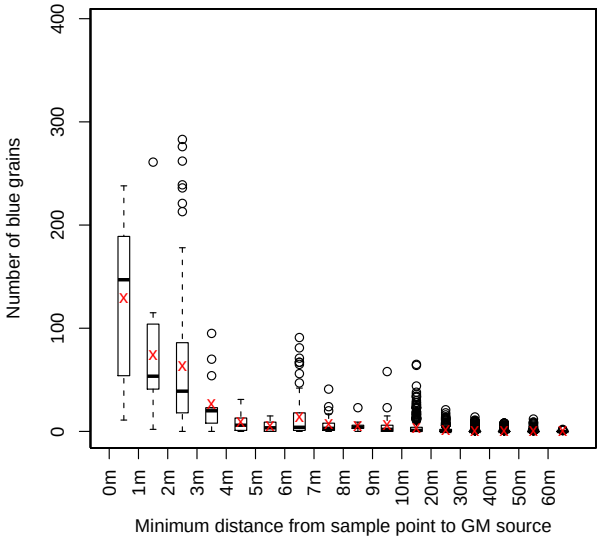


Figure 1: Boxplot of the number of blue grains on each sampled cob as a function of the distance from the sampling point to the closest GM pollen source. Red crosses correspond to the average number of blue grains by distance interval.

2.1.2 Meteorological Data

We used data for wind direction and intensity collected 10 m above ground at 3 hours interval by Meteo France. The meteorological station nearest to the experiment was 70 km West of the field (Orléans). We then calculated the distribution of wind direction over the pollination period from wind data between 8 a.m. and 7 p.m., when pollination occurs, and deduced the main wind direction. A comparison between data from Meteo France (70 km West of the field) and the local data resulting from a meteorological station located inside the maize field in 1999 showed little difference over the 15 days period dispersal.

2.2 Model

2.2.1 Observation Model

The scale of observation here is the non-GM plot and the variable to be predicted is the cross pollination rate on each plant of that plot. Let y_s denote the number of grains carrying the transgene in a cob located at point s and K the total number of grains of a sampled cob. d_s represents the minimum distance from the pollen source to the s^{th} sampling point. The Poisson distribution is often used to model counts of rare events Besag et al. (1991).

$$y_s \sim \mathcal{P}(K\mu_s^{'}), \tag{1}$$

where $\mu_s^{'}$ is a random variable (defined below) which represents the expectation of the cross pollination rate.

However counts of GM seed in a conventional maize cob exhibit a high variability, with a few exceptional cob having almost 100% of GM seeds and lots of cobs with zero GM seed. Therefore over dispersion with respect to the Poisson distribution is strong and especially an excess of zero. The Zero Inflated Poisson distribution (ZIP) is a good candidate to cope with that excess as it consists of a mixture of a Poisson and a Dirac distribution in zero. The ZIP assumes that, with probability p, the only possible observation is zero and, with probability $q = 1 - p$, the observation model is a Poisson distribution

$$y_s \sim ZI\mathcal{P}(1-q, K\mu_s^{'}). \tag{2}$$

To estimate the weight of the zeros in the ZIP mixture we defined Z to be a hidden variable distributed as a Bernoulli variable

$$Z_s \sim \mathcal{B}\text{ern}(q_s), \tag{3}$$

with

$$\begin{cases} y_s = 0 & \text{if } Z_s = 0 \\ y_s \sim \mathcal{P}(K\mu_s^{'}) & \text{if } Z_s = 1 \end{cases}$$

We consider a logit link to distance d_s from the closest pollen source to the sampling point

$$\text{logit}(q_s) = \beta_1(\beta_2 - d_s), \tag{4}$$

where β_2 is the abscissa of the inflection point and β_1 is equal to $-4\times$ the slope of the tangent to the logistic curve at the inflection point. Thus the *ZIP* model has two parameters (β_1 and β_2) whose priors are listed in Table 1.

In order to take into account the remaining variability observed in the data, we considered the expectation parameter of the Zero Inflated Poisson model to be a random variable

$$\mu_s^{'} \sim \mathcal{N}(\mu_s, \sigma^2). \tag{5}$$

where μ_s is the output of a dispersal function to be precisely defined below. This added only one parameter (σ^2) whose prior is also listed in Table 1.

Parameter	Distribution
β_1	$\mathcal{U}(0,10)$
β_2	$\mathcal{U}(-150,150)$
σ^2	$InvGamma(0.001,0.001)$

Table 1: Prior distributions For the *ZIP* and random expectation models

2.2.2 Individual Dispersal Functions and Dispersal Frameworks

An individual dispersal function $\gamma(s,s')$ is a four-dimensional probability density function. It gives the probability that a pollen grain emitted at point s' falls and pollinates a plant at point s (see Lavigne et al., 1998, Klein et al., 2003, for details).

Two frameworks have been defined to compute cross pollination rates from an individual dispersal function. The individual dispersal framework defined in Lavigne et al. (1998) also known as dispersal kernels framework (Klein et al., 2006a), allows one to compute with a given

dispersal kernel, for each plant (pixel) of the conventional field, the expected impurity rate (i.e. expected the proportion of GM grain). The proportion is computed as:

$$\mu_s = \frac{\sum_{s' \in GM} \gamma(s,s')}{\sum_{s' \in GM} \gamma(s,s') + \sum_{s' \in nonGM} \gamma(s,s')} \tag{6}$$

where *GM* refers to the set of GM maize plants in the landscape and *nonGM* to the set of all non-GM maize plants.

The global dispersal framework is a simplification of the individual framework. It assumes that the impurity rate of a plant located at point s depends only on the relative position between this point and the closest GM point

$$\mu_s = \gamma(s,s'), \tag{7}$$

where s' represents the coordinates of the closest GM plant. This framework has also been used by Damgaard and Kjellson (2005) to model oilseed rape gene flow. In our attempt to simplify model dispersal, we chose to adopt the latter one, not only for its simplicity but also for the speed of computation time which would be too large using the individual framework.

2.2.3 Two Dispersal Functions

Cross pollination rate decreases as a function of the distance between pollen source and receptor field. There has been considerable effort on defining the shape of the dispersal curve. Klein et al. (2006b), Clark (1998), Damgaard and Kjellson (2005) proposed various forms of dispersal function. In our attempt to define simple dispersal model, we chose to adopt an exponential function. Moreover, cross pollination decreases rapidly in the first meters and is then characterized by a fat tail. This is why we proposed to model the decrease of cross-pollination rate close to the pollen donor differently from the decrease of cross-pollination rate more distant from the source, as proposed by Damgaard and Kjellson (2005) for oilseed rape.

The kernel is a compound exponentially decreasing function similar to the one used in Damgaard and Kjellson (2005). For more conciseness and clarity, we use the notation $\gamma(d_s)$, where d_s is the distance between s and s'. We have

$$\gamma(d_s) = \begin{cases} Ke \times e^{-a_1 d_s} & d_s \leq D \\ Ke \times e^{-a_1 D - a_2(d_s - D)} & d_s \geq D \end{cases} \quad (8)$$

where $d_s = \sqrt{x^2 + y^2}$.

This kernel has the advantage to be very simple but, as formulated in Damgaard and Kjellson (2005), it doesn't take wind effect into account. However, wind effect has been identified in Klein et al. (2003) as a key factor of pollen dispersal. He observed that dispersal patterns were shaped primarily by the major wind direction. Higher cross-pollination rates were observed in the downwind direction. To overcome this limitation, we assumed that wind affects distance from sampling point to pollen source. We modeled wind effect through interaction with distance, considering only the prevailing wind direction. We incorporated a so-called *effective distance*, which results from an interaction between distance and wind effect.

$$\gamma(d_s, \omega) = \begin{cases} Ke \times e^{-a_1 d_s^*} & d_s^* \leq D \\ Ke \times e^{-a_1 D - a_2(d_s^* - D)} & d_s^* \geq D \end{cases} \quad (9)$$

where $d_s^* = d_s \times (1 - \theta_v \cos(\omega - \omega_0))$, ω_0 is the main wind direction and ω is the angle between the vector $(0, s)$ and the vector $(0, s')$.

In order for the model to be able to adapt to the level of available in situ information, we keep the first kernel as the *default kernel* whose only input variable is the distance. If the prevailing wind direction is available through measurement or time series, the model defined in equation (9) is used.

2.2.4 Prior Distributions

We chose fairly non-informative prior distributions. No information was available for the parameters, except for $a2$ which, in the model formulation, is lower than $a1$ to model the quick decrease of cross pollination in the first meters and slower decrease after. The parameter Ke, which reflects the cross pollination rate at distance 0, is another exception; given that it represents a rate, it should lie between 0 and 1. We chose to consider uniform distributions for all parameters of the dispersal kernel. The prior distributions are summarized in the table 2.

Parameter	Lower bound	Upper bound	Mean	SD	CV
Ke	0	1	0.5	0.29	0.58
a_1	0	2	1	0.58	0.58
a_2	0	a_1	$a_1/2$	$a_1/\sqrt{12}$	0.58
D	1	10	5.5	2.6	0.47
θ_v	0	1	0.5	0.29	0.58

Table 2: Prior distributions used for parameters of the dispersal kernel

2.2.5 Bayesian Inference

In the Bayesian approach, model parameters are treated as random variables. The fundamental equation is $P(\theta|Y) \propto P(\theta) \times P(Y|\theta)$. Here θ is the vector of the parameters in the gene flow model, Y is the vector that includes all the observed data. The above equation says that the posterior distribution $P(\theta|Y)$, which specifies our knowledge about θ after invoking the data, is proportional to the product of the prior distribution $P(\theta)$, which represents our knowledge of the parameters before using the data, and the likelihood function $P(Y|\theta)$, which specifies the probability of observing Y given the values θ. The Bayesian approach thus allows for probabilistic predictions which are easily derived from posterior distribution of the model parameters. This methodology is particularly relevant in our case given the variability of the observations and our interest to assess the uncertainty of the predicted cross pollination rates.

Estimation of the observation model and the dispersal kernel parameters was achieved using Markov Chain Monte-Carlo (MCMC) methodology. Bayesian inferences were performed using software JAGS (Plummer, 2012). Simulated data produced by JAGS were processed using the CODA statistical R package (Plummer et al., 2009). After an adaptation phase (called *burn-in*) of 2×10^4 iterations, the convergence of the MCMC algorithm was checked by visually analysing three independent MCMC chains using three different initial values for parameters. Gelman and Rubin convergence statistics (Gelman and Rubin, 1992) were also calculated and examined. This criterion indicates that we can assume convergence of the Markov Chain to the posterior distribution with the 150000 iterations following the burn-in period. We then thinned the chain by using only one value out of every 25 in the Markov chain. Thinnig reduces auto-

correlations in the chain and reduces also computing time when using the posterior distribution to make predictions. Overall, that lefts 6000 parameters vectors for the inference.

2.2.6 Evaluation of Calibrated Model

Statistical Criteria

Models were firstly compared using scoring rules, which asses the quality of a probabilistic forecast (Gneiting and Raftery, 2007). Since the forecast is probabilistic, it can be represented by its cumulative distribution function F for an obsevation y_s. The continuous ranked probability score is defined as

$$CRPS(F, y_s) = -\int_{-\infty}^{+\infty} (F(x) - \mathbb{1}_{x \geqslant y_s})^2 \mathrm{dx}$$

where $\mathbb{1}_{x \geqslant y_s}$ takes values 1 if $x \geqslant y_s$ and 0 otherwise. However solutions to this intergral can be hard to compute. Fortunately, the CRPS can be expressed in a readily computable expression as

$$CRPS(F, y_s) = \frac{1}{2}\mathbb{E}_F|Y - Y'| - \mathbb{E}_F|Y - y_s|$$

We also used more classical criteria on the mean response. Let $Y = \{y_1, y_2, \ldots, y_N\}$ be the vector of all the observed data and $\hat{Y} = \{\hat{y_1}, \hat{y_2}, \ldots, \hat{y_N}\}$ be the vector of predicted mean response. Correlation (r), root mean squared error ($RMSE$) and modelling efficiency (EF) were calculated as follows:

Correlation: $r = \frac{\sigma_{Y\hat{Y}}}{\sigma_Y \sigma_{\hat{Y}}}$

Root mean squared error: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

Modelling efficiency: $EF = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \mathbb{E}_Y)^2}$

All those criteria were firstly calculated with the training dataset: After parameter estimation, parameters posterior distributions were used to predict this dataset. This allows to asses the quality of adjustment. Then the same posterior distributions were used to predict the validation dataset (i.e. a dataset that was not used for parameter estimation) and the same criteria were calculated. This provided an estimation of the quality of prediction.

ROC Analysis

In a context of coexistence the most important feature of a gene flow model is not necessarily the capacity to predict the cross pollination rate with the lowest possible error but rahter to predict with accuracy whether or not a non-GM plot is above or below the legal theshold. The problem is therefore a classification problem. Receiver operating characteristic (ROC) curves are often used as a means of evaluating diagnostic tests for decision making. ROC analysis is a procedure, derived from statistical decision theory, that was developed in the context of electronic signal detection. It became widely used in agronomic applications to asses the accuracy of a diagnostic or a model.

The ROC curve represents a plot of sensitivity values as a function of (1-specificity) values. Sensitivity is the rate of true positives and 1-specificity is the rate of false positives. The area under the ROC curve is a popular index of the overall performance of a test. This synthetic index is equal to one if the classification of non-GM plots cross pollination rates is perfect (i.e. no differences between the real observed classification and the classification obtained with model predictions) and equal to 0.5 if the classification is not better than a random classification. Area under ROC curves is usually calculated from predictions of a deterministic model. When the model is stochastic, one often calculates the mean for all predictions and the area under ROC curve is then calculated from those means. In this paper we tried to take full advantage of the stochasticity of the model. Indeed model predictions are represented by distributions derived from posterior distribution of model parameters. So we can calculate the area under ROC curve (Aurocc) for all the elements of those predictive distribution and obtain a distribution of Aurocc. In this way we can asses the quality of a decision resulting from model prediction but also the uncertainty in this quantified quality, this allows to balance a very good quality with a lot of uncertainty and poorer quality but with more confidence. Here the cross pollination rate represents our gold standard i.e. the variable of reference used to asses the accuracy of the model ranking. The threshold was set to 0.9%, which the EU legal threshold, so that a good model (a model with a high Aurocc) is a model which can segregate plants that are above or below this threshold.

3 Results

3.1 Parameter Estimation

For all the parameters, the posterior distribution was much narrower than the prior. The data allowed us to considerably reduce our uncertainty about those parameters. Table 3 summarizes the marginal posterior distribution for each parameter through its mean, standard deviation (SD) and coefficient of variation (CV). Not shown, but also important, is the fact that the prior distributions are all independent, whereas in the posterior distribution the parameters are correlated, and this correlation structure was used to make prediction.

