Application de méthodes bayésiennes pour l'estimation des paramètres et l'évaluation de l'incertitude d'un modèle de prévision de la septoriose sur blé

Arnaud Bensadoun

2010

Rev. T. Bayes

MEMOIRE D'INGENIEUR AGRONOME
PRESENTE PAR
ARNAUD BENSADOUN

POUR L'OBTENTION DU
DIPLOME D'AGRONOMIE APPROFONDIE

SPECIALITE
SYSTEME DE PRODUCTION, ENVIRONNEMENT ET TERRITOIRE

APPLICATION DE METHODES BAYESIENNES POUR L'EVALUATION DE L'INCERTITUDE D'UN MODELE DE PREVISION DE LA SEPTORIOSE SUR BLE : *SEPTO-LIS*®

Stage effectué sous la direction de :
Daniel Wallach (INRA UMR AGIR Toulouse),
François Piraux (Arvalis Institut du végétal),
Et David Gouache (Arvalis Institut du végétal)

En interaction avec le groupe de suivi et de réflexion :
François Brun (ACTA ingénieur du RMT modélisation)
David Makowski (INRA/INA-PG UMR Agronomie Thiverval-Grignon)

Tuteur: Christophe Laplanche (ENSAT)
2ème correcteur: Philippe Grieu (ENSAT)

RESUME

La septoriose est la maladie fongique foliaire la plus dommageable sur blé tendre en France. La lutte chimique contre cette maladie est actuellement quasi systématique. Pour répondre aux exigences environnementales sans réduire la production de blé, Arvalis à développé un modèle mécaniste de prévisions de la contamination d'une parcelle de blé par la septoriose. L'objectif de l'étude est d'évaluer l'incertitude de ce modèle lorsqu'il est utilisé pour décider de la date de traitement optimale. Pour cela, nous appliquons des méthodes bayesiennes à l'estimation des paramètres et de la variance résiduelle du modèle. Nous avons testées plusieurs approches, la plus prometteuse est l'approche MCMC. Nous proposons ensuite une méthode pour évaluer l'incertitude dans la décision à partir des distributions obtenues dans l'estimation. Cette méthode permet une meilleure prise en compte de l'incertitude dans le conseil et la prise de décision.

ABSTRACT

Septoria leaf blotch (SLB) is the most damaging foliar fungal disease on wheat in France. Chemical control against this disease is currently almost systematic. In order to meet the environmental requirements without reducing wheat production, *Arvalis* (a French farmer-financed institution for research in agriculture) has developed a mechanistic model to predict the contamination of a wheat plot by SLB. The aim of this study is then to assess uncertainty of this model when it is used to decide the optimal date of treatment. To proceed we apply bayesian methods to the model for parameter and residual variance estimation. We tested several approaches; the most promising is the MCMC approach. We then propose a method to assess uncertainty in the decision from the distributions we dispose. The method allows a better account of the uncertainty in advising and decision making.

REMERCIEMENTS

Je tiens tout d'abord à remercier sincèrement Daniel Wallach et François Brun pour m'avoir intégré dans ce projet mais aussi pour leur soutien aussi bien statistique qu'informatique. Leur expérience et l'apport de bases méthodologiques m'ont permis de mener à bien cette étude.

Merci également à Christophe Laplanche sans qui je n'aurais pas pu découvrir l'utilité, la puissance et la beauté des statistiques bayesienne.

Je tiens tout particulièrement à adresser de chaleureux remerciements à Jérôme Thiard (développeur informatique à l'INRA et ancien de l'ENSAT) pour ses conseils toujours avisés dans des domaines aussi variés que la programmation ou le dosage optimal des apéritifs. Son parcours autodidacte et son intérêt pour mon projet m'ont permis de renforcer ma confiance en moi et ma persévérance notamment pour l'implémentation des algorithmes nécessaires à l'étude.

Merci à Raphael et Houcine de m'avoir sollicité pour du soutien statistiques. Ils m'ont permis d'élargir mes compétences (notamment à l'analyse multi varié et l'analyse de variance) et de renforcer mon attirance pour les applications statistiques aux problèmes agronomiques.

Merci à Areski, Abdé, David, Damien et plus globalement à toutes les personnes qui prennent leurs pauses devant le bâtiment A de l'UMR AGIR pour toutes discussions passionnantes ou parfois incompréhensibles que nous avons pu avoir.

Un très grand merci à la R Development Core Team pour leur merveilleux logiciel, gratuit et open source qui plus est !

Enfin merci à Bayes, Gauss, Markov, Kolmogorov, Fisher, Box, Metropolis, Hastings et Gibbs d'avoir autant fait avancer le domaines des statistiques et d'avoir publié leurs théories assez clairement pour que je puisse m'en servir.

PREAMBULE

Mon stage s'est déroulé au sein de l'UMR AGIR du centre INRA Toulouse pour le compte d'ARVALIS-Institut du végétal.

ARVALIS-Institut du végétal est un institut de recherche appliquée en agriculture. La mission principale de cet institut est de mettre au point et de diffuser des informations et des techniques permettant aux producteurs de céréales à paille (blé, orge, avoine, triticale, seigle, sorgho ...), de protéagineux(pois, féverole, lupin), de pomme de terre, de maïs et de fourrages, de s'adapter à l'évolution des marchés agro-alimentaires et de rester compétitifs au plan international, tout en respectant l'environnement. Son programme d'activités est réalisé avec le soutien financier des filières (Intercéréales, GNIS, FNPSMS, UNIP, ITPT, ...) et la participation financière du Compte d'Affectation Spécial pour le Développement Agricole et Rural (CASDAR) géré par le Ministère de l'Agriculture et de la Pêche.

Pour permettre aux agriculteurs de concilier rentabilité, qualité, environnement et adaptation à l'évolution des marchés, ARVALIS-Institut du végétal met au point et teste méthodes ou solutions innovantes permettant d'optimiser les facteurs de production et de réduire les coûts. Ces méthodes ou solutions innovantes prennent ensuite, le plus souvent, la forme d'outils d'aides à la décision.

Une part grandissante de ces outils d'aides à la décision est aujourd'hui représentée par des modèles mathématiques de prévision (de risques, de rendement, de maladie, d'efficacité d'une pratique…). L'utilisation de ces modèles nécessite évidemment une grande connaissance sur le comportement du système modélisé. ARVALIS dispose de ces connaissances cependant ce n'est pas toujours suffisant ; Des considérations d'ordres mathématiques, statistiques et informatiques sont en général aussi importantes que le reste. C'est donc pour prendre en compte ce type de considérations qu'ARVALIS et l'INRA ont formé un partenariat autour de questions méthodologiques sur l'utilisation des modèles mathématiques en agronomie.

C'est dans ce contexte que s'intègre le projet d'appliquer des méthodes bayesienne pour l'évaluation de l'incertitude de modèle dont j'ai eu la charge et que je vais présenter dans ce mémoire.

TABLE DES MATIERES

INTRODUCTION

En France, la septoriose, maladie fongique foliaire causée par *Septoria tritici* et *Septoria nodorum,* est la plus dommageable sur blé tendre (*Triticum aestivum*), pouvant engendrer des pertes de rendement de l'ordre de 50 q/ha dans les situations les plus exposées (Jorgensen et al. 2008). Le contrôle de la maladie dépend essentiellement de la lutte chimique. Cependant, cette lutte chimique est doublement remise en cause. Le développement de résistances aux matières actives par le pathogène a induit des baisses d'efficacité importantes (Maufras et al., 2006). De plus, la pression sociétale s'est concrétisée par l'engagement pris lors du Grenelle de l'Environnement de réduire l'utilisation des produits phytopharmaceutiques de 50%. Dans un tel contexte, il est important de rechercher un maximum d'efficacité et donc de rentabilité pour un programme de protection donné, et d'exploiter toutes les opportunités de réduction éventuelle du nombre de traitements. Le positionnement des traitements fongicides est une des voies d'optimisation de l'efficacité d'un programme. C'est la voie qu'a choisi *Arvalis-Institut du végétal* en développant un modèle mathématique de prévision de la contamination d'une parcelle de blé par la septoriose en fonction de variables agro-climatiques. Ce modèle est utilisé pour déterminer la date optimale du premier traitement contre la septoriose, à partir de laquelle découle le reste du programme de traitement. *Septo-LIS®* est un modèle mécaniste dans le sens où les processus biologiques étudiés sont explicitement représentés dans la structure même des équations. Ce type de modèle peut être traité à deux niveaux ; on peut étudier les équations individuellement, parce que chaque équation représente un processus que l'on peut étudier indépendamment du système global, mais on peut aussi étudier le modèle dans sa totalité, en tant qu'outil qui prédit le comportement du système complet. Dans notre cas, les processus comprennent par exemple la croissance des lésions et le transfert d'inoculum. Le modèle complet prédit le pourcentage de dégâts sur chaque feuille en fonction du temps. La formulation des équations vient de l'étude des processus individuels, c'est l'essence d'un modèle mécaniste. L'importance relative des deux niveaux dépend en grande partie des données et de leur relation avec l'utilisation prévue du modèle. Ainsi, *Septo-LIS®*, par son caractère mécaniste n'est pas adapté pour l'application de méthodes statistiques standards permettant d'évaluer son incertitude. L'objectif de notre étude est donc d'améliorer le modèle *Septo-LIS®* en proposant des méthodes permettant de quantifier l'incertitude qui existe sur ses prédictions. On utilise ici les méthodes bayesiennes car elles permettent d'intégrer différents types de variabilité dans l'estimation des paramètres. Ces méthodes sont bien connues, mais rarement appliquées dans leur ensemble aux modèles mécanistes.

On présentera dans une première partie le contexte, les enjeux et les objectifs de l'étude, avant de décrire dans une seconde partie, les matériels et méthodes employés pour réaliser l'étude. Ce stage ayant une composante méthodologique très prononcée, certaines des méthodes peuvent être considérées comme des résultats à part entière. Cependant pour plus de clarté nous avons décidé de conserver le schéma classique d'un article scientifique. La troisième partie sera consacrée à la présentation, l'interprétation et la discussion des résultats obtenus jusqu'ici. Enfin, avant de conclure ce mémoire nous dresserons les perspectives et potentiels de notre étude.

I. CONTEXTE ET OBJECTIFS

1. La septoriose

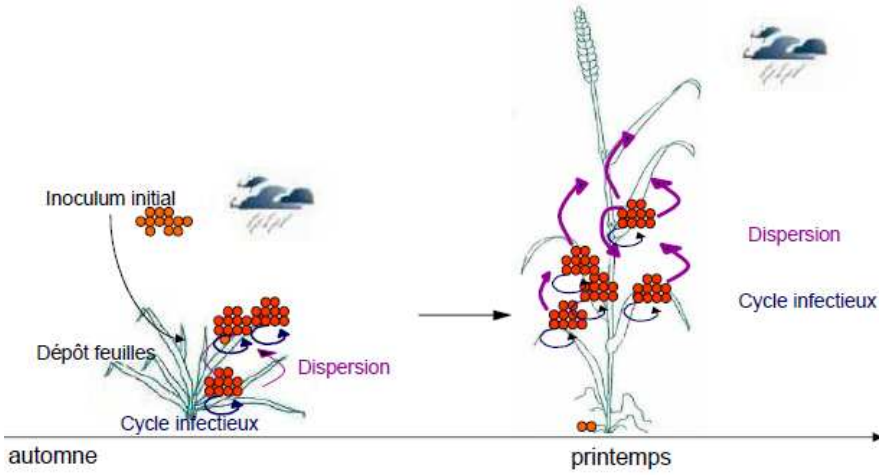
La septoriose est une maladie cryptogamique très fréquente dans le Nord de la France et très dépendante du climat. Elle est provoquée par deux parasites fongiques connus sous les noms de *Septoria tritici* et *Septoria nodorum*. Ces champignons provoquent des taches foliaires brunes et peuvent entraîner des baisses de rendement importantes. Actuellement, l'espèce *S. tritici* est largement dominante en France, alors que *S. nodorum*, qui peut également infecter les épis et les semences, est devenue très rare. Les régions les plus touchées par le développement de *S. tritici* sont celles qui connaissent le plus d'épisodes pluvieux, car la pluie et l'action éclaboussante des gouttes sur les feuilles du blé favorisent la dissémination des spores. Ainsi, il existe en France un gradient croissant de l'est vers l'ouest, où les précipitations et le climat océanique favorisent le développement de *S. tritici.*