The other characteristic to point out and maybe the most important in a Bayesian analysis is that the distribution of the parameters in the *Distance+Wind* model have a coefficient of variation lower than or equal to the distribution of the same parameters in the *Distance* model. The parameters of the model that take wind direction effect into account are more certain than the others. This can easily be interpreted by the fact that we added data between the two models. Indeed the parameters of the *Distance* model ignore the wind direction, thus the dispersal model is isotropic and therefore the pollen cloud is distributed evenly around the GM plots. While we know and observe in the data that the pollen cloud is mainly oriented in the direction of the wind. To find the best tradeoff in the estimation process, the parameters of the *Distance* model must have larger variances than the same parameters of the *Distance+Wind* model.

Model	Parameter	Mean	SD	CV
Distance	Ke	0.235	0.054	0.23
Distance+Wind	Ke	0.231	0.038	0.16
Distance	a_1	0.440	0.072	0.16
Distance+Wind	a_1	0.458	0.047	0.10
Distance	a_2	0.041	0.004	0.09
Distance+Wind	a_2	0.061	0.006	0.09
Distance	D	7.023	0.714	0.10
Distance+Wind	D	7.042	0.592	0.08
Distance	β_1	0.071	0.003	0.04
Distance+Wind	β_1	0.103	0.004	0.03
Distance	β_2	20.519	0.658	0.03
Distance+Wind	β_2	18.731	0.071	0.003
Distance	σ^2	1.300	0.089	0.06
Distance+Wind	σ^2	1.036	0.071	0.06
Distance+Wind	θ_v	0.535	0.018	0.03

Table 3: Posterior distributions of the model parameters

3.2 Predictions

First of all, we have looked at the ability of the model to reproduce the overall variability observed in the data. Figure 2 shows boxplots of observed and predicted cross pollination rates for different classes of distances. Those predictions are derived from the *Distance+Wind* model. One can realizes that the retranscription in the predictions of the overall observed variability in the data is satisfactory. Not shown but interesting to note, is the fact that the *Distance* model allows a satisfactory retranscription of the overall variability as well. This indicates that this behavior is mainly due to the randomness of the expectation in the Poisson model and much less to the integration of additional variables like wind direction.

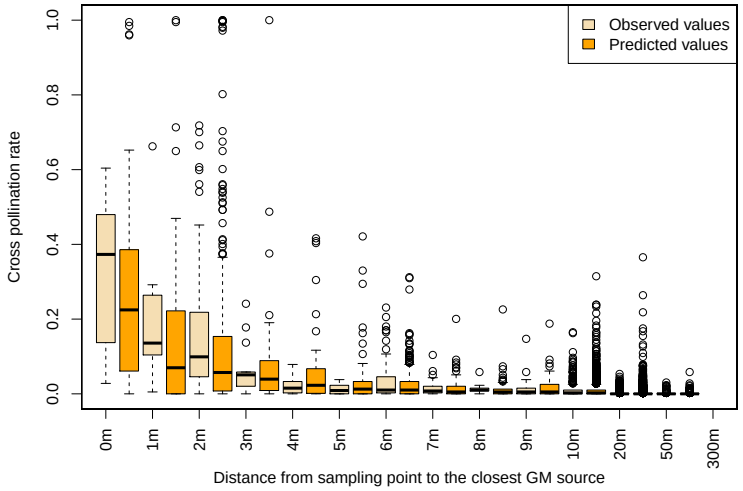


Figure 2: Boxplots of observed and predicted cross pollination rates as a function of distance to the closest GM source. Predictions are derived from the *Distance+Wind* model.

3.2.1 Goodness of Fit

The quality of the model was assessed by the goodness of fit, in other words, how well the model can reproduce the data. For all criteria defined above, the *Distance+Wind* model outperforms the simpler one. Indeed *CRPS* and *RMSE* are to be minimised whereas Correlation and Modelling efficiency are to be maximised. Table 4 summarizes the criteria values calculated with the training dataset for the *Distance* and the *Distance+Wind* models. Table 5 summarizes the same criteria calculated with the validation dataset

	Model	
Criterion	Distance	Distance + Wind
CRPS	-2.720	-2.400
r	0.696	0.746
RMSE	14.234	13.254
EF	0.465	0.536

Table 4: Criteria values for the two models calculated with the training dataset

We can observe here that the best model is patently the *Distance+Wind* model, for adjustment as well as for validation. The fact that the rank of models is the same with the training and validation datasets is not obvious. This is the case here and this is reassuring because this means that the model is not overfitted.

	Model	
Criterion	Distance	Distance + Wind
CRPS	-0.964	-0.944
r	0.702	0.720
RMSE	8.333	7.776
EF	0.393	0.471

Table 5: Criteria values for the two models calculated with the validation dataset

3.2.2 Quality of Decision

As stated before, in the coexistence and decision making context the most important feature of a gene flow model is not necessarily the capacity to predict the cross pollination rate with the lowest possible error but rather to predict with accuracy whether or not a non-GM plot is above or below the legal theshold. The criterion Aurocc (Area Under ROC Curve) was calculated for the two models in two different ways. The first way is the classic deterministic-like ROC analysis. As the model is not deterministic, the mean of each prediction was calculated and the the analysis was made on those means. The second way was to perform the classical ROC analysis but on each element of predictive distributions in order to obtain a distribution of the Aurocc criterion. This distribution allowed us to assess not only the quality of the decision resulting from model outputs but also the confidence we can have on this estimated quality. These results are summarized in Table 6. The *Distance+Wind* model was more accurate, but this appeared only when the stochasticity was taken into account.

	Model	
Criterion	Distance	Distance + Wind
Aurocc of the Mean	0.980	0.981
Mean of Aurocc	0.813	0.863
SD of Aurocc	0.029	0.026

Table 6: Aurocc values calculated for the two models

3.2.3 Benefits of Adding Data

We have looked at the capacity of our models to predict cross pollination rate with the lowest possible error (Goodness of fit) and their capacity to correctly rank plots or plants (Quality of decision). Another interesting and relatively basic feature of our models is the capacity to predict a cross pollination rate with the lowest possible variance. Indeed, as every prediction is characterized by a probability distribution, one wants to have a mean close to the real value but also to have a small dispersion around this mean. So, after assessing the goodness of fit and the quality of the decision resulting from model outputs, we tried to evaluate which model makes the less uncertain predictions. We therefore calculated the variance of each prediction for the two models and substracted the variances of the prediction from the *Distance* model to the variances of the prediction from the *Distance+Wind* model. We then looked at the sign of the calculated differences of variances.

In the training dataset, 70% (2064 out of 2937) variances are lower with the *Distance+Wind* model. The other 30% correspond to very high values of cross pollination rate, and thus located downwind. This can be interpreted by the fact that the *Distance* model is isotropic. It follows that the predictions of points located downwind are underestimated and their predicted variance is excessively optimistic. In contrast, with the *Distance+Wind* model, those points are better predicted in terms of means but as the value is larger, the variance is also larger. In the validation dataset, there are 99.95% (4428 out of 4430) variances are lower with the *Distance+Wind* model. The 0.05 other percent (i.e. 2 points) correspond to the farthest points from the GM plot.

4 Discussion

We have presented a Bayesian method to predict cross pollination rate using spatial and climatic data (distances between plots and main wind direction). The emphasis in this study was on the uncertainty in model predictions and, in particular, on the contribution of additional input data to the overall uncertainty. This study shows that the Bayesian method allows the integration of additional input data such as climatic variables and thus improves the accuracy of model predictions. Further improvements need to be studied to reach our objective which is to be able to take all available information into account. One of the ongoing step is to integrate flowering dynamics as a factor influencing the cross pollination rate. Indeed, flowering delay that could occur between GM and non-GM plants can significantly reduce cross pollination and is therefore important to consider. Another step, which is in progress, is to integrate real agricultural landscape description in the model. That is to say being able for the model to take into account the multiplicity of GM sources.

5 Aknowledgments

This study was partially funded by the European Union project PRICE (PRactical Implementation of Coexistence in Europe), contract number 289157.

References

Angevin, F., Klein, E. K., Choimet, C., Gauffreteau, A., Lavigne, C., Messean, A., and Meynard, J. M. (2008). Modelling impacts of cropping systems and climate on maize cross-pollination in agricultural landscapes: The MAPOD model. *European Journal of Agronomy*, 28(3):471–484.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–59.

Clark, J. (1998). Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. *The American Naturalist*, 152:204–224.

Colbach, N., Clermont-Dauphin, C., and Meynard, J. (2001). GeneSys: a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers II. genetic exchanges among volunteer and cropped populations in a small region. *Agriculture, Ecosystems and Environment*, 83:255–270.

Damgaard, C. and Kjellson, G. (2005). Gene flow of oilseed rape (*Brassica napus*) according to isolation distance and buffer zone. *Agriculture, Ecosystems and Environment*, 108:291–301.

Devos, Y., Reheul, D., and De Schrijver, A. (2005). The co-existence between transgenic and non-transgenic maize in the european union: a focus on pollen flow and cross-fertilization. *Environmental and Biosafety Research*.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511.

Gneiting, T. and Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

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

Klein, E., Lavigne, C., Picault, H., Renard, M., and Gouyon, P. (2006a). Pollen dispersal of oilseed rape: estimation of the dispersal function and effects of field dimension. *Journal of Applied Ecology*, 43:141–151.

Klein, K., Lavigne, C., and Gouyon, P. (2006b). Mixing of propagules from discrete sources at long distance: comparing a dispersal tail to an exponential. *BMC Ecology*, 6:3.

Larédo, C. and Grimaud, A. (2007). Stochastic models and statistical inference for plant pollen dispersal. *Journal de la Société Française de Statistique*, 148:77–105.

Lavigne, C., Klein, E., Vallee, P., Pierre, J., Godelle, B., and Renard, M. (1998). A pollen-dispersal experiment with transgenic oilseed rape. estimation of the average pollen dispersal of an individual plant within a fiel. *Theoretical and Applied Genetics*, 96:886–896.

Messéan, A., Squire, G., Perry, J. Angevin, F., Gomez, M. Townend, P., Sauuse, C., Breckling, B., Langrell, S., Dzeroski, S., and Sweet, J. (2009). Sustainable introduction of GM crops into european agriculture: a summary report of the FP6 SIGMEA research project. *OCL*, 16:37–51.

Plummer, M. (2012). *JAGS Version 3.3 user manual*.

Plummer, M., Best, N., and Cowles, K. abd Vines, K. (2009). *The coda Package: Output analysis and diagnostics for MCMC*.

Annexes 3

Sampling strategies to evaluate rate of transgenic adventitious presence in non-genetically modified crop fields. Bancal, R., Bensadoun, A., Messéan, A., Monod, H., Makowski, D. Soumis à *Risk Analysis*, 2014.

Sampling strategies for evaluating the rate of adventitious transgene presence in non-genetically modified crop fields

Abstract

According to European Union regulations, the maximum allowable rate of adventitious transgene presence in non-genetically modified (GM) crops is 0.9%. We compared four sampling methods for the detection of transgenic material in agricultural non-GM maize fields: random sampling, stratified sampling, random sampling + ratio reweighting, random sampling + regression reweighting. Random sampling involves simply sampling maize grains from different locations selected at random from the field concerned. The stratified and reweighting sampling methods make use of an auxiliary variable corresponding to the output of a gene-flow model (a zero-inflated Poisson model) simulating cross-pollination as a function of wind speed, wind direction, and distance to the closest GM maize field. With the stratified sampling method, an auxiliary variable is used to define several strata with contrasting transgene presence rates, and grains are then sampled at random from each stratum. With the two methods involving reweighting, grains are first sampled at random from various locations within the field, and the observations are then reweighted according to the auxiliary variable. Data collected from three maize fields were used to compare the four sampling methods, and the results were used to determine the extent to which transgene presence rate estimation was improved by the use of stratified and reweighting sampling methods. We found that transgene rate estimates were more accurate and that substantially smaller samples could be used with sampling strategies based on an auxiliary variable derived from a gene-flow model.

Key-words: Genetically modified crop, gene-flow model, sampling, stratification.

1. INTRODUCTION

The coexistence of genetically modified (GM) and conventional crops in the same area has become an important issue since the introduction of GM crops a few decades ago, due to gene flow between fields during pollination [1, 2, 3, 4]. It remains highly controversial, because of the risk of transgenes passing from outcrossing GM crops to non-GM crops, which may be costly to non-GM crop growers, as reported for the GM maize [5]. Pollen emitted by GM maize may be dispersed over nearby non-GM maize fields, resulting in the fertilization of non-GM maize plants [1, 3]. A fraction of the grains produced by the non-GM maize crops may thus contain the transgene from the GM maize crop [4]. If this fraction is large enough, then the commercial product derived from non-GM maize may lose its GM-free status and any related price premium.

In the European Union, regulations have been established to guarantee the traceability and labeling of genetically modified organisms (GMOs) and the products generated from them throughout the food chain (EC No 1829/2003). The main purpose of this regulation is to provide consumers with information, through the compulsory labeling of food products, so that they can exercise free choice in the selection of food products containing GMOs or GMO-free products. According to these regulations, the words 'genetically modified' or 'produced from genetically modified (name of the organism)' must be clearly indicated on the labels of food and feed products containing GMOs. Only traces (less than 0.9%) of GMOs are permitted, provided that their presence is adventitious and technically unavoidable.

Thus, according to the EU regulations, a product with an adventitious content of GM material of less than 0.9% could still be labeled as non-GM, and sold as such. On the contrary, if the tolerance threshold of 0.9% is exceeded, the product must be labeled as containing a GMO. This may render the product less acceptable to consumers and may be associated with economic losses. However, the implementation of EU regulations requires efficient detection techniques for the accurate identification of GM material in non-GM products. For maize, the adventitious presence of transgenic material in non-GM crop fields can be detected by applying a transgene detection method (e.g., PCR) to a sample of maize grains taken from different locations within the field tested, before harvest [7, 8]. The rate of

transgene presence in the tested field can then be estimated from the results, and the estimated rate value can be compared to the 0.9% threshold defined in the EU regulations. The accuracy of the classification obtained with this approach depends on both the accuracy of the detection method and the grain sampling strategy. Methods of transgene detection have been evaluated elsewhere [7]. Allnutt *et al.* [9] recommended the use of simple random sampling for estimation of the rate of adventitious transgene presence, but this method has never been compared with more sophisticated sampling methods based on the use of an auxiliary variable.

Gene flow in agricultural areas is influenced by wind direction, wind speed, distance between fields, and farmers' practices [1, 10, 11], and the rate of transgene presence can vary widely within a given field [12]. This high within-field variability makes it difficult to define an efficient sampling strategy. We hypothesized that it would be possible to make use of this information to design efficient sampling strategies in which a smaller sample size could be used without reducing the accuracy of the estimated rate of transgene presence. Statistical and semi-mechanistic gene flow models (e.g., MAPOD, [13]) have been developed to predict the rate of transgene presence in conventional non-GM fields. Some of these models predict the number of transgenic grains in each maize ear from a non-GM maize field as a function of a limited number of input variables. Their outputs could be used as auxiliary variables, for the design of new sampling strategies to decrease sample size. However, the degree to which this approach yields an improvement over random sampling depends on the predictive capacity of the model concerned. If the gene-flow model can predict the spatial distribution of transgene presence rates within a maize field, then its outputs could be used to define strata with contrasting presence rates before the sampling date, and grains could then be sampled within each stratum. This approach should improve transgene presence rate estimation if within-stratum variability is smaller than between-stratum variability. An alternative approach involves the use of model outputs after the sampling date, for the reweighting of transgene presence rates measured at different locations selected at random within the field concerned. An updated and more accurate transgene presence rate can then be derived for the whole field from the reweighted measurements. These two approaches (i.e.,

stratification and reweighting) are commonly used in social science and ecology surveys ([14, 15]), but have never before been used to estimate the rate of transgene presence in non-GM crop fields.

In this study, we analyzed the extent to which the estimation of adventitious transgene presence rates in agricultural fields could be improved by the use of sampling strategies based on an auxiliary variable derived from a gene-flow model. We compared four sampling methods for the detection of transgene presence: i) random sampling, ii) stratified sampling, iii) random sampling + ratio reweighting, iv) random sampling + regression reweighting. The last three of these methods use the output of a gene flow model as an auxiliary variable. We compared the four sampling methods, using real data collected from three maize fields, assuming that the presence of the transgene in grains was correctly detected.

2. SAMPLING METHODS

The sampling methods presented below were designed to estimate the mean transgene rate $\bar{Y}$ in a field U including N maize ears, through the use of a sample $t = \{t_1, \ldots, t_n\}$ including n ears selected in U. We use y_k to denote the transgene rate of the k^{th} ear, k=1, ..., N. We assume that the transgene rate is measured for each one of the n ears sampled. We also assume that the transgene rate can be predicted by a gene-flow model for each ear of the field concerned, and we use x to denote the model output and x_k to denote the value of the model output obtained for the k^{th} ear, k=1, ..., N.

2.1. Simple random sampling (SRS)

With this method, ears are selected at random, without replacement, from the entire field, and the transgene rate in the field U is estimated as follows:

$$\hat{Y}_{SRS} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_{t_i} \qquad (1)$$

The mean square error and the variance of $\hat{Y}_{SRS}$ ([14], chapter 2; [16], chapter 2) are both equal to:

$$var\left(\hat{Y}_{SRS}\right) = \left(1 - \frac{n}{N}\right)\frac{1}{n}S_y^2 \quad (2)$$

where S_y^2 is the variance of the transgene rate over the entire field, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$. In this study, we used the SRS method as a benchmark.

2.2. Ratio reweighting

This method is based on the predicted transgene rate *x*. The mean of x over the whole field *U* is denoted by $\bar{X}$, its variance is denoted by $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$, and its mean over the sample t is denoted by $\bar{x}$. The ratio reweighting method assumes that x is proportional to y. The coefficient of proportionality $R = \bar{Y}/\bar{X}$ estimated as $\hat{R} = \bar{y}/\bar{x}$, and the estimator of $\bar{Y}$, for a sample selected at random, are defined as follows ([14], chapter 6, [15], [16], chapter 7):

$$\hat{Y}_{Ratio} = \hat{R}\bar{X} = \bar{y}\frac{\bar{X}}{\bar{x}} \quad (3)$$

For the ratio estimator, the squared bias is smaller than the variance, and the following formula can be given for the approximate mean square error or variance of $\hat{Y}_{Ratio}$ ([16], chapter 7):

$$var\left(\hat{Y}_{Ratio}\right) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\left(S_y^2 + R^2S_x^2 - 2RS_{xy}\right) \quad (4)$$

where $S_{xy} = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{Y})(x_k - \bar{X})$. When $R^2S_x^2 - 2RS_{xy} < 0$, the upper limit of $var\left(\hat{Y}_{Ratio}\right)$ is $var\left(\hat{Y}_{SRS}\right)$. This condition is satisfied when $\rho > \frac{1}{2}\frac{CV(x)}{CV(y)}$ (Cochran, 1977), where $\rho = S_{xy}\Big/\sqrt{S_y^2S_x^2}$ is the correlation between *x* and *y*, $CV(x) = \frac{\sqrt{S_x^2}}{\bar{X}}$, and $CV(y) = \frac{\sqrt{S_y^2}}{\bar{Y}}$, i.e., when the correlation coefficient ρ is sufficiently high with respect to the ratio of the coefficients of variation. The estimator $\hat{Y}_{Ratio}$ converges on $\bar{Y}$ when *n* and *N* - *n* tend to infinity.

2.3. Regression reweighting

This approach can be seen as a generalization of the ratio reweighting method described above. It assumes that x and y are linearly related, but that y is not necessarily equal to zero when x=0. The regression-based estimator $\hat{Y}_{Reg}$ of $\bar{Y}$ is defined as follows ([14] chapter 7; [16] chapter 8):

$$\hat{Y}_{Reg} = \bar{y} + \hat{B}(\bar{X} - \bar{x}) \quad (5)$$

where $\hat{B} = \frac{s_{xy}}{s_x^2}$, $s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{Y})(x_k - \bar{X})$ and $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$. A large-sample approximation of the mean square error and of the variance of $\hat{Y}_{Reg}$ is given by ([14] chapter 7):

$$var(\hat{Y}_{Reg}) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\left(S_y^2 - \frac{{s_{xy}}^2}{s_x^2}\right) = \left(1 - \frac{n}{N}\right)\frac{1}{n}S_y^2(1 - \rho^2) \quad (6)$$

where ρ is the coefficient of the correlation between x and y. The variance (6) tends toward $var(\hat{Y}_{SRS})$ when ρ tends towards 0, and it takes a value close to zero when ρ tends towards 1.

2.4. Stratified sampling based on an auxiliary variable

In this approach, the field is split into strata characterized by contrasting values of x, and samples are taken at random within each stratum. Note that H is the number of strata, N_h the number of ears of the stratum h, n_h the number of ears selected from stratum h ($\sum_{h=1}^{H} n_h = n$ and $\sum_{h=1}^{H} N_h = N$), and $\overline{y_h}$ the mean value of the n_h measurements collected from stratum h. The Horvitz-Thomson estimator of $\bar{Y}$ for stratified sampling ([14] chapter 5; [16] chapter 11) is defined as follows:

$$\hat{Y}_{HT} = \sum_{h=1}^{H}\frac{N_h}{N}\overline{y_h} \quad (7)$$

This estimator is unbiased and its variance can be defined as follows:

$$var(\hat{Y}_{HT}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1 - \frac{n_h}{N_h}\right)\frac{S_{h,y}^2}{n_h} \quad (8)$$

where $S_{h,y}^2$ is the within-stratum variance of y in stratum h, defined by $\frac{1}{N_h - 1}\sum_{i \in h}(y_i - \overline{Y_h})^2$, where $\overline{Y_h}$ is the mean of y in stratum h. In this study, the H strata are defined on the basis of the auxiliary variable x. The h^{th} stratum, h=1, ..., H, includes the maize ears with a predicted transgene rate x falling

within the interval $\left[q_{h-1/H}(x); q_{h/H}(x)\right]$, where $q_k(x)$ is the quantile k of the distribution of the values of x in the field U. The number of ears sampled in the h^{th} stratum is defined according to the Neyman optimal allocation rule, $n_h = n \times \frac{s_{h,x}^2}{\sum_h s_{h,x}^2}, \forall h \in [\![1, H]\!]$. With this allocation rule, the strata with highest levels of dispersion are the most intensively sampled. This rule was defined to minimize $var(\hat{Y}_{HT})$ ([14] chapter 5). However, as n_h is calculated with an auxiliary variable, the Neyman rule is not necessarily optimal here.

3. COMPARISON OF SAMPLING METHODS

3.1. Experiments

Sampling methods were tested with data from three field experiments (Fig. 1): two experiments carried out in 1998 and 1999 at Montargis, France [17], and an experiment carried out in 2004 at Mas Cebria in Spain [18]. The three experimental fields studied had different rates of transgene presence (Table I). In each experiment, two maize cultivars, with similar flowering dynamics but different grain colors were grown in a receptor field and an emitter field. The experimental designs are described in Figure 1.

In the Montargis 1998 experiment, a 120 m x 120 m plot was sown with maize such that a cultivar with blue grains (a cultivar closely related to Adonis, but homozygous for the "blue" allele) was located in a central area (20 m x 20 m), the rest of the plot being sown with a variety with yellow grains (the hybrid variety Adonis). The blue-grained maize crop was used as the emitter and the yellow-grain maize crop was used as the receptor. The blue maize was used as a substitute for GM maize. The ear silks in the receptor field fertilized by blue maize pollen displayed a blue coloration. Ears were sampled from a regular grid of 1.6 m x 2 m (10% sampling ratio) up to 20 m from the blue area and 2.4 m x 4 m (3.3% sampling ratio) further away (one ear was sampled per grid cell). The final sample size was 2937 ears and the rate of transgene (or blue gene, in this case) presence was calculated for each ear as the percentage of grains with a blue color. The mean rate of blue-gene

presence in the yellow-maize field was 1.12%. The total number of grains on each ear, which is required for calculation of the rate of transgene presence, rather than the number of transgenic grains, was not measured ear-by-ear but was instead estimated at 394 (mean value for 34 randomly selected ears) [17].

In the Montargis 1999 experiment, a 200 m x 100 m plot was sown with blue maize grains, and this plot was surrounded by a field sown with yellow maize grains. The blue-grain plot was not in the central part of the field, but was upwind of the yellow-grain field with respect to the prevailing wind direction. The sampling grid was 0.8 m x 4 m (10% sampling ratio) in size, up to 20 m from the blue spot and 1.6 m x 4 m (5% sampling ratio) further away. The final sample size was 4430 ears and the rate of transgene (represented here by the blue gene) presence was calculated for each ear as the percentage of grains with a blue color, with a mean of 0.36% blue grains in the yellow field. As in the preceding experiment, we used an estimation of 394 grains for the mean total number of grains per ear.

In Mas Cebria in 2004, the experimental design consisted of a central plot sown with four different hybrids of the GM cultivar Mon810, all with yellow grains. This central plot was surrounded by a conventional cultivar with white grains (Pioneer Hi Bred). The transgenic yellow-grain (this color being dominant) maize was used as the emitter and the white-grain maize was used as the receptor. The grains in the receptor field resulting from cross-pollination by the emitter were yellow. We used a sampling procedure described elsewhere [17] to select 708 points, at each of which, we sampled three ears. The rate of transgene presence was calculated for each ear as the percentage of grains with a yellow color. The mean transgene ratio was 1.90% yellow grains in the white-maize field. In this experiment, both the number of yellow kernels and the total number of grains were determined, by counting, for each ear.

3.2. Gene dispersion model

A zero-inflated Poisson statistical regression model was used to predict the rate of transgene presence for each ear in the three experiments presented above [19]. The model output variable was used as an auxiliary variable for the implementation of three of our sampling strategies. The model includes three

inputs: wind direction and speed during the flowering period, and the spatial configuration of the field. The model computes an efficient distance r*:

$$r^* = r \times (1 - \theta cos(\omega - \omega_0))$$

where ω is the angle between the receptor point and the emitter point, θ is the wind effect parameter, ω_0 is the wind direction, and r is the distance between the emitter and the receptor. Gene flow was simulated with the dispersion function $\gamma(r^*)$ as defined in a previous study [20]:

$$\gamma(r^*) = \begin{cases} Ce^{-ar^*}, r^* \leq D \\ Ce^{-(a-b)D-br^*}, r^* \geq D \end{cases}$$

where a, b, C, and D are the dispersion function parameters. Based on this dispersion function, the expected number of transgenic grains in the receptor was defined as:

$$\mu = K \times \gamma(r^*)$$

where K is the total number of grains on the receptor ear. This model considers only one GM emitter, that closest to the receptor, and the number of transgenic grains on the receptor ear was calculated as follows:

$$y \sim ZIP(p, \mu)$$

where ZIP is a zero-inflated Poisson distribution [21, 22], such that y takes the value 0 with a probability p, and takes a value drawn from a Poisson distribution of parameter μ with a probability $1 - p$, with logit(1-p)= $\beta_1(\beta_2$-r*).

The parameters (a, b, C, D, θ, β_1, β_2) of the gene-flow model were estimated with training datasets independent of the predicted datasets. The data obtained at Montargis in 1998 were predicted with parameters estimated in the Montargis 1999 experiment. The data obtained in the Montargis 1999 and Mas Cebria 2004 experiments were predicted with the parameters estimated from the Montargis 1998 experiment. Parameters were estimated with a Markov chain Monte Carlo (MCMC) algorithm implemented in JAGS software [23]. The prior parameter probability distributions were defined as

independent uniform distributions: a~Unif(0, 2), b~Unif(0, a), C~Unif(0, 1), D~Unif(1, 10), θ~Unif(0, 1), β_1~Unif(0, 10), β_2~ Unif(-150, 150). The transgene rate of each ear in each experiment was predicted as the mean value obtained for 2000 MCMC runs. The coefficient of correlation between the real data and the dispersion model predictions was 0.751 for Montargis 1998, 0.671 for Montargis 1999 and 0.701 for Mas Cebria 2004 (Table I). The measured and predicted transgene presence rates are shown in Fig. 2.

3.3. Accuracy of the transgene rate estimates obtained with the four sampling methods

The theoretical standard deviations of the estimated mean transgene rates $\hat{Y}_{SRS}$, $\hat{Y}_{Ratio}$, and $\hat{Y}_{Reg}$ were calculated with Eqs.(2, 4, 6) for sample sizes ranging from n=5 to 200 ears. In these equations, N was set at the total number of ears in each experiment, and S_y^2, S_x^2, S_{xy}^2, ρ and R were calculated for each experiment from the N available values of x and y.

We also assessed the four sampling methods described above by simulations, for the same range of sample sizes. Ten thousand samples were randomly drawn for each of the three experiments and for each sample size. Each of the samples generated was used to estimate the mean transgene rate with Eqs.(1, 3, 5). This approach yielded three series of 10,000 estimated mean transgene rate values for each experiment and each sample size. Ratio and regression reweighting was implemented with the output of the transgene model described in section 3.2 used as an auxiliary variable.