1.1. Physiologie et modalité de contamination

Par reproduction asexuée, *S.tritici* produit des pycnides, se développant sur des nécroses rectangulaires, apparaissant elles-mêmes sur les feuilles les plus âgées de la céréale. Elles renferment des picniospores (ou pycnidiospores), exsudées puis projetées sur les feuilles supérieures grâce aux pluies automnales et printanières. Les contaminations se poursuivent tant que les pluies persistent. Mais il existe des limites physiques à la projection des picniospores : on estime qu'elle dépasse rarement 30 cm. La dernière feuille peut donc en théorie se trouver hors de portée d'une contamination, si l'inoculum est situé trop bas sur la plante. Mais le champignon présente également une phase de reproduction sexuée. D'une importance mineure, celle-ci contribue au brassage génétique des souches fongiques. Les ascospores sont dispersées par le vent, à l'automne mais aussi au printemps et plus tôt qu'on ne le pensait : du stade 2 noeuds environ jusqu'à la récolte.

La pluie reste, malgré tout, le moteur principal de l'épidémie sur lequel s'appuient tous les modèles agro-climatiques de prévision du risque septoriose (Shaw, 1987 ; Shaw et Royle, 1993). Des températures fraîches peuvent, par ailleurs, rendre moins efficaces les contaminations, ralentir les incubations mais à l'évidence, elles ne permettent pas de stopper totalement la maladie.

Figure 1:Développement de l'épidémie au fur et à mesure que le couvert se développe. (adaptée de Robert et al. 2008)



1.2. Pression et protection

Une forte présence de septoriose se traduit par des pertes de rendements de quelques quintaux à 30 q/ha, voire même 60 q/ha dans les cas extrêmes (Jorgensen et al. 2008). Ces pertes sont la conséquence de la disparition prématurée des parties vertes de la plante. La maladie sera donc d'autant plus nuisible que la perte de surface verte est précoce. Les dégâts les plus importants résultent de la sénescence des 2 feuilles les plus importantes : la F1 et la F2 qui contribuent le plus au remplissage des grains. Parmi les composantes du rendement, le PMG (poids de mille grains) est la composante la plus touchée.

Notons qu'il peut s'écouler 20 à 30 jours entre la contamination et l'apparition des symptômes, mais que les traitements les plus efficaces sont ceux appliqués à une date proche de la contamination. Bien positionné, le traitement est plus efficace et rend la protection globalement plus efficiente. Traiter trop tôt peut s'avérer inutile, et trop tard, désastreux. De cette date de première intervention, découle aussi le reste du programme et parfois le nombre d'applications total de la saison.

2. L'évolution du cadre législatif

2.1. Une prise de conscience des impacts

L'utilisation des pesticides en France a permis depuis 50 ans une augmentation considérable des rendements ainsi qu'une régularité accrue de la production. Cependant, la prise de conscience des impacts écologiques et sanitaires liés aux pesticides a amené la question de leur utilisation au cœur des débats sur l'agriculture et l'environnement. Cette évolution a déjà conduit à une réduction progressive de leur utilisation par certains agriculteurs et à un renouvellement des politiques publiques les concernant : au-delà d'un renforcement de l'évaluation des pesticides eux-mêmes, le besoin d'une évaluation et d'un encadrement de leur utilisation s'est peu à peu imposé. Cela s'est traduit en particulier au niveau communautaire avec le "paquet pesticides" et au niveau national avec le Plan Interministériel de Réduction des Risques liés aux Pesticides (PIRRP) et le Grenelle de l'environnement.

2.2. Un plan ambitieux

Ce Grenelle a marqué une nouvelle étape dans les décisions publiques relatives aux pesticides, en conduisant à l'élaboration d'un plan visant une réduction de 50% des usages des pesticides dans un délai de 10 ans, "si possible" : le plan Ecophyto 2018. Ce plan s'inscrit dans un cadre d'action communautaire : la directive du 21 octobre 2009, qui a pour objectif une utilisation des pesticides compatible avec le développement durable, impose en effet aux États membres d'adopter des plans d'action nationaux. Ces plans doivent viser la réduction des risques et des effets de l'utilisation des pesticides sur la santé humaine et l'environnement, ainsi que la mise en oeuvre des principes de la lutte intégrée contre les ennemis des cultures à compter du 1er janvier 2014, en vue de réduire la dépendance à l'égard de l'utilisation des pesticides. Par ailleurs, la réforme de l'homologation européenne sur les substances actives risque d'entraîner une réduction des substances disponibles. La réglementation incite à une utilisation plus parcimonieuse des produits.

Pour répondre à ces exigences réglementaires et environnementales les leviers d'actions sont d'une part de choisir des variétés résistantes et des conditions de cultures optimales, et d'autre part de bien détecter les maladies afin de ne traiter qu'en cas de besoin. Concernant la septoriose, la réactivité des producteurs est d'autant plus nécessaire que la maladie dépend fortement de la pluviométrie. C'est dans ce contexte qu'Arvalis a développé un outil d'aide à la décision, mis à disposition des acteurs du conseil agricole et des producteurs, permettant d'agir rapidement pour décider du premier traitement afin d'optimiser la lutte contre cette maladie et limiter au maximum les pertes de production.

3. L'outil d'aide à la décision d'Arvalis : *Septo-LIS®*

Comme nous l'avons vu en section 1.2., le positionnement des traitements fongicides contre la septoriose est un levier d'action contribuant à répondre au double enjeu environnemental et de gestion des résistances de la protection fongicide. Le positionnement des traitements permet effectivement d'améliorer la rentabilité d'un programme à dose égale, et peut même permettre d'éviter certains traitements.

Les traitements fongicides sont plus efficaces avant l'apparition des symptômes de septoriose. Cependant, il peut s'écouler 20 à 30 jours entre la contamination des plantes et l'apparition des symptômes. Dès lors, comment savoir si l'état sanitaire de la culture demande un traitement ? *Septo-LIS®* répond à ce besoin en donnant les prévisions du nombre de jours avant la date de déclenchement du premier traitement, basées sur la prévision des contaminations quotidiennes et les règles de décision mises au point par les spécialistes d'Arvalis. Le déploiement de modèles de prévision comme *Septo-LIS®* doit permettre d'ajuster au mieux le positionnement du premier traitement.

3.1. Un modèle mécaniste pour des processus complexes

Dans la classification (parfois controversée) des modèles mathématiques, *Septo-LIS®* est considéré comme un modèle mécaniste déterministe ; Par modèle mécaniste on entend un modèle qui considère explicitement l'objet modélisé comme un ensemble d'éléments ou de processus qui interagissent, et qui est basé au moins pour partie sur le comportement de ces éléments. Les modèles mécanistes se basent sur des lois physiques, chimiques et biologiques et évitent les approches empiriques de type fonctionnelles. Le modèle est déterministe (par opposition à stochastique) car ces paramètres été jusqu'à présent supposé fixes ; un seul résultat est obtenu avec le modèle.

3.2. Un conseil opérationnel à 3 critères

L'outil de calcul prend en compte 3 facteurs qui influencent fortement le développement de la maladie : date de semis, variété et conditions climatiques. Plus la date de semis est tardive, moins le risque de septoriose est élevé. Concernant la variété, le modèle intègre deux aspects : le rythme de développement de la culture selon la précocité et sa sensibilité aux maladies. A partir des données climatiques, il simule la sortie et le développement des feuilles successives du blé, la contamination sur chaque étage foliaire et l'incubation de chacune des contaminations.

3.3. Spatialisation de l'information

En début de campagne, l'utilisateur définit plusieurs combinaisons « variétés / date de semis », dites « situations agronomiques », selon les spécificités de son secteur. Pour chaque situation agronomique, il obtient des cartes actualisées au quotidien, lui offrant une vision spatialisée du développement de la maladie. Il peut également suivre l'évolution de la septoriose sur 10 jours et comparer les dates de traitements conseillées pour l'année en cours et les dates optimales des années précédentes.

4. Les modèles mécanistes en agronomie

Comme nous avons déjà pu le comprendre, les modèles mécanistes pour l'agronomie s'imposent de plus en plus comme des outils incontournables pour les chercheurs, ingénieurs et acteurs du conseil et développement agricole. Dans notre cas le modèle a été développé par un institut technique agricole pour l'usage des ingénieurs et/ou conseillers, dans d'autres cas ils sont développés dans le cadre de collaborations INRA-ITA. Ils peuvent être utilisés pour fournir des références, pour préparer des avertissements ou pour explorer des scénarios.

4.1. Connaître le niveau de fiabilité des prédictions

Quel que soit le domaine d'application, il est important de bien connaître le niveau de fiabilité des prédictions ou préconisations dérivées du modèle utilisé. D'une part, les concepteurs des modèles ont besoin d'avoir une estimation de la fiabilité des modèles pour en mesurer la qualité, déterminer l'intérêt ou la nécessité de leur amélioration et ainsi orienter leurs travaux. D'autre part, les utilisateurs des modèles ont besoin de connaître le niveau de précision des modèles afin de prendre en considération cette information dans l'analyse ou la prise de décision.

4.2. Les pratiques actuelles

Pour des modèles statistiques classiques, l'estimation de la fiabilité des résultats est une partie importante de l'analyse. En pratique on calcule un intervalle de confiance, qui peut ensuite être systématiquement affiché avec les résultats du modèle. En général, il est difficile de transposer cette approche directement aux modèles mécanistes. L'approche générale pour ces modèles est alors de comparer, dans un certain nombre de cas, prédictions du modèle avec observations et de déduire un niveau moyen d'erreur. C'est pourtant loin d'être satisfaisant. D'abord, le modèle est souvent ajusté aux données. Dans ce cas, le niveau d'erreur constaté représente l'erreur d'ajustement, et sous estime l'erreur de prédiction. Ensuite, il y a rarement une réflexion avancée sur la représentativité des données, bien que le niveau d'erreur puisse être très différent suivant les situations où le modèle est appliqué. De plus, ce n'est pas seulement l'erreur totale qui est intéressante mais aussi des contributions individuelles à l'erreur. Enfin, le niveau d'erreur est très rarement affiché avec le modèle. Une fois testé et validé, le modèle est adopté sans plus de mention du niveau d'erreur. Tout cela entraîne un manque de transparence sur la fiabilité de ces modèles qui nuit à leur utilisation pratique.

Il existe, dans la littérature statistique ou la littérature de la modélisation, un ensemble important de travaux sur l'évaluation des modèles mécanistes. Des méthodes permettant d'évaluer les modèles mécanistes ont été proposées, mais le passage des connaissances académiques à une démarche opérationnelle, applicable à des cas réels et adaptée aux différentes situations, reste à réaliser. Parmi les méthodes permettant de quantifier l'incertitude d'un modèle, nous avons sélectionné les méthodes Bayesiennes que nous allons maintenant décrire.

5. Les méthodes Bayesiennes

5.1. Modalités générales

Les méthodes Bayesiennes tiennent leur nom du pasteur et mathématicien anglais Thomas Bayes (1702-1761) qui fut le premier à formuler le théorème des probabilités conditionnelles, aussi connu sous le nom de Théorème de Bayes.

Ce théorème constitue un des fondements de la théorie des probabilités conditionnelles. Dans son unique article, Bayes cherchait à déterminer ce que l'on appelle actuellement la distribution a posteriori de la probabilité p d'une loi binomiale. Ce théorème est utilisé dans l'inférence statistique pour estimer ou actualiser les estimations d'une probabilité ou d'un paramètre quelconque, à partir des observations et des lois de probabilité de ces observations. Il existe une version discrète et une version continue du théorème.

La version continue du théorème se déduit simplement de la densité jointe des observations et des paramètres, produit de la vraisemblance par la densité a priori sur les paramètres, par application de la définition des lois et des densités conditionnelles. Cette application montre que la distribution a posteriori s'obtient en multipliant la distribution a priori, par la vraisemblance, et en effectuant une normalisation (du fait qu'il s'agit d'une densité de probabilité). En calcul bayésien, on prend donc l'habitude de travailler avec des signes de proportionnalité plutôt que des égalités pour diminuer la complexité des expressions puisque les constantes manquantes se retrouvent par intégration (en principe). Les techniques de simulation de type Monte Carlo et MCMC, que nous emploierons ici et que nous décrirons dans la partie matériels et méthodes, n'utilisent d'ailleurs pas ces constantes de normalisation.