The mean transgene rate of each experiment was also estimated by stratified sampling, with Eq.(7). We considered two very different numbers of strata successively: H=5 and H=50. The strata were defined from quantiles calculated from transgene dispersion model predictions. Ten thousand samples were generated for each of the two numbers of strata, and for sample sizes ranging from 5 to 200 ears (for H=5) and from 50 to 200 ears (for H=50).

Each series of 10,000 estimated transgene rate values was compared to a reference value corresponding to the mean transgene value calculated from all the measurements collected in the experimental field (Table I). This comparison was made for each method, each experiment, and each sample size, by calculating two criteria:

- The root mean square error (RMSE, square root of the mean of the squared differences between the estimated mean transgene rate values and the reference mean value for the experiment considered),
- The misclassification frequency, assessed by calculating the proportion of false positives for Montargis 1999, and the proportions of false negatives for Montargis 1998 and Mas Cebria 2004. The proportion of false positives is equal to the proportion of mean transgene rate estimates above the threshold of 0.9% in the experimental fields in which the actual mean transgene rate is below 0.9%. The converse is also true.

A low RMSE value indicates that the estimate of mean transgene rate is accurate. We calculated the proportion of false positives for Montargis 1999, because, in this field, the reference mean transgene rate was below 0.9%, and misclassification thus occurred when the estimated rate was above this threshold. On the contrary, false negative proportions were calculated for the other two fields, because their reference mean transgene rates were higher than 0.9%.

4. RESULTS

The theoretical standard deviations of the estimated mean transgene rates $\hat{Y}_{SRS}$, $\hat{Y}_{Ratio}$, and $\hat{Y}_{Reg}$ decrease with increasing sample size (Fig. 3). In the Montargis 98 experiment, the standard deviations of $\hat{Y}_{Ratio}$ and $\hat{Y}_{Reg}$ were undistinguishable and smaller than the standard deviations of $\hat{Y}_{SRS}$ for all tested sample sizes (Fig. 3A). In the Montargis 99 experiment, the standard deviations of $\hat{Y}_{Ratio}$ and $\hat{Y}_{Reg}$ were also smaller than those of $\hat{Y}_{SRS}$, but the standard deviations obtained for ratio reweighting were slightly higher than those obtained for regression reweighting (Fig. 3B). The results obtained for Mas Cebria 2004 were different (Fig. 3C): the standard deviations obtained for ratio reweighting were slightly higher than those obtained for simple random sampling, and those for regression reweighting were smaller. The poor performance of ratio reweighting in the Mas Cebria 2004 resulted from the condition $\rho > \frac{1}{2}\frac{CV(x)}{CV(y)}$ not being satisfied in this experiment; the coefficient of variation of the model predictions was greater than the coefficient of variation for the observations in this experiment (Table

I) and the correlation between predictions and observations (coefficient =0.701) was below $\frac{1}{2}\frac{CV(x)}{CV(y)}$ (0.73). For this reason, the standard deviations of $\hat{Y}_{Ratio}$ were higher than the standard deviations of $\hat{Y}_{SRS}$ in Mas Cebria 2004.

These results were confirmed by the RMSE values calculated for the 10,000 samples drawn from the data. In Montargis 98 and Montargis 99 (Fig. 4AB), the RMSE values of $\hat{Y}_{Ratio}$ and $\hat{Y}_{Reg}$ were very similar and were smaller than those obtained for $\hat{Y}_{SRS}$. A different situation was observed for the Mas Cebria 2004 experiment, in which the RMSE of $\hat{Y}_{Ratio}$ was higher than the RMSE of $\hat{Y}_{SRS}$ for all sample sizes (Fig. 4C) and higher than the RMSE of $\hat{Y}_{Reg}$. The RMSE values of $\hat{Y}_{Reg}$ were higher than those obtained by simple random sampling for samples of less than 75 ears, but were lower if the sample size exceeded this threshold (Fig. 4C). In Mas Cebria 2004, the mean value of the model output for the sample ($\bar{x}$) and its variance were, in most cases, zero for sample sizes of less than 40. As a result, ratio and regression reweighting approaches performed very poorly in these situations.

The RMSE values obtained by stratified sampling were lower than those obtained by simple random sampling for all experiments, all sample sizes, and the two numbers of strata (5 and 50) tested (Fig. 4). Stratified sampling thus yielded more accurate transgene rate estimates than simple random sampling, even if the number of strata was no higher than 5. Stratified sampling almost always outperformed ratio and regression reweighting. The only two exceptions concerned sample sizes of less than 20 ears for the Montargis 1999 experiment (for which stratified sampling with 5 strata yielded a higher RMSE than ratio and regression reweighting) and the Mas Cebria 2004 experiment (in which regression reweighting and stratified sampling with 5 strata gave similar RMSE values for sample sizes of more than 75 ears) (Fig.4C). The number of strata affected the accuracy of the transgene rate estimates; the most accurate estimates were obtained with 50 strata, once the sample size exceeded 60 (Fig. 4).

In the Montargis 98 experiment, false negative rates ranged from 50 to 75% for a sample size of 5 ears, depending on the method used (Fig. 5A). With simple random sampling, the false negative rate decreased to 35% when the sample size was increased to 200 ears. This value is much higher than the false negative rates obtained for ratio and regression reweighting (about 20% for a sample size of 200

ears) and the false negative rates obtained for stratified sampling (from about zero to 6%, depending on the number of strata). With 50 strata, false negative rates for stratified sampling fell below 5% as soon as the sample size exceeded 100 ears (Fig. 5A, Table II). For all the other methods, more than 240 ears were required to reach a misclassification rate of 5% (Table II). Overall, the false negative rates obtained by ratio and regression reweighting were lower than those obtained by simple random sampling. Stratified sampling gave even smaller misclassification rates, particularly for 50 strata.

Misclassification rates (false positives) were lower for Montargis 99 than for Montargis 98 (Fig.5B). The highest false positive rate (12%) obtained was that for simple random sampling, for a sample size of 50. All methods gave misclassification rates below 5% for sample sizes of more than 100 ears (Table II). Regression and ratio reweighting performed better than simple random sampling for almost all sample sizes. With both methods, a misclassification rate of 5% was reached for samples of fewer than 10 ears (Table II). Stratified sampling gave lower false positive rates than the other methods, for sample sizes of more than 50 ears.

The same ranking was obtained in the Mas Cebria 2004 experiment: the regression and ratio reweighting methods outperformed simple random sampling, and stratified sampling yielded the lowest misclassification rates (Fig. 5C). Stratified sampling with five strata gave a false negative rate below 5% for samples of more than 10 ears (Table II).

5. DISCUSSION

The adventitious presence of GM material in non-GM fields mostly results from cross-pollination, seed impurity and volunteers [24]. Maize pollen can be transported over long distances by the wind, making some degree of cross-pollination almost inevitable [24, 25]. For the purposes of consumer information, it is important to establish clear grain sampling strategies, to determine whether the rate of GM presence exceeds the regulatory tolerance threshold (0.9% in the EU). Sampling strategies are also required for evaluating the efficacy of proposed coexistence measures to prevent the

contamination of non-GM crop fields with GM material. In several parts of the world (including the EU), the coexistence of GM and non-GM is indeed a matter of concern, and several measures have been recommended to reduce the presence of GM material in non-GM fields (e.g., flowering delays between GM and non-GM crops, minimum distances between GM and non-GM fields) [24, 25]. Transgene detection in field experiments can be used to assess the efficacy of the proposed coexistence measures [24,25,26], but the reliability of the conclusions of such assessments depends on several factors, including grain sampling strategy, in particular [9].

Allnutt *et al*. [9] recommended the use of a simple random sampling scheme to estimate the frequency at which GM material was present in non-GM fields. However, the authors acknowledged that large sample sizes might be required to achieve a 95% probability of correct non-GM field classification with this sampling strategy, particularly in situations in which the true transgene presence rate is close to the tolerance threshold [9]. Our results show that estimation of the rate of adventitious transgene presence is improved by the use of output from a gene-flow model as an auxiliary variable. For a given sample size, sampling methods making use of gene-flow model output outperformed simple random sampling in most of the situations considered. Regression reweighting, ratio reweighting, and stratified sampling systematically yielded lower rates of misclassification than simple random sampling, in all three experiments considered. The sample size required to achieve a 95% probability of correct classification was reduced by a factor from 5 to 20, by the use of sampling methods based on an auxiliary variable rather than random sampling. These methods also yielded a lower RMSE, except at Mas Cebria in 2004 for ratio reweighting. In this experiment, the RMSE values obtained for ratio reweighting were higher than those for simple random sampling. The correlation between model predictions and observations was strong in this experiment (coefficient of 0.7), but the model predictions were much more variable than the observations, and ratio reweighting was not efficient in this case. The theoretical standard deviations calculated for simple random sampling, for regression reweighting, and for ratio reweighting were not always identical to the RMSE values calculated by simulation, but their comparison led to similar conclusions.

The highest rate of misclassification was that for the Montargis 98 experiment, due to a rate of transgene presence (1.12%) close to the threshold of 0.9%. For this experiment, the stratified method based on 50 strata yielded a false negative rate below 5% only for samples sizes of more than 100 ears. With the other methods, sample sizes needed to exceed 240 ears to obtain a misclassification rate below 5%. In the Montargis 99 and Mas Cebria 2004 experiments, the stratified method based on five strata gave a false positive rate below 5% for relatively small sample sizes: 10 and 20 ears for Mas Cebria 2004 and Montargis 99, respectively. In these two experiments, the actual transgene rates were very different from the 0.9% threshold (either much lower or much higher), and it was thus possible to obtain accurate classifications with small sample sizes.

In practice, methods using gene-flow model output as an auxiliary variable can be used in different ways. Ratio and regression reweighting methods can be used for the reweighting of ear samples. In our application, both methods gave similar results in terms of misclassification rates, but regression reweighting tended to lead to lower RMSE values and theoretical variances in Mas Cebria 2004. Both methods gave lower misclassification rates than simple random sampling without reweighting, even with small sample sizes (less than 10 ears). Stratified sampling is useful for the identification of several strata with contrasting predicted transgene presence rates. Each stratum can then be sampled separately, to obtain an estimate of the overall rate of transgene presence for the whole field. It is also possible to use the strata for the reweighting of samples generated by random sampling. We found that stratified sampling based on an auxiliary variable made it possible either to reduce the sample size to reach a given misclassification rate (e.g., 5% chance of misclassification) or to increase the accuracy of the transgene presence rate estimates for a given sample size.

The performance of the methods tested in this paper depends on several factors. One key factor is the correlation between gene-flow model outputs and observations. In our application, the coefficient for this correlation ranged from 0.67 to 0.75, depending on the experiment, and was sufficiently high to conclude that sampling methods using the model output as an auxiliary variable were advantageous. A second key factor is sample size: larger sample sizes are associated with lower RMSE and misclassification rate. A third factor is the actual rate of transgene presence. Misclassification rates

tend to be higher when the actual rate of transgene presence is close to the reference threshold (here, 0.9%). The accuracy of the detection method is also important. Here, the rate of transgene presence was assessed by counting the number of colored grains, and this method is highly accurate. However, the use of other methods, such as polymerase chain reaction (PCR) techniques [7] would be expected to generate less accurate measurements. The RMSE and misclassification rates calculated here should therefore be seen as minimal values.

Finally, the performance of stratified sampling is influenced by another factor: the number of strata. In our application, better results were generally obtained with 50 strata than with five strata, but the use of a large number of strata was beneficial only if several ears were sampled from each stratum. The use of a large number of strata is thus only recommended if transgene presence rates are estimated with large samples.

Due to the numerous factors influencing the performance of sampling strategies, it is difficult to give precise recommendations concerning sample size. However, our results indicate that, provided that the actual rate of transgene presence differs markedly from the tolerance threshold (by at least 50%), it is possible to achieve a 95% probability of correct classification with samples of about 10 to 20 ears per field. In situations in which the actual rate is closer to the tolerance threshold, sample sizes of more than 100 ears may be required to reach the same level of confidence.

REFERENCES

1. Demont M, Devos Y. Regulating coexistence of GM and non-GM crops without jeopardizing economic incentives. Trends in Biotechnology, 2008; 26, 353-358.
2. Messéan A, Angevin F, Champolivier J, Colbach N. Impact of GMOs within cropping systems: towards a more systemic approach. Aspects of Applied Biology, 2005; 191-196.
3. Messeguer J, Penas G, Ballester J, Bas M, Serra J, Salvia J, Palaudelmas M, Mele E. Pollen-mediated gene flow in maize in real situations of coexistence. Plant Biotechnology Journal, 2006; 4, 633-645.

4. Njontie C, Foueillassar X, Christov NK, Hüsken A. The impact of GM seed admixture on the non-GM harvest product in maize (*Zea mays* L.) Euphytica, 2011; 180: 163-172.
5. Wu F. Explaining public resistance to genetically modified corn: an analysis of the distribution of benefits and risks. Risk Analysis, 2004; 24: 715-726.
6. CE regulation (EC) no. 1830/2003 of the European Parliament and of the Council of 22 September 03 concerning the traceability and labelling of genetically modified organisms and the traceability of food and feed products produced from genetically modified organisms and amending Directive 2001/18/EC. Official Journal of the European Union, 2003, 46 L268: 24-28.
7. Pla M, Paz JL, Penas G, Garcia N, Palaudelmas M, Esteve T, Messeguer J, Mele E. Assessment of real-time PCR based methods for quantification of pollen-mediated gene flow from GM to conventional maize in a field study. Transgenic Research, 2006; 15: 219-228.
8. MacArthur R, Feinberg M, Bertheau Y. Construction of measurement uncertainty profiles for quantitative analysis of genetically modified organisms based on interlaboratory validation data. J AOAC Int., 2010; 93:1046-56.
9. Allnutt TR, Dwyer M, McMillan J, Henry C, Langrell S. Sampling and modeling for the quantification of adventitious genetically modified presence in maize. Journal of Agricultural and Food Chemistry, 2008; 56: 3232-3237.
10. Jarosz N, Loubet B, Durand B, McCartney A, Foueillassar X, Huber L. Field measurements of airborne concentration and deposition rate of maize pollen. Agricultural and Forest Meteorology, 2003; 119: 37 -51
11. Jarosz N, Loubet B, Durand B, Foueillassar X, Huber L. (2005). Variations in maize pollen emission and deposition in relation to microclimate. Environmental Science & Technology, 2005; 39: 4377-4384.