5.2. Intérêt pour notre étude

L'originalité de ces méthodes par rapport aux méthodes classiques, dites fréquentistes, repose sur deux grands principes. Premièrement, les méthodes Bayesiennes permettent d'intégrer de l'information a priori sur les paramètres d'un modèle (p.e issue de l'étude de processus individuels) et de combiner cette information avec des données expérimentales sur le système complet. En 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 paramètres, qui sont ici considérés comme variables aléatoire, plutôt que d'une valeur. En 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.

De ces deux grands principes découle une meilleure prise en compte de la variabilité des paramètres. La connaissance de cette variabilité doit pouvoir permettre de quantifier plus précisément l'incertitude de notre modèle.

6. Objectifs à long terme

Les objectifs de notre étude sont vastes et peuvent se diviser en deux catégories ; des objectifs fondamentaux et des objectifs appliqués.

D'un point de vue appliqué, l'objectif est d'améliorer le modèle *Septo-LIS®* pour optimiser le conseil dans les dates de traitement. Cet objectif s'inscrit totalement dans le cadre du plan Ecophyto 2018 puisque l'optimisation à long terme des dates de traitement devrait conduire à une réduction de l'utilisation des fongicides sur le territoire français.

D'un point de vue fondamental, nous cherchons à comprendre comment appliquer des méthodes Bayesiennes à des systèmes d'équations dynamiques non linéaires. Au-delà des objectifs liés spécifiquement à *Septo-LIS®*, ce projet représente une occasion pour tenter de définir une démarche d'analyse applicable à d'autres modèles semblables. On essaiera notamment de proposer des pistes pour comprendre comment et à quel point le nombre de données et la connaissance a priori sur la variabilité des paramètres influencent les résultats de l'estimation.

7. Objectifs opérationnels

A court terme et d'un point de vue opérationnel, mon objectif est d'appliquer et d'implémenter des méthodes Bayesiennes pour l'estimation des paramètres du modèle afin de mettre au point et d'implémenter une procédure d'évaluation de l'incertitude des prédictions de *Septo-LIS®*. En terme statistique cela signifie que nous allons chercher la distribution des paramètres et de la variance résiduelle de notre modèle pour en déduire une distribution des ses prédictions.

Cet objectif d'obtenir une distribution de prédictions plutôt qu'une valeur prédite doit permettre une meilleur prise en compte de l'erreur du modèle par les spécialistes qui établissent les règles de décisions. Pour acteurs du conseil agricole, notre méthode d'évaluation doit pouvoir faciliter la prise de décisions car elle permettra une visualisation suffisamment intuitive de la date de traitement considérée par le modèle comme optimale et de la dispersion des dates possibles autour de celle-ci. Cette approche est totalement novatrice pour ce type de modèle. On rappelle qu'à l'heure actuelle, les sorties de ce type de modèles sont représentées par une valeur dont on ne sait pas, le plus souvent, ce qu'elle vaut d'un point de vue statistique !

II. MATERIEL ET METHODES

1. Le modèle Septo-LIS®

Le modèle *Septo-LIS®* est dérivé du modèle *« Septoria tritici blotch »* (STB) décrit dans Audsley et al. (2005). Il simule au pas de temps journalier le développement de la maladie sur les six feuilles les plus jeunes à partir de la date d'émergence de F6 (F1 étant la feuille la plus jeune).

Le modèle initial prend en compte une quantité fixe d'inoculum considérée comme la source d'inoculum du sol, qui peut être transférée aux feuilles supérieures, à mesure qu'elles se développent, et les infecter. La proportion d'inoculum transférée dépend de l'abondance des précipitations et de la distance entre les feuilles (fonction linéaire du temps thermique et de la position de la feuille). La proportion d'inoculum transférée causant des infections est aussi fonctions des précipitations. Les infections entrent ensuite dans une phase d'incubation pendant une période qui dépend de la température journalière moyenne. Après cette phase, les infections vont évoluer en lésions dont l'augmentation de taille est modélisée par une croissance logistique en fonction du temps thermique. Enfin les lésions produisent des spores et agissent donc comme de nouvelles sources d'inoculum.

Un certain nombre d'élément ont été modifiés dans *Septo-LIS®* par rapport au modèle original. Le modèle de croissance du blé avec lequel le modèle de septoriose est couplé a été simplifié et adapté au climat français. Ici, seules l'émergence et l'expansion des feuilles sont simulées. Un taux d'émergence des feuilles est calculé d'après le modèle phyllochron développé pour la France par Gate (1995) à partir de celui proposé par Weir (1984). L'expansion des feuilles est considérée comme une fonction linéaire du temps thermique. La fonction d'incubation a été remplacée par celle proposé par Lovell et al. (2004). Par ailleurs, un module de développement hivernal de l'inoculum a été développé. En effet, le modèle d'Audsley et al. (2005) démarre à partir de la F6, et c'est une observation en cours de montaison qui permet d'ajuster le niveau d'inoculum de début de montaison. Or, de nombreux auteurs ont mis en avant l'importance des conditions hivernales, et en particulier les températures basses, dans le développement du potentiel épidémique d'une parcelle (Coakley *et al.*, 1985; Gladders *et al.*, 2001; Pietravalle *et al.*, 2003; te Beest *et al.*, 2009). Une fonction polynomiale de la température et de la pluie journalière à partir du stade début tallage a donc été développée et couplée au modèle d'Audsley *et al.* (2005).

1.1. Equations et formalismes

Nous décrivons ici les principales équations du modèle sans rentrer dans tous les détails qui ne contribueraient pas à la compréhension de notre étude.

- Croissance des lésions : Fonction logistique du temps thermique

$$\frac{dy}{dT} = ky(1-y)\ avec\ y(0) = a$$

- Fraction de feuille l infectée au jour t

$$Y(t,l) = \sum_{i=1}^{t} I(i,l)y[T(t) - T(i)]$$

$I(i,l) =$ Nombre d'infections jour i sur la feuille l

$y[T(t) - T(i)] =$ Taille d'une infection au jour i

- Nombre d'infections

$$I(t,l) = F(t,l)f_{pluie}f_{niveau\ maladie}f_{fraction\ visible}f_{fraction\ saine}$$

  - Spores qui atteignent la feuille l

$$F(t,l) = \textit{Spores du sol+ Spores des autres feuilles+ Auto-infection}$$

$$F(t,l) = U_g e^{-kh_g(t)} + \sum_{j=l+1}^{6} U(t,j)\, e^{-kh_j(t)} + \rho U(t,l)$$

$Avec\ k(r_t) = ae^{-br_t}$

$r_t =$ Pluviométrie au jour t
$h_j(t) =$ Distance entre la feuille l et la feuille j au jour t
$h_g(t) =$ Distance entre la feuille l et le sol au jour t
$k =$ Efficacité du transfert d'inoculum(fonction de la pluviométrie)
$a, b =$ Paramètres

- Production de spores (dépend de la surface infectée)

$$U(t,l) = 1 - e^{-a[Y(t,l)+I_s\delta(l-6)]}$$

$$Avec \quad \delta(l-6) = \begin{cases} 1 & l = 6 \\ 0 & sinon \end{cases}$$

- Pluie

$$f_{pluie} = \begin{cases} a + br_t & r_t < \Theta \\ 0 & r_t \geq \Theta \end{cases}$$

Le modèle comporte 17 paramètres au total. Sur ces 17, 14 sont intégrés à des équations qui représentent un processus biophysique réel et sont donc mesurables en conditions contrôlées. Les trois paramètres restant n'ont pas réellement de sens biologique et représentent plutôt des paramètres d'ajustement empiriques du système global.

2. Le modèle statistique

Contrairement au modèle *Septo-LIS®*, notre modèle statistique s'écrit très simplement de la façon suivante :
Soit Y: la variable expliquée, X: l'ensemble des variables explicatives, θ: les paramètres du modèle et $\boldsymbol{\varepsilon_k}$: l'erreur du modèle

$$Y_k = f(X_k, \theta) + \varepsilon_k$$

Avec l'hypothèse de normalité des résidus : $\varepsilon_k \sim N(0, \sigma_\varepsilon^2)$

Cette équation formalise simplement le fait que le modèle $f(X_k, \theta)$ prédit notre variable expliquée Y_k à une composante aléatoire près ε_k. L'hypothèse de normalité des résidus signifie que nous supposons cette composante aléatoire comme étant distribuée selon une loi normale centrée sur 0 avec une variance σ_ε^2 qui est la variance résiduelle du modèle.

3. Les données

Toutes les données dont nous disposons pour l'estimation des paramètres du modèle proviennent d'au total 17 départements localisés dans la moitié nord de la France (Calvados, Morbihan, Côte d'Armor, Loire-Atlantique, Mayenne, Charente Maritime, Indre, Cher, Côte d'Or, Loir et Cher, Seine et Marne, Yonne, Seine Maritime, Eure, Somme, Essonne et Marne). Pour la plupart des sites d'observation, plusieurs années sont disponibles. Chaque combinaison d'une année et d'un site est décrite comme un « site-année ». Toutes les données de sévérité de septoriose sont relevées dans les parcelles (non traitées avec des fongicides) du réseau expérimental d'Arvalis.

Deux jeux de données différents ont été utilisés ; un pour réaliser l'estimation des paramètres (appelé jeu d'entraînement) et un second pour évaluer la qualité de prédiction du modèle avec les paramètres estimés (jeu de validation). L'utilisation de deux jeu de données distincts permet d'analyser plus rigoureusement les composantes de la qualité du modèle ; d'une part la qualité d'ajustement c'est-à-dire l'aptitude du modèle à reproduire des données utilisées pour l'estimation et d'autre part la qualité de prédiction c'est-à-dire l'aptitude du modèle à faire des prédictions proche de la réalité dans de nouvelles situations.

3.1. Jeu d'entraînement

Les données constituant le jeu d'entraînement sont composées de 157 sites-année, couvrant 38 sites et 34 variétés sur une période s'étalant de 1998 à 2007. Au total, il y a 2001 données pour l'ajustement du modèle.

3.2. Jeu de validation

Les données du jeu de validation (pour la prédiction du modèle) ont été récoltées sur les essais de l'année 2008. Elles couvrent 18 sites et 11 variétés (dont 5 qui n'étaient pas représentées dans le jeu d'entraînement). Au total, il y a 270 données pour la validation du modèle.

4. Application de méthodes Bayesiennes à notre cas

L'équation de Bayes est une équation de base en statistique (fréquentiste ou bayesienne)

$$P(U|V) = \frac{P(V|U) \times \pi(U)}{p(V)}$$

où U et V sont deux variables aléatoires. En revanche, il n'y a que dans l'approche bayesienne que l'on applique cette formule avec $U=\theta$, le vecteur des paramètres, et $V=Y$, le vecteur des mesures. La forme de l'équation qui nous intéresse est alors :

$$P(\theta|Y) = \frac{P(Y|\theta) \times \pi(\theta)}{p(Y)}$$

$P(\theta|Y)$ est la distribution a posteriori. La notation est un raccourci pour $P(\theta = \theta_0|Y = Y_0)$, la probabilité que θ ait une valeur particulière θ_0 sachant que le vecteur des dégâts observés a une valeur particulière Y_0. On voit 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 aussi différent que l'approche fréquentiste. Dans l'approche fréquentiste, 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 observé de Y. Dans une approche bayesienne on parle d'intervalle crédible, qui est l'équivalent d'un intervalle de confiance mais qui est basé sur l'incertitude de θ et non pas sur l'incertitude des mesures.

Le dénominateur $p(Y)$ dans l'équation est la probabilité d'observer la valeur Y. Il peut s'écrire $p(Y) = \int P(Y|\theta) \times \pi(\theta)\, d\theta$. C'est-à-dire, c'est la probabilité d'observer Y pour chaque valeur de θ, pondérée par la probabilité de cette valeur de θ et intégrée sur θ. 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 ;

$$P(\theta|Y) \propto P(Y|\theta) \times \pi(\theta)$$

On connaît donc la distribution a posteriori à une constante près, égale à *1/p(Y)*. Par contre, en général on ne connaît pas cette constante, parce que l'intégration $\int P(Y|\theta) \times \pi(\theta)\, d\theta$ est très difficile à réaliser surtout si θ est multidimensionnel. Etant donné que l'on connaît la distribution a posteriori à une constante près, on peut calculer le maximum de cette distribution. Si la distribution a priori est une distribution non-informative ($\pi(\theta)$ proportionnelle à une constante), le maximum correspond au maximum de la vraisemblance. C'est-à-dire, en absence d'information a priori, la valeur de θ la plus probable dans une approche bayesienne est la valeur de maximum de vraisemblance.