12. Bannert M, Vogler A, Stamp P. Short-distance cross-pollination of maize in a small-field landscape as monitored by grain color markers. European Journal of Agronomy, 2008: 29: 29-32.

13. Angevin F, Klein EK, Choimet C, Gauffreteau A, Lavigne C, Messéan A, Meynard J-M. Modelling impacts of cropping systems and climate on maize cross-pollination in agricultural landscapes: The MAPOD model. European Journal of Agronomy, 2008; 28: 471-484.

14. Cochran WG. Sampling Techniques. New York: John Wiley & Sons, third edition, 1977.

15. Gregoire TG, Salas C. 2009. Ratio estimation with measurement error with auxiliary variate. Biometrics, 2009; 65: 590-598.

16. Thompson SK. Sampling. Hoboken, New Jersey: John Wiley & Sons, Inc, 2012, third edition.

17. Klein EK, Lavigne C, Foueillassar X, Gouyon PH, Laredo C. Corn pollen dispersal: Quasi mechanistic models and field experiments. Ecological Monographs, 2003; 73: 131-150.

18. Palaudelmas M, Mele E, Monfort A, Serra J, Salvia J, Messeguer J. Assessment of the influence of field size on maize gene flow using SSR analysis. Transgenic Research, 2012; 21:471–483.

19. Bensadoun A, Monod H, Angevin F, Makowski D, Messéan A. Modeling of gene flow by a Bayesian approach: a new perspective for decision support. AgBioForum, 2014, 17, in press.

20. Damgaard C, Kjellson G. Gene flow of oilseed rape (*Brassica napus*) according to isolation distance and buffer zone. Agriculture, Ecosystems and Environment, 2005; 108 :291-301.

21. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 1992; 34: 1-15.

22. Kuhnert PM, Martin TG, Mengersen K, Possinghma HP. Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. Environmetrics, 2005; 16: 717-747.

23. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria, 2003.

24. Palaudelmas M, Melé E, Penas G, Pla M, Nadal A, Serra J, Salvia J, Messeguer J. Sowing and flowering delays can be an efficient strategy to improve coexistence of genetically modified and conventional maize. Crop Science, 2008; 48: 2404-2413.

25. Devos Y, Reheul D, De Schijver A. The co-existence between transgenic and non-transgenic maize in the European Union: a focus on pollen flow and cross-fertilization. Biosafety Res., 2005; 4: 71-87.

26. Rühl G, Langhof M. Coexistence in maize: effect of the genetically modified maize final depth on pollen-mediated gene flow. Crop Science, 2011; 2186-2193.

Table I. Measured and predicted rates of transgene presence in the three experimental maize fields. Transgene presence rates correspond to the percentages of blue grains in Montargis 1998 and 1999, and to the percentage of yellow grains in Mas-Cebria 2004. The rate of transgene presence was measured for each maize ear sampled in the fields. The mean value, standard deviation, and coefficient of variation were calculated for each experimental field. The rate of transgene presence was also predicted, ear-by-ear, with a gene-flow model. The mean value, standard deviation, and coefficient of variation of the model predictions were calculated for each field, and their correlations with the measured presence rates were estimated.

	Field 1 (Montargis 1998)	Field 2 (Montargis 1999)	Field 3 (Mas-Cebria 2004)
Mean value for measured rate of transgene presence (%)	1.12	0.36	1.90
Mean value for predicted rate of transgene presence (%)	1.02	0.79	1.53
Correlation between measured and predicted rates	0.75	0.67	0.70
Coefficient of variation of the measurements (%)	442	751	262
Coefficient of variation of the predictions (%)	313	299	385
Standard deviation of the measurement (%)	4.94	2.71	4.99
Standard deviation of the prediction (%)	3.18	2.37	5.91

Table II. Minimum sample sizes (numbers of ears) required for a misclassification rate of 5% (95% chance of correct classification) for the three experimental fields. Misclassification rates were assessed for a detection threshold of 0.9% of GM grains. Sample sizes larger than the reported values gave misclassification rates below 5%.

Method	Field 1 (Montargis 98)	Field 2 (Montargis 99)	Field 3 (Mas Cebria 2004)
Simple random sampling	900	100	50
Ratio reweighting	450	8	20
Regression reweighting	500	5	25
Stratification (5 strata)	240	25	10
Stratification (50 strata)	100	51	50

Figure 1. Characteristics of the three maize fields used for the comparison of sampling methods.

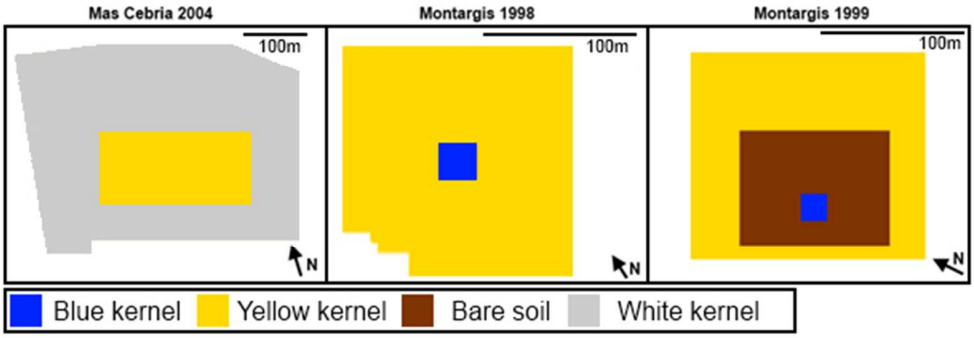
Figure 2. Observed and predicted rates of transgene presence (% of maize grains) in maize ears from three maize fields: Montargis 1998 (a), Montargis 1999 (b), and Mas Cebria 2004 (c). Each point corresponds to one maize ear. All observations are presented.

Figure 3. Theoretical standard deviation of the rate of transgene presence, as estimated by simple random sampling, ratio and regression reweighting for three maize fields, as a function of sample size (5 to 200 ears); Montargis 1998 (a), Montargis 1999 (b), and Mas Cebria 2004 (c).

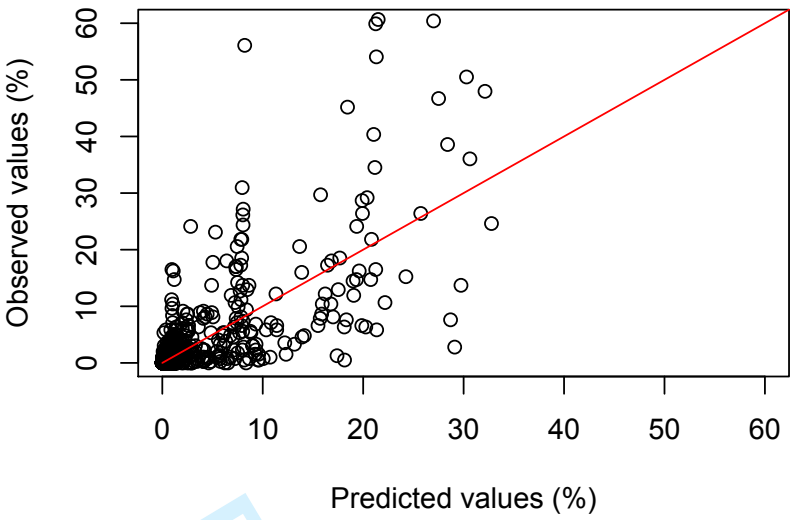
Figure 4. Root mean square error (RMSE, %) of estimated rates of transgene presence, as obtained by simple random sampling, ratio and regression reweighting, stratified sampling (with 5 or 50 strata) for three maize fields: Montargis 1998 (a), Montargis 1999 (b), and Mas Cebria 2004 (c). Calculations were performed for sample sizes of 5 to 200 ears.

Figure 5. Misclassification rates (false-negative or false-positive, depending on the field considered) calculated for a detection threshold of 0.9% GM grains. Calculations were performed for sample sizes of 5 to 200 ears, for simple random sampling, ratio and regression reweighting, and stratified sampling (with 5 or 50 strata) for three maize fields: Montargis 1998 (a), Montargis 1999 (b), and Mas Cebria 2004 (c). Horizontal

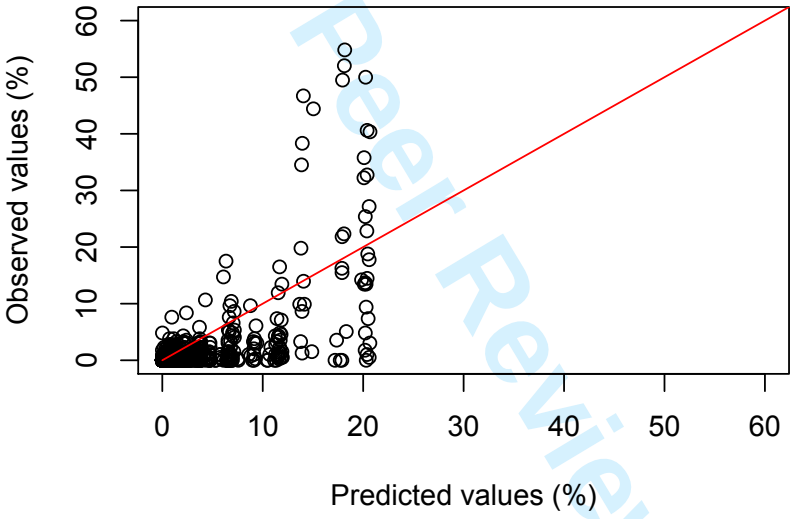
dashed lines indicate the rate of 5% misclassification (95% chance of correct classification).



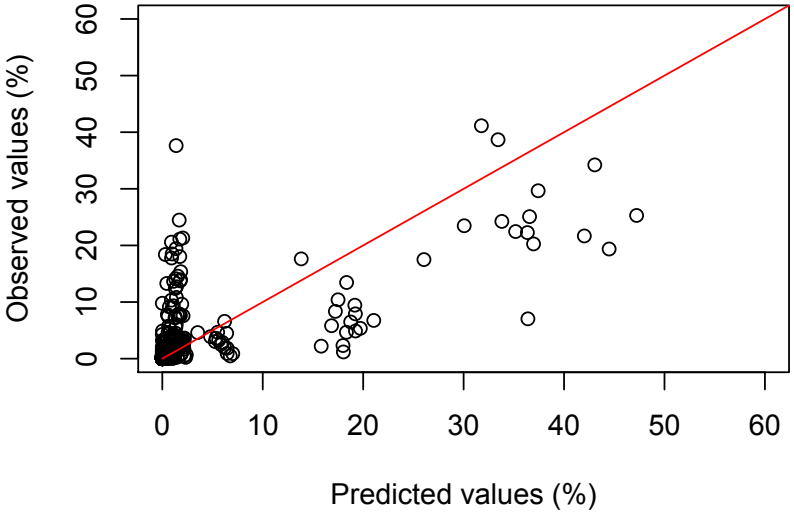
(a)



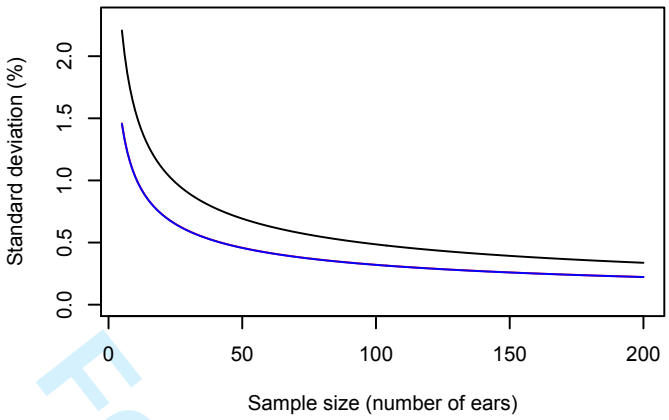
(b)



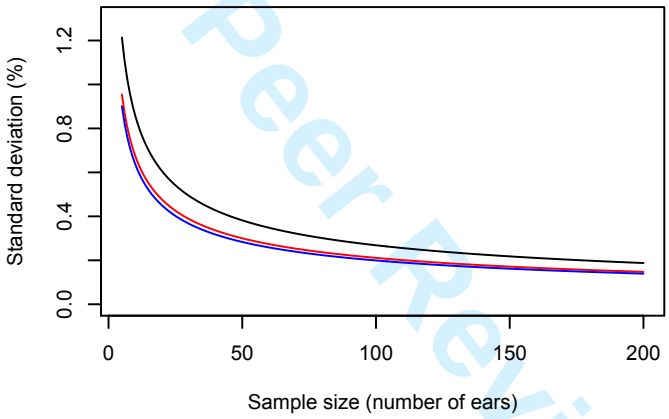
(c)



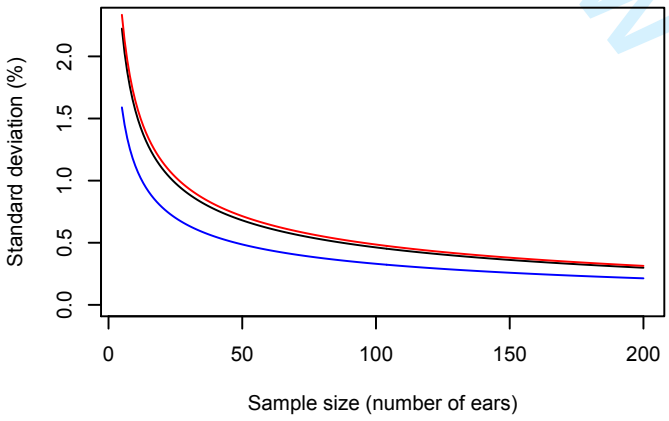
(a)



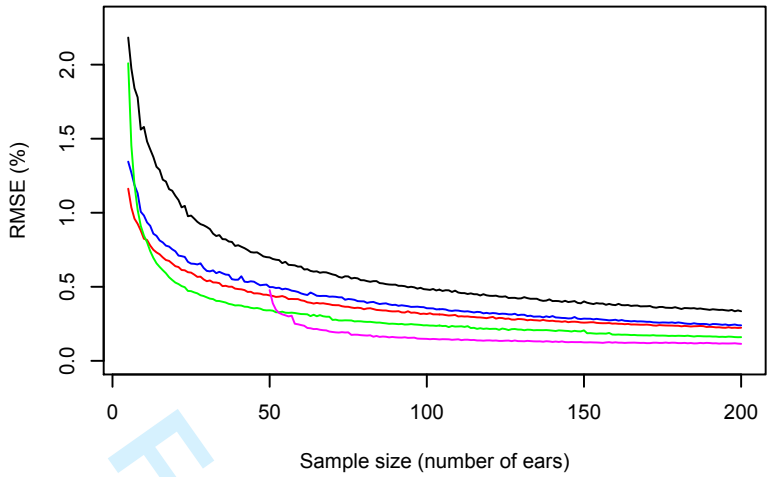
(b)



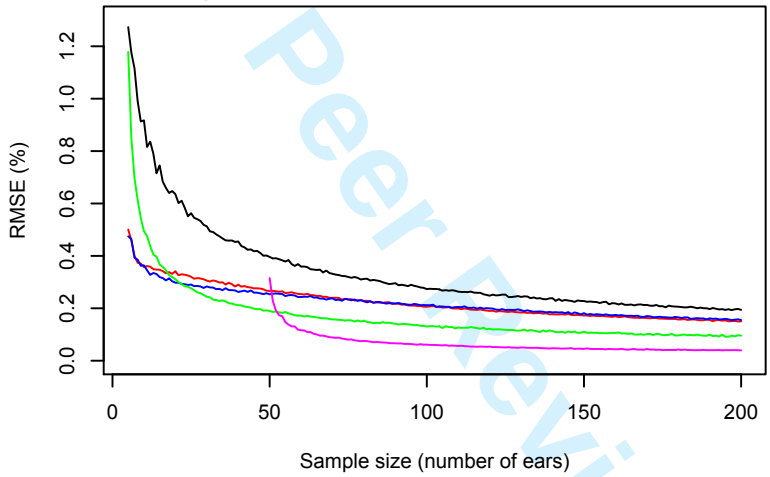
(c)



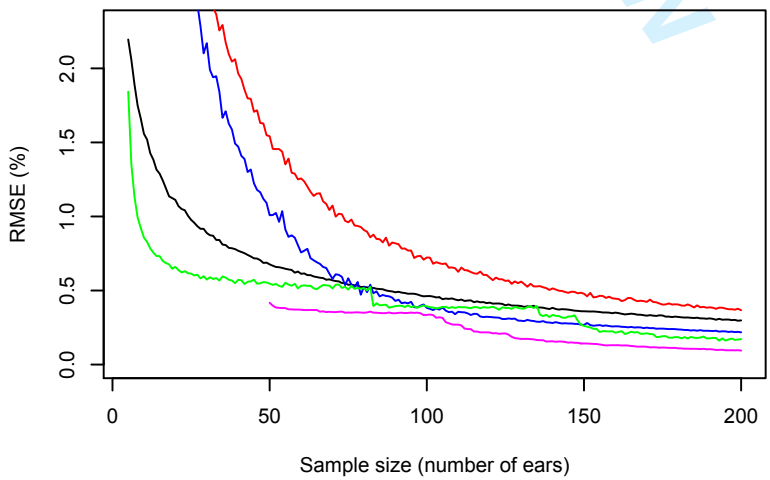
(a)



(b)

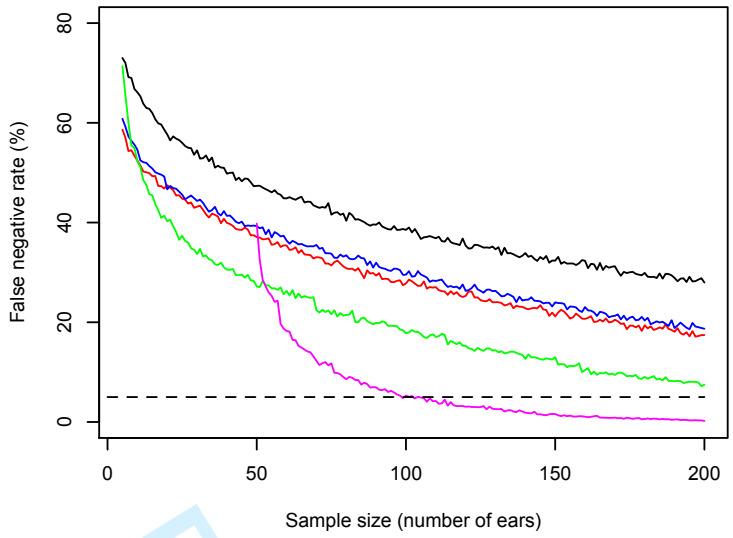


(c)

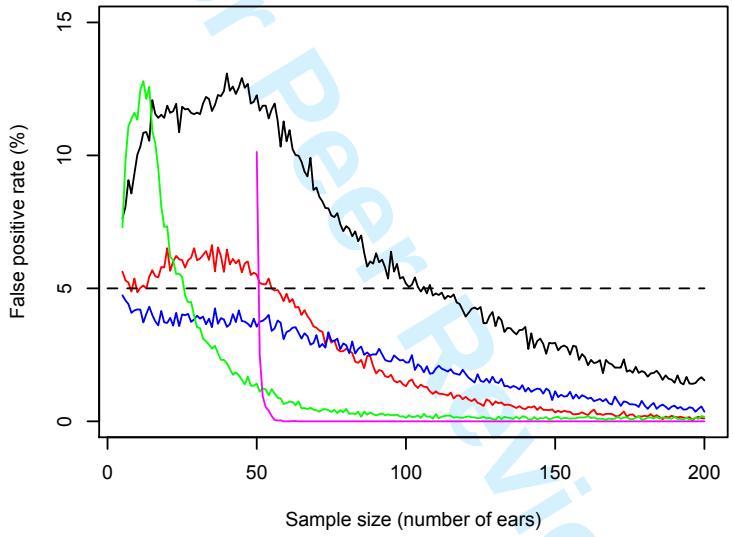


Simple random Ratio Regression 5 strata 50 strata

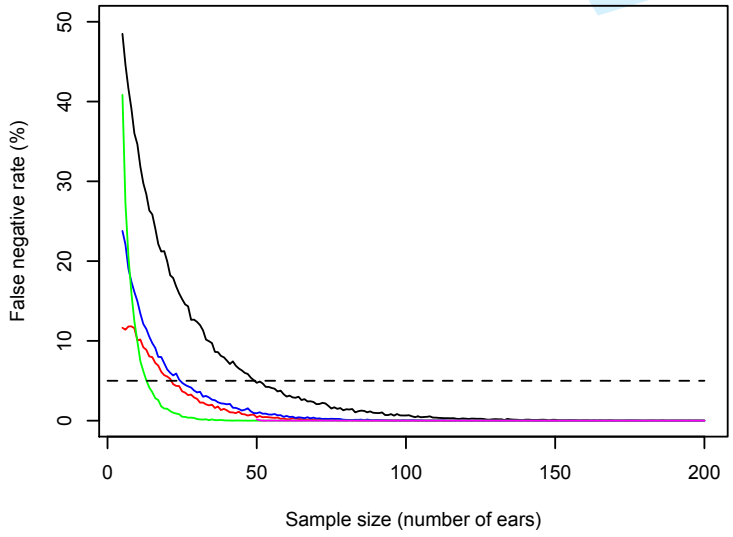
(a)



(b)



(c)



Simple random Ratio Regression 5 strata 50 strata

Annexes 4

Design of a decision support tool for coexistence at farm and regional level. Meillet, A., Angevin, F., Bensadoun, A., Cathary, A., Huby, G., Monod, H., Messéan, A. In *GMCC 2013 - 6th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains* (Lisbon, France, November 2013).

Design of a decision support tool for coexistence at farm and regional level

Anne Meillet[a,*], Frédérique Angevin[a], Arnaud Bensadoun[b], Agnès Cathary[c], Guillaume Huby[c], Hervé Monod[b], Antoine Messéan[a]

a. *INRA, UAR 1240 ECOINNOV, F-78850 Grignon, France*

b. *INRA, UR 341 MIAJ, 78352 Jouy-en-Josas, France*

c. *GEOSYS, Balma, France*

E-mail: anne.meillet@grignon.inra.fr*, tel: +33 5 61 28 50 95

ABSTRACT

The European Commission has commissioned several studies on coexistence that have used simulation results of spatially explicit gene flow models, such as MAPOD® for maize, or GeneSys® for rapeseed. These models predict the adventitious presence (AP) of GM grains in non-GM fields at the landscape scale. However the results' uncertainty is not quantified. Moreover, most of the models require input data on climate, land use and crop management practice which might not always be available. Our aim is to design a dynamic and flexible decision support tool (DST), based on MAPOD®, for stakeholders in the agricultural and food production sectors. It can compute the expected AP and its probability distribution in non-GM maize fields at different times of the season and under different management scenarios. Integrated through a web interface, the DST is designed to be an operational help to manage coexistence, used by stakeholders from farmers to policy makers.

KEY WORDS: Adventitious Presence, Coexistence, Genetically Modified Organisms, Decision-Support-Tool, Gene Flow modelling, maize, Public GMO register, Risk management

INTRODUCTION

Since the first cultivation of GM crops in Europe and given the public opinion about this subject, the feasibility of a GMO free sector has been explored. In 2003, the European Commission (EC) defined coexistence as the ability of farmers to make a practical choice between conventional, organic, and GM crop production, in compliance with the legal obligations for labelling and/or purity standards (EC, 2003a). Recommendations were made by the EC to give farmers freedom of choice (EC, 2003 & 2010). Given the diversity of agricultural production systems, of farms, and of economic and physical conditions in the European Union (EU), the legal measures have to be defined by each Member State.

In 2009, 15 of the 27 member states (MS) enacted measures on coexistence (EC, 2009). The measures mainly included individual measures to be implemented at plot level by the farmer growing GM plants, in order to respect the tolerance threshold of 0.9% set out in community legislation (EC, 2003b). These measures mostly define the required isolation distance, which can be partially or fully replaced by buffer zones between GM and non-GM fields in which non-GM varieties are grown and treated as GM crops.

Research projects related to coexistence were conducted in the Framework Programme for Community Research. The SIGMEA project, which ran from April 2004 to May 2007, studied the temporal and spatial gene flow of GMOs across Europe in crop production systems to help competent authorities establish appropriate coexistence measures. The Co-Extra project ran from 2005 until 2009 and studied and validated biological containment methods, modelled the organization of supply chains, and provided practical tools and methods for implementing coexistence. The PRICE project (PRactical Implementation of Coexistence in Europe), which began in December 2011 and will run

until December 2014, aims to assess whether cost-effective coexistence strategies are feasible for farmers, agro-food supply chains, and consumers.

One of the EC recommendations is that "the choice of measures should take into account the regional and local constraints and situations, as well as the specific nature of the crop concerned"(EC, 2003a). In this context, several approaches to gene flow modelling have been developed to help determine which field scale practices would make it possible to comply with existing purity thresholds at harvest.

Angevin et al., (2008) designed a spatially explicit model of maize pollen dispersal, MAPOD®, which takes into account climate conditions, some varietal characteristics, crop management, and the shape and position of the plot. It predicts the adventitious presence (AP) of GM maize in non GMO fields by taking into account local conditions, like distance between fields, wind direction, flowering dates, fields' shape and size. It was previously used as a decision support tool in coexistence studies and also used for deliberations to define coexistence rules and decision rules (Bock, Lheureux, Libeau-Dulos, Nilsagård, & Rodriguez-Cerezo, 2002; EC, 2003c; Messéan, Angevin, Gómez-Barbero, Menrad, & Rodríguez-Cerezo, 2006).

MAPOD® is specifically designed to be used by trained users who have all relevant data at their disposal: climatic data, variety characteristics, variety allocation in the agricultural landscape. Because in practice not all users have the necessary data, our project aims to design a cost-effective and practical decision-support tool (DST) for end-users, including regulatory bodies, crop advisors, and farmers. As they depend on biological phenomenon, input data are surrounded with uncertainties. The DST will account for the uncertainties linked to the input data by giving the distribution of probability of AP in non-GM maize fields. The users will thus be aware of the risk that the AP might be above a

given threshold. The DST also had to be responsive enough to facilitate exchanges between users. To facilitate its use, the DST will not need powerful computers and will respond quickly enough to allow real time discussion between users.

GENERAL SPECIFICATIONS

As mentioned above, the DST is designed to be used at different periods of maize growing season: before sowing ("*ex ante*"), after sowing and before flowering ("*ex post* 1"), after flowering and before harvesting ("*ex post* 2"). The adaptability of the DST to local conditions, time of use and available inputs will be achieved through a library of Bayesian models. The most suitable model for each use will be chosen. The models come from a set of gene flow models developed through a Bayesian approach, allowing probabilistic prediction of AP (Bensadoun, Monod, Angevin, Makowski, & Messean, 2013). Throughout the growing season and depending on the organisation which will use the DST, the amount and certainty of data will increase and thus the uncertainties of the results will be reduced.

Over time, the data will become more and more certain. Milestones beyond which uncertainty will decrease (due to new data added or a decrease in data uncertainties) have been identified:

- before sowing, during cropping system's planning;
- just before sowing, when the sowing dates are more certain;
- between sowing and flowering, when sowing dates are known;
- between flowering and harvesting, when flowering dates are known.

These milestones and the evolution of the DST' use is illustrated in Figure 1.

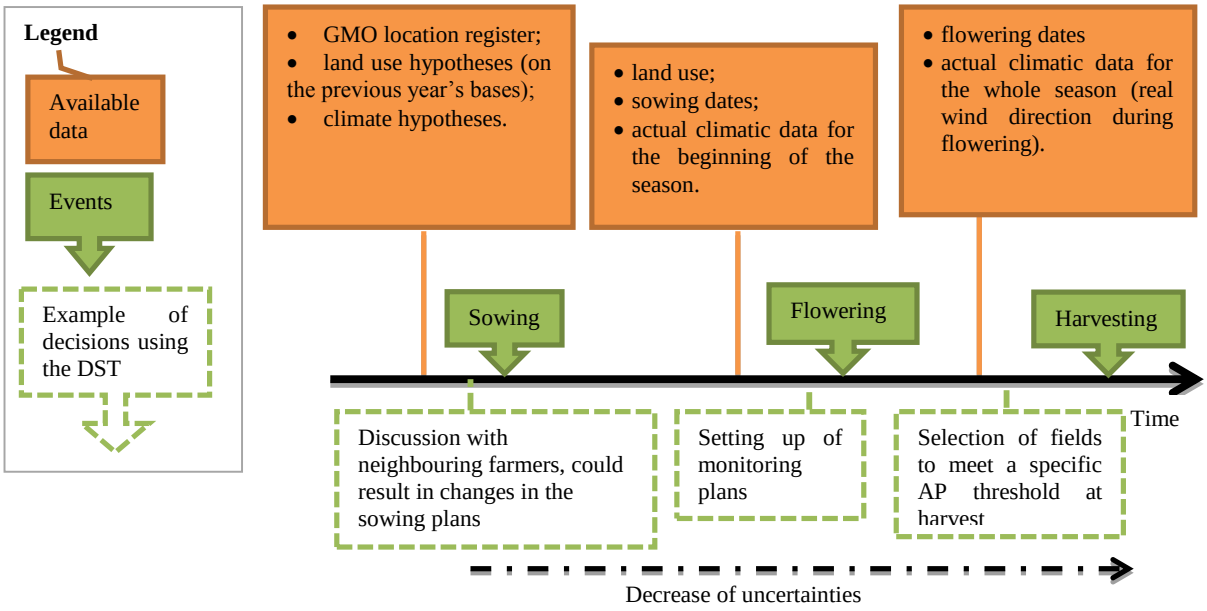


Figure 1: Evolution of the DST use throughout the growing season.