4.1. Vraisemblance $P(Y|\theta)$

La vraisemblance est la fonction qui relie les données aux paramètres du modèle. Elle représente la probabilité que les données aient été observées sachant la valeur des paramètres. Comme dans l'approche fréquentiste, la vraisemblance est fondamentale dans l'approche bayesienne.

Le terme $P(Y|\theta)$ est donc la probabilité d'observer la valeur Y, quand le paramètre prend la valeur θ. C'est exactement la vraisemblance classique. Nous faisons ici les mêmes hypothèses que dans le cas fréquentiste, c'est-à-dire $Y_k = f(X_k, \theta) + \varepsilon_k$ avec $\varepsilon_k \sim N(0, \sigma_\varepsilon^2)$ et indépendance des ε_k. On a alors :

$$P(Y|\theta) = \prod P(Y_k|\theta)$$

$$P(Y|\theta) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}^{N/2}} \times e^{-\sum_{k=1}^{N}[Y_k - f(X_k,\theta)]^2/2\sigma_\varepsilon^2}$$

Dans le cas où on considère la variance résiduelle comme connue, on remplacera σ_ε^2 par sa valeur au maximum de vraisemblance σ_{MV}^2. Sinon on estimera σ_ε^2

4.2. Loi a priori π

4.2.1. Pour les paramètres $\pi(\theta)$

La loi, ou distribution a priori, représente la connaissance du modélisateur et des experts sur les paramètres d'un modèle avant qu'ils n'aient accès aux données. Cette connaissance peut avoir différentes origines ; l'expertise, la littérature, les expériences passées… La notion de distribution a priori n'existe pas dans approche fréquentiste, étant donné que θ n'est pas,dans cette approche une variable aléatoire. Dans l'approche bayesienne par contre la distribution a priori est obligatoire. Dans notre cas, nous avons affaire à deux types de paramètres ; certains paramètres sont considérés comme bien connus car ils représentent chacun un processus biophysique réel mesurable en conditions contrôlées (exemple : Vitesse d'expansion des lésions, Précipitations seuil pour le développement du champignon…). Les autres paramètres n'ont pas réellement de sens biologique, ils représentent plutôt des paramètres empiriques de calage du système complet et sont considérés comme incertains. Cette différence de sens entre les paramètres nous oblige à définir différents types de loi a priori.

Pour les paramètres jugés incertains (noté θ_A) on attribue des lois uniformes dans un intervalle traduisant le fait qu'aucune valeur n'est plus probable que les autres dans cet intervalle.

Pour les paramètres considérés comme bien connus (noté θ_B) on dispose de valeurs nominales fixes mais pas d'incertitude autour de ces valeurs. On construit donc pour chacun de ces paramètres une loi a priori normale centrée sur la valeur dont on dispose et on définit deux niveaux d'incertitude: 5% et 25% d'écart-type autour de la valeur nominale.

$$\theta_A \sim U\left(\theta_{A_{min\,j}}, \theta_{A_{max\,j}}\right) \qquad \pi(\theta_A) = \prod_{j=1}^{3} \frac{1}{\left(\theta_{A_{max\,j}} - \theta_{A_{min\,j}}\right)}$$

$$\theta_B \sim N\left(\hat{\mu}_{\theta_{B_i}} = \theta_{B\,i}, \hat{\sigma}^2_{\theta_{B_i}}\right) \qquad \pi(\theta_B) = \prod_{i=1}^{14} \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\theta_{B_i}}}} \times e^{-\frac{\left(\theta_{B_i} - \hat{\mu}_{\theta_{B_i}}\right)^2}{2\hat{\sigma}^2_{\theta_{B_i}}}}$$

$$\pi(\theta) = \pi(\theta_A) \times \pi(\theta_B) = \prod_{j=1}^{3} \frac{1}{\left(\theta_{A_{max\,j}} - \theta_{A_{min\,j}}\right)} \times \prod_{i=1}^{14} \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\theta_{B_i}}}} \times e^{-\frac{\left(\theta_{B_i} - \hat{\mu}_{\theta_{B_i}}\right)^2}{2\hat{\sigma}^2_{\theta_{B_i}}}}$$

4.2.2. Pour la variance résiduelle $\pi(\sigma_\varepsilon^2)$

Lorsque la variance résiduelle est inconnue, on cherche la distribution suivante

$$P(\theta, \sigma_\varepsilon^2 | Y) = \frac{P(Y|\theta, \sigma_\varepsilon^2) \times \pi(\theta, \sigma_\varepsilon^2)}{p(Y)}$$

Il nous faut donc définir un élément nouveau ; la distribution a priori $\pi(\theta, \sigma_\varepsilon^2)$. On supposera que les distributions a priori pour θ et σ_ε^2 sont indépendantes, d'où

$$\pi(\theta, \sigma_\varepsilon^2) = \pi(\theta)\pi(\sigma_\varepsilon^2)$$

et que la distribution a priori pour σ_ε^2 est

$$\pi(\sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2$$

Cette distribution est la loi de Jeffreys. Elle est souvent utilisée pour représenter une distribution non-informative. (Gelman et al., 2004):

5. Stratégie globale

Pour avoir la distribution complète des paramètres et de la variance résiduelle, il y a trois types d'approche. Dans certains cas particuliers, on peut obtenir une expression analytique pour la distribution a posteriori, mais ce n'est pas notre cas ici. On ne peut

donc qu'obtenir une approximation. « Importance sampling » est une approche pour obtenir cette approximation. L'approximation à la distribution a posteriori dans ce cas prend la forme d'un ensemble de valeurs θ, et chaque valeur est associée à une pondération. L'autre famille d'approches est celle des méthodes MCMC (Monte Carlo Markov Chain). Dans ce cas le résultat est une chaîne de valeurs de θ, et l'ensemble des valeurs de la chaîne représente l'approximation à la distribution a posteriori.