REQUIRED INPUTS

Spatial data are required as an input. Since the format needed by the DST for spatial data is widely used, users will also be able to use their own geographical data. They could also use land-parcel identification systems (LPIS), which were established in the framework of the Common Agricultural Policy to identify land use. The LPIS' access rights and characteristics depend on the MS (Milenov & Kay, 2006). If their access rights allow it, it will be possible to enter them in the DST.

At the beginning of the season, it's quite unlikely that the whole land use information is contained in the spatial data. Users could use historical data to make assumptions about next season's land use.

Information about GM crops fields is more likely to be available before sowing because MS are required by law to establish public registers in which the location of cultivated GMOs is recorded[1]. Each MS establishes a timeline for their register. For the sake of decision-making, the earlier is the better. Within PRICE, a harmonized European location register for GMOs, in the form of a geographical information system-based software, is being developed by the German Federal Office of Consumer Protection and Food Safety (BVL). This software will be interfaced with the DST; location of GM fields will thus be available.

Spatial data (field boundaries and parcels attributes such as land use, type of crop, and date of sowing) will be editable through the DST interface. The field manager (see Figure 2) will also allow adding new parcels.

[1] Directive 2001/18/EC of the European Parliament and of the Council – Article 31

Figure 2: Field manager.

Climatic data provided by GEOSYS and available for whole Europe will be extracted from an external climatic database. It is based upon a forecast model on a 10 km × 10 km grid allowing weather forecasts 9 days ahead. The DST will be able to make direct queries on the database; the user will only have to provide the location and dates of the simulation.

The other input data will be entered by the users and the DST will allow data sharing between users.

USE OF THE BAYESIAN MODELS

Different kinds of gene flow models will be implemented in the DST. The statistical part of the work is described in Bensadoun et al. (2013). The models are selected upon the available data entered in the DST.

The first kind of Bayesian models only requires the distance between fields, i.e. spatial data. One of these models has been implemented in the first DST's prototype (see Figure 3). Being the models that require the least amount of data, its output is associated with the highest uncertainties. If the user doesn't have information on land use other than GMO, he could make assumptions about crop allocation, for example by using historical land use data or by using the "worst case scenario", i.e. fields where land use is unknown would be considered as non-maize fields (no conventional pollen available around to reduce the mitigate the GM pollen pressure).

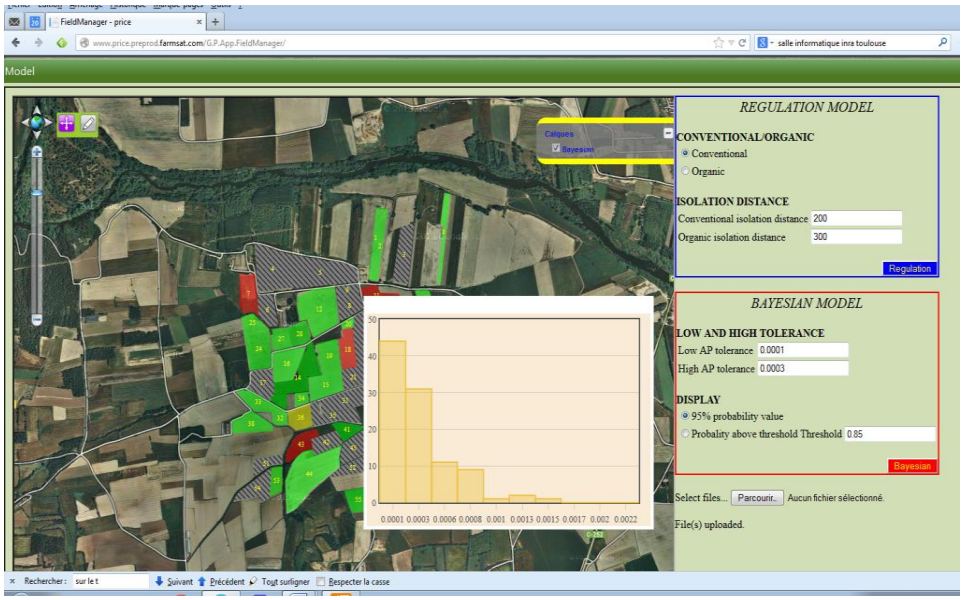


Figure 3: Output from a simulation made with the prototype in Foixà, Spain. The graph represents the probability distribution in a conventional maize-field.

When not only spatial data but also climatic data (wind effect and/or temperature) are available, a second kind of models will be used, considering both distances and wind distribution as input variables. When used early in the growing season, historical wind data could be used in the model. Uncertainty will decrease when actual climatic data will become available for simulations carried out later in the season.

A third kind of models will integrate flowering dynamics in addition to the variables listed above, decreasing the uncertainty of the results compared to the previous two kinds of models. Nevertheless they require data on crop management practices and varietal characteristics besides spatial and climatic data.

As stated before, models will be selected depending on the data available at the time of use of the DST. The tables below illustrate this selection strategy for three periods in the growing season: before sowing (Table 1), between sowing and flowering (Table 2) and between flowering and harvesting (Table 3).

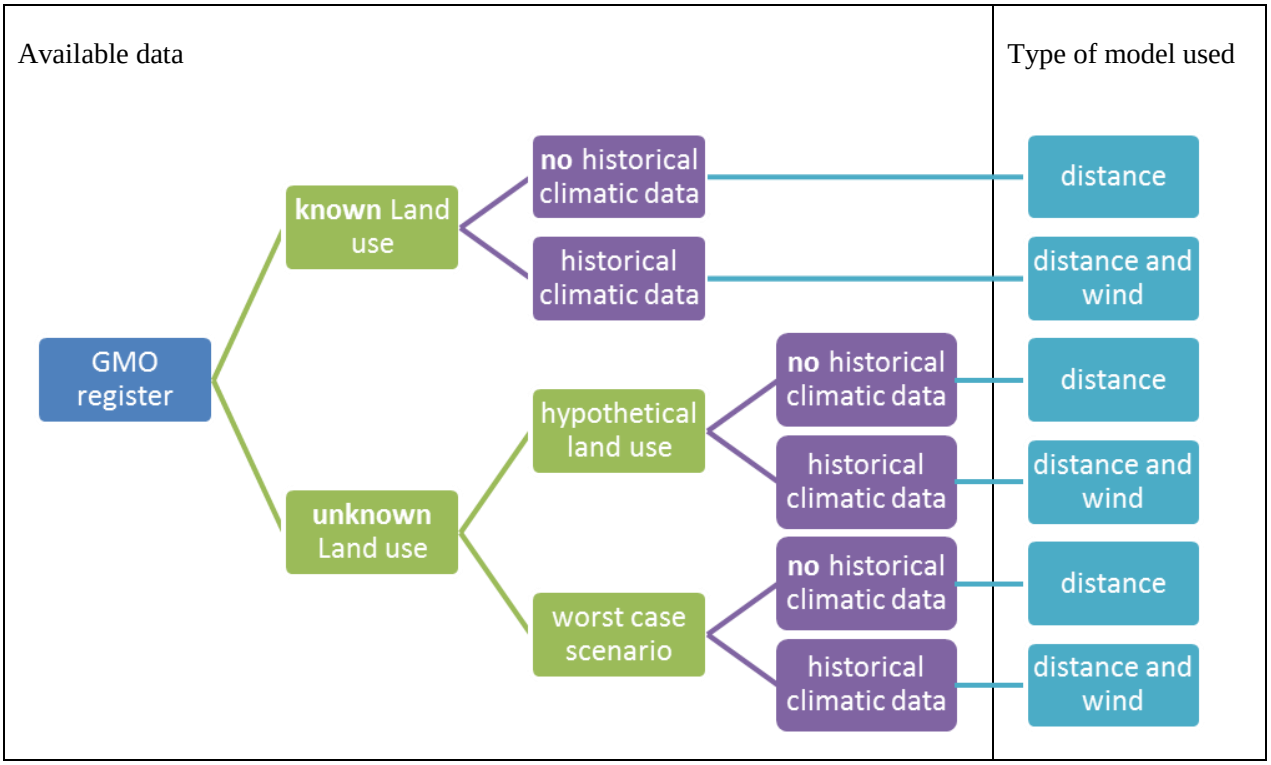


Table 1: Choice of model depending on available data for an "*ex ante*" situation, before sowing.

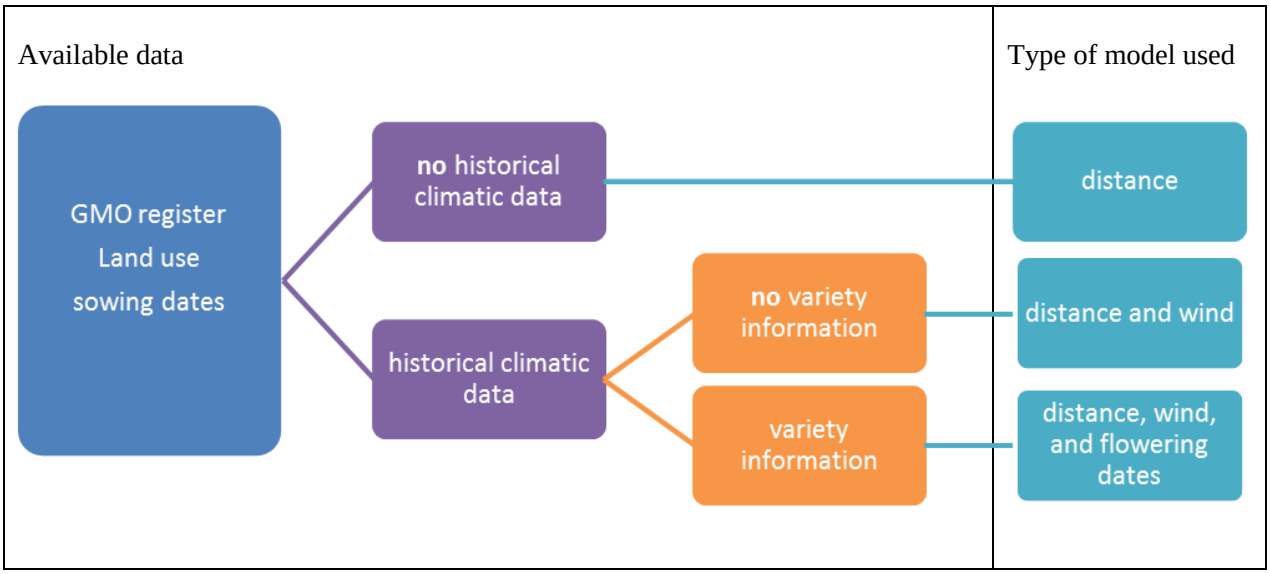


Table 2: Choice of model depending on available data for an "*ex post* 1"situation, between sowing and flowering.

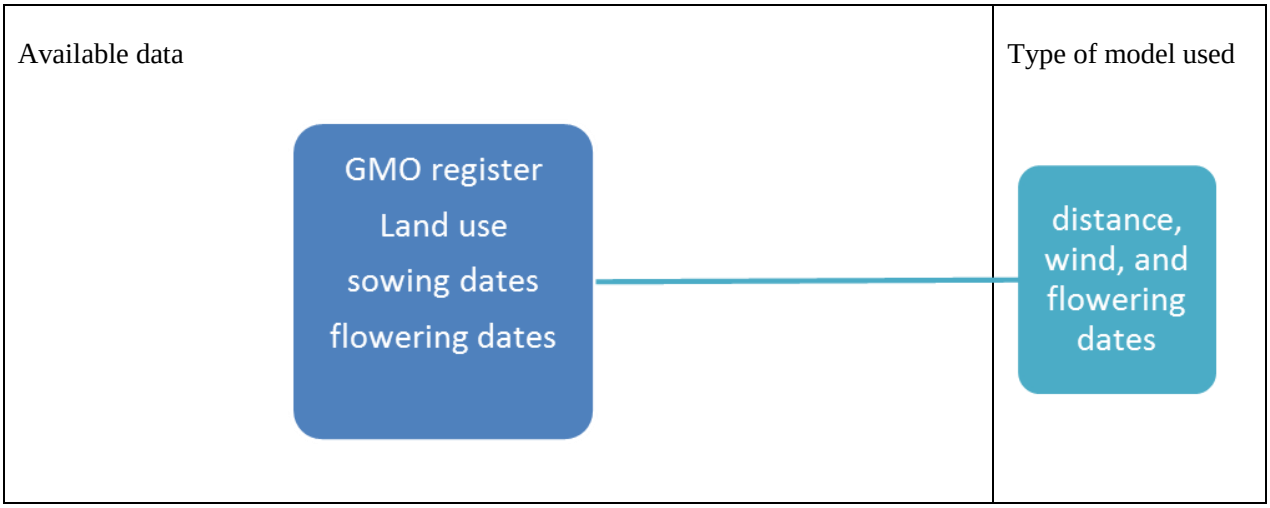


Table 3: Choice of model depending on available data for an "*ex post* 2"situation, between flowering and harvesting.

ERGONOMICS

The DST runs on a web server which means that users only need a computer with a web navigator and an internet connection to access it. This will help widen the possibilities of use, because powerful computers might not be always available on-site. The algorithms used for the models are optimized so that response time is as low as possible, allowing real-time discussion between users.

Furthermore, the company developing the DST, GEOSYS, was created 25 years ago and has since then developed software using geographic information. The DST is partly based on existing products, which are currently used by agricultural professionals in many countries. It will thus capitalise on GEOSYS' experience in designing tools that can be used by growers, farmer advisors, or decision makers. The DST will benefit from the use of the up to date web technologies such as HTML5 that improve the user experience.

STAKEHOLDERS' NEEDS

Stakeholders' needs to manage coexistence have already been addressed in previous studies. For example, Hannachi, Coleno, and Assens (2009) analysed strategies undertaken by cooperatives to manage coexistence in France before 2008 and concluded that coexistence was possible if grain merchants coordinate. To further enhance the knowledge of stakeholders needs, semi-structured interviews have been conducted, so far mainly with policy makers in France. A major expectation is to help them plan monitoring of compliance. In MS where GM crops are not grown, however, few potential stakeholders agreed to be interviewed because they didn't feel directly concerned by the

issue. Addressing stakeholders from MS where GM maize is actually sown (e.g. Spain, Portugal, Ukraine), or planning to grow GM maize (e.g., UK) will add elements to our analysis.