De ce fait nous avons recours à une phase exploratoire pour tester l'efficacité des différentes approches existantes sur notre modèle. Après sélection de l'algorithme jugé le plus performant, on détermine une stratégie spécifique pour ce dernier. L'idée générale est de pouvoir disposer de distributions pour tous les paramètres du modèle et d'utiliser ces distributions en simulation pour obtenir une distribution des sorties.

En effet, une fois que l'on a la distribution a posteriori des paramètres ou une approximation à cette distribution, toute l'information utile est disponible. Par exemple, si l'on dispose d'une chaîne de paramètres et que l'on s'intéresse à la distribution des prédictions. Il suffit simplement de réaliser les prédictions avec toutes les valeurs de θ de la chaîne. Cela donne une approximation à la distribution des prédictions. Si on veut un intervalle qui a une probabilité d'inclure la vraie valeur égale à 0.9, on prendra l'intervalle le plus court qui couvre 90% de ces prédictions.

6. Plan d'expérience

Notre stratégie d'estimation a été formalisée en plan d'expérience factoriel à deux facteurs ; Les deux niveaux d'incertitudes a priori (5 et 25% de la valeur nominale) sur les 14 paramètres représentant un processus biophysique constitue le premier facteur (2 modalités). Le deuxième facteur est le nombre de site-année (quantité de données) utilisés pour l'estimation. Pour ce facteur on a défini 3 modalités : 1, 8 et 157 sites-année. Pour les deux facteurs, nous avons ajouté une modalité correspondant à un témoin (respectivement 0% d'incertitude et 0 données). La Figure 2 montre une représentation schématique de notre plan d'expérience

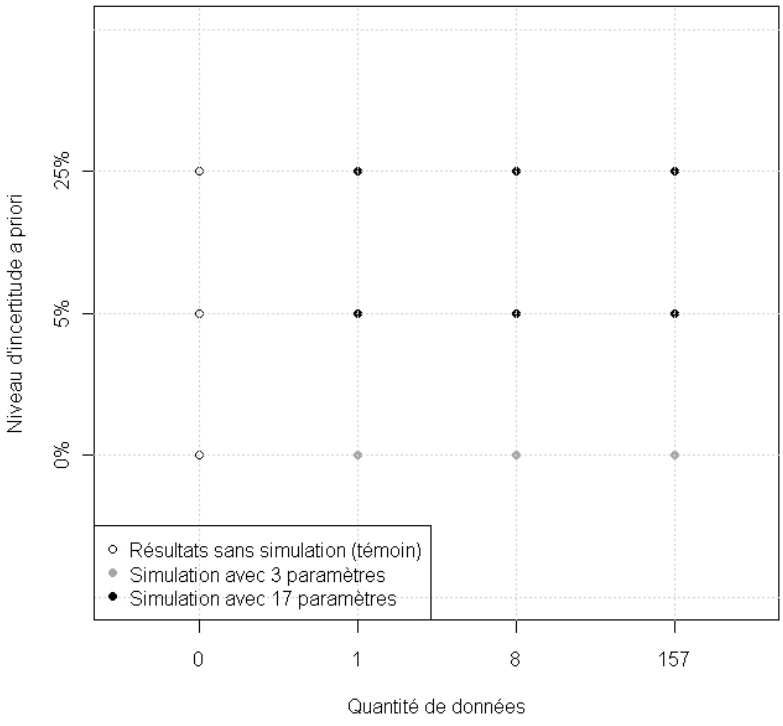


Figure 2: Représentation schématique de notre plan d'expérience

7. Les méthodes de Monte-Carlo

Depuis environ 15 ans, on assiste à une véritable explosion des travaux de recherche en statistique Bayesienne. Comme on l'a vu précédemment, en statistique Bayesienne on choisit une distribution a priori dans l'espace des paramètres et on détermine ensuite, à l'aide des données, la distribution a posteriori. Une fois que l'on a cette distribution a posteriori, les calculs inférentiels sont relativement simples. La partie la plus délicate réside donc dans le choix de la distribution a priori et la « recherche » de la distribution a posteriori. Dans le passé, le plus gros problème avec les méthodes bayesiennes était l'absence de méthodes robustes permettant d'approximer la distribution a posteriori. Ces dernières années, ce problème a été en partie résolu en utilisant les méthodes de Monte-Carlo pour simuler cette distribution.

Les méthodes de Monte Carlo sont des méthodes numériques, qui utilisent des tirages aléatoires pour réaliser le calcul d'une quantité déterministe. Ces méthodes permettent de générer des échantillons aléatoires indépendants et identiquement distribués pour simuler des processus aléatoires. Elles ont comme intérêt d'être « insensible à la dimension » des problèmes étudiés, contrairement à des méthodes classiques d'analyse numérique qui ne sont performantes qu'en « petite dimension ».

Du point de vue des applications, ces méthodes sont aujourd'hui indispensables dans des domaines aussi variés et différents que la finance, la mise au point de nouveaux microcomposants électroniques, la sismologie, les télécommunications, en ingénierie ou en physique, mais aussi en biologie, en sciences sociales, etc. Par exemple, en chimie, en physique et en biologie, de nombreux problèmes exigent l'analyse des propriétés dynamiques d'un nombre tellement grand d'objets (particules atomiques, atomes, molécules, macromolécules, organes, organismes) et de leurs éventuelles interactions, que ceci ne peut se faire que par des techniques de type Monte Carlo.

7.1. Importance sampling

7.1.1. Modalités générales

L'idée générale ici est de définir une distribution « instrumentale » permettant de générer aléatoirement des échantillons (valeurs de paramètres) indépendants et identiquement distribués (iid). Ces échantillons sont ensuite pondérés par une valeur proportionnelle au ratio des probabilités a posteriori et instrumentale notée ω. L'échantillon pondéré est une approximation à la distribution a posteriori.

En théorie la distribution instrumentale peut être quelconque à la seule condition que la densité soit différente de zéro partout où la distribution a posteriori est différente de zéro. On note cependant que le choix de la distribution instrumentale n'est pas totalement anodin, il va considérablement influencer l'efficacité de l'algorithme en termes de temps de calcul et d'importance des valeurs échantillonnées dans la définition de la distribution a posteriori. En pratique, on a intérêt à utiliser une bonne approximation à la distribution a posteriori si possible. Si on n'a pas d'approximation, on peut utiliser la loi a priori comme distribution instrumentale. Par construction, la densité a priori satisfait la condition