DISCUSSION AND CONCLUSION

The DST is adaptable to different situations, depending on its time of use and on the availability of data. It is designed to be easy to use and to meet the needs of potential users. The users can be policy makers, cooperatives or elevators, advisers, as well as farmers in MS where GM maize is currently grown. The DST has also been designed for potential users from MS which might allow GM maize cultivation in the future. It is hence important to design an evolving DST.

The probability distributions of AP take into account the uncertainty of biological phenomena as well as the amount of available data. The DST is thus flexible because its outputs are adapted to the input data. Uncertainties on the AP are therefore explicitly expressed and users therefore have to interpret probability distributions rather than a single value. This could make decision-making more difficult but more reliable.

The next step will be to use the DST in real agricultural landscapes under different scenarios of coexistence: percentage of GM maize in the landscape, more or less scattered fields, various climatic conditions, and different coexistence measures, such as buffer zones, isolation distance or flowering shifts. Prototype changes will also be made based on identified stakeholder needs. The DST's final prototype is to be delivered at the end of PRICE, in December 2014.

ACKNOWLEDGEMENT

This work is funded by the EU project PRICE (PRactical Implementation of Coexistence in Europe), contract number 289157.

REFERENCES

Angevin, F., Klein, E. K., Choimet, C., Gauffreteau, A., Lavigne, C., Messéan, A., & Meynard, J. M. (2008). Modelling impacts of cropping systems and climate on maize cross-pollination in agricultural landscapes: The MAPOD model. *European Journal of Agronomy*, *28*(3), 471–484. doi:10.1016/j.eja.2007.11.010

Bensadoun, A., Monod, H., Angevin, F., Makowski, D., & Messean, A. (2013). Modeling of gene flow by a Bayesian approach: A new perspective for decision support. In *GMCC-13 proceedings*. Lisbon 2013.

Bock, A.-K., Lheureux, K., Libeau-Dulos, M., Nilsagård, H., & Rodriguez-Cerezo, E. (2002). *Scenarios for co-existence of genetically modified, conventional and organic crops in European agriculture* (p. 133). Retrieved from http://ftp.jrc.es/EURdoc/eur20394en.pdf

EC. (2003a). Commission recommendations of 23 July 2003 on guidelines for the development of national strategies and best practices to ensure the co-existence of

genetically modified crops with conventional and organic farming. *Official journal of the European Union*, *46*(July), 36 – 47.

EC. (2003b). Regulation (EC) No 1830/2003 of the European Parliament and of the council of 22 September 2003 concerning the traceability and labelling of genetically modified organisms and the traceability of food and feed products produced from genetically modified o. *Official journal of the European Union*, *31*(1830), 24–28.

EC. (2003c). Regulation (EC) No 1829/2003 of the European Parliament and of the council of 22 September 2003 on genetically modified food and feed. *Official journal of the European Union*, *L 268*, 1 – 23.

EC. (2009). Report from the commission to the council and the european parliament on the coexistence of genetically modified crops with conventional and organic farming. *SEK*, 12 p. Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0153:FIN:en:PDF

Hannachi, M., Coléno, F., & Assens, C. (2009). Collective strategies and coordination for the management of coexistence : the case studies of Alsace and western South of France . In *4th international conference on coexistence between genetically modified (GM) and non-GM based agricultural supply chains (GMCC09)*. Melbourne Australia.

15

Messéan, A., Angevin, F., Gómez-Barbero, M., Menrad, K., & Rodríguez-Cerezo. (2006). *New case studies on the coexistence of GM and non-GM crops in European agriculture.* (pp. 1 – 112). Retrieved from http://www.jrc.es/home/pages/eur22102enfinal.pdf

Milenov, P., & Kay, S. (2006). Status of the implementation of LPIS in the EU member states. In JRC Scientific and Technical Reports (Ed.), *12th MARS PAC Annual Conference* (pp. 41 – 47). Toulouse.

Annexes 5

Livrable D4.25 du projet Européen PRICE : *Upadated database of geneflow datasets*

Deliverable D4.25

Deliverable D4.25

Updated database of geneflow datasets

PRICE

PRactical Implementation of Coexistence in Europe

PRICE Project

Project no: FP7-KBBE-2011-5

Expected delivery date: August 2012

Actual delivery date: September 2012, release 1.0

SEVENTH FRAMEWORK PROGRAMME

This deliverable provides a table describing the validated datasets that will be used for the Bayesian meta-analysis of maize gene flow studies. The existing SIGMEA database has been updated and completed with datasets from Spain (commercial fields), Germany (field experiments) and Portugal (experimental fields). Special attention has been paid to the gathering of relevant landscape-level and multi-source gene flow datasets. As for other crops, the existing SIGMEA database will be used as no relevant and usable gene flow datasets has been produced since then (www.inra/SIGMEA).

Source	Site-Year	Type	Sampling	Data				PRICE	Ref.
				Distance	Wind	Flowering	AP Estimation		
SIGMEA-JKI	Germany 2000	DC1	96 sampling points (8 to 30 cobs per point)	From the sampling point to the edge of the donor field (3m to 50m)	Direction and speed known (every half hour)	Male flowering period known * donor : 21 to 30 of july * receptor: 14 to 24 of july	Sowing of sampled grain and herbicide application, => AP : proportion of resistant plants	Yes - ready	1,2
JKI	Germany 2001	DC1				Number of flowering plants at diffrent dates (over 3 weeks) at sampling points			
JKI	Germany 2001	DD2	30 sampling points (~30 cobs per point)	From the sampling point to the edge of the donor field (50m to 200m)		No			
AIP OGM	France (Montargis) 1998	DC1	31 cobs per row 101 row=> 3131 cobs	From the sampling point to the edge of the donor field (0.4 to 71.9m)	Direction and speed known (every 3 hours)	Flowering date male and female	Counting blue grains in yellow ears	Yes - ready	3
AIP OGM	France (Montargis) 1999	DC1 DD1	DC1 :31 to 139 cobs per zone, 52 zones DD1 : 10 to 272 cobs per zone, 42 zones	From the sampling point to the edge of the donor field (0 to 140m)	Direction and speed known (every hour)	Number of flowering plants (out of 100 plants)			10
WUR	The Netherlands 2006-2007	DD1	5 cobs/Blocs 21 blocs	From the sampling point to the edge of the donor field (25 to 106m)	Direction and speed known (every hour)	Flowering date male and female (40 plants evaluated 3 times a week)	PCR	Yes - ready	4
	The Netherlands 2006-2007	DD1	5 cobs/Blocs 16 blocs	From the sampling point to the edge of the donor field (250 to 300m)					

JKI	Germany 2005-2006-2007	DD2	94 to 280 sampling points (20 cobs per point)	From the sampling point to the edge of the donor field (25 to 300m)	Direction and speed known (every hour)	Number of plants at a certain stage (20 to 60 plants in the donor, 53 to 96 in the receptor, evaluated every 2 or 3 days)	PCR	Yes - ready	5
	Germany2007-2008	DD2	20 cobs/points 4 points/distance 9 distances =>720 cobs	From the sampling point to the edge of the donor field (25 to 85m)		Number of plants at a certain stage (12 to 16 plants in the donor, 16 in the receptor, evaluated every 2 or 3 days)		Yes – but don't have	6
IPSS	Portugal 2002-2003	DD2	10 cobs per row, 6 sampled rows (only odd rows)	Edge to edge distance (40m to 250m)	Direction and speed known (mean and max for flowering period)	Indirectly observed with frequencies of yellow grains in each part of the ear	Counting yellow grains in white ears + PCR	Yes – but don't have	7
	Portugal 2005-2007	DC2	2 cobs/point on a grid from 9x9 to 36x18 or 24x12m	From the sampling point to the edge of the donor field (40 to 300m)			PCR		
IRTA	Spain 2004	DC1	*	From the sampling point to the edge of the donor field	Measured in « wind run values » expressed in wind km	Flowering dates male and female visually evaluated on 150 plants for each field	Counting yellow grains in white ears + PCR	Yes – but don't have	8
IRTA	Spain 2004-2008	MSR	**	From the sampling point to the edge of the donor field	Direction and speed known (mean and max every day)	Flowering dates relative to the first field in bloom	PCR	Yes – not ready	9

* :3 sampling methods :

- For the field in the main wind direction:

  - Regular grid 10x10m (sampling point every 10m in both orientation, abscises and ordinates, 3 cobs per points) and additional samples at 0, 2, 5 and 10m from the nearest edge of the GM field.
  - Transect sampling in the prevailing wind direction (Samples at 0, 2, 5, 10, 20, 40, 80 and 120m)

-For others fields

Regular grid 20x20m (sampling point every 20m in both orientation, abscises and ordinates, 3 cobs per points) and additional sample at 0, 2, 5, and 10m from the nearest edge of the GM field.

** : Stratified sampling : 4 transect are drawn, 1 ears is sampled on the transect at 0, 3 and 10m from the border (4x2x3=24 ears). An additional ear is sampled at each croos transect (4 ears). In total, there are 28 ears per field.

Type :

DC1 : The receptor surrounds the donor continuously

DC2 : Receptor and donor side by side continuously

DD1 : The receptor surrounds the donor with discontinuity

DD2 : Receptor and donor side by side with discontinuity

MSR : Multisource design at landscape scale in real coexistence situations

PRICE :

Yes : The data set will be used for PRICE

Ready : Raw data were collated, the data set is clean and ready to use

Not yet ready : Raw data were not cleaned, the data set is not ready to use

Don't have yet: The dataset is interesting and could be retrieved from publications but still to retrieve raw data.

References:

1. Loos, C., Seppelt, R., Meier-Bethke, S., Schiemann, J., Richter, O., 2003. Spatially explicit modelling of transgenic maize pollen dispersal and cross-pollination.Journal of Theoretical Biology 225, 241-255.

2. Meier-Bethke, S., Schiemann, J., 2002. Cross pollination of GM corn in adjacent non-transgenic corn fields. In: Proceedings of the 7th International Symposium on the Biosafety of Genetically Modified Organisms. Beijing, China, pp. 295.

3. 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.

4. Van De Wiel, C.C.M., Groeneveld, R.M.W., Dolstra, O., Kok E.J., Sholtens, I.M.J., Thissen, J.T.N.M., Smulders M.J.M., Lotz, L.A.P., 2009. Pollen-mediated gene flow in maize tested for coexistence of GM and non-GM crops in Netherlands: Effect of isolation distances between field. NJAS – Wageningen Journal of Life Sciences 56-4, 405-423.

5. Langhof, M., Hommel, B., Hüsken, A., Njontie, C., Schiemann, J., Wehling, P., Wilhelm, R., Rühl, G., 2010. Coexistence in maize: Isolation distance in dependence on non-GM field depth and separate edge harvest. Crop Science 50, 1496–1508.

6. Rühl, G., Langhof, M., 2011. Coexistence in maize: Effect of the genetically modified maize field depth on pollen-mediated gene flow. Crop Science 51, 2186–2193.

7. Quedas, F., Andrade Silva, E., 2009. Maize Pollen Flow in Contiguous and Noncontiguous Field Plots. Proceedings 4th International Conference on Coexistence between Genetically Modified (GM) and non-GM based Agricultural Supply Chains. 10 - 12 November 2009, Melbourne, Australia.

8. Palaudelmàs, M., Melé, E., Monfort, A., Serra, J., Salvia, J., Messeguer, J., 2012. Assessment of the influence of field size on maize gene flow using SSR analysis. Transgenic Research 21, 471-483.

9. Messeguer, J., Penas, G., Ballester, J., Bas, M., Serra, J., Salvia, J., Palaudelmàs, M., Melé, E., 2006. Pollen-mediated gene flow in maize in real situations of coexistence. Plant Biotechnology Journal 4, 633–645.

10. INRA, 1998. Internal report from Etienne Klein.

Résumé

Modélisation des flux de gènes par approches Bayésiennes. Application à l'aide à la décision pour la coexistence de cultures OGM et non OGM.

Dans le débat autour des OGM en Europe, la question d'une possible coexistence entre cultures OGM et non-OGM tient une place centrale. Est-il possible de cultiver des OGM sans pénaliser les cultures conventionnelles voisines en raison des risques de flux de gènes à l'échelle des paysages ? Les règles actuellement envisagées reposent uniquement sur des distances de séparation fixées à l'échelle nationale et ne permettent pas d'adaptation à la diversité des situations possibles. Le travail de thèse présenté dans ce mémoire a pour objectif de développer et étudier une approche par modélisation, alternative aux règles plus rigides actuellement envisagées, permettant d'adapter les mesures de coexistence, à la diversité des situations rencontrées en pratique. La démarche repose sur la conception d'un modèle prédictif, permettant la simulation des flux de pollen entre champs cultivés et la prédiction du taux de pollinisation croisée dans les champs non OGM, en valorisant au maximum l'information locale lorsqu'elle est disponible. L'inférence des paramètres du modèle est réalisée dans un cadre bayésien de manière à intégrer de façon explicite aussi bien la variabilité intrinsèque du système étudié que les approximations et incertitudes sur les processus modélisés. Ce cadre probabiliste offre l'avantage non seulement de mieux prendre en compte l'incertitude dans l'évaluation des risques mais aussi de permettre une mise à jour aisée des valeurs des paramètres du modèle à mesure que de nouvelles données sont disponibles.

Mot-clés : statistique, agronomie, statistique Bayésienne, coexistence, échantillonnage, aide à la décision

Abstract

Modeling of gene flow at the landscape scale by a Bayesian approach. Application to decision support for the coexistence between GM and non GM crops.

In the European debate about GMOs, the coexistence between GM and non-GM crops is a major stake. Is it possible to cultivate GMOs without penalizing neighboring conventional crops because of the risk of gene flow across landscapes ? The rules that are presently considered only rely on separation distances that are fixed at a national scale. The study presented in this memoire aims at developing a model-based approach, alternative to rigid rules currently under consideration, to adapt the coexistence measures to the diversity of situations encountered in practice. The approach is based on the design of a predictive model for the simulation of pollen flow between cultivated fields and prediction of outcrossing rates in non-GM fields, valuing to the highest local information when available. The inference of the model parameters is performed in a Bayesian framework to explicitly integrate both the intrinsic variability of the studied system as well as approximations and uncertainties in the modeled process. This probabilistic framework has the advantage not only to better take into account the uncertainty in the risk assessment but also to allow easy updating of model parameter values as new data are available.

Keywords : statistics, agronomy, Bayesian statistics, coexistence, sampling, decision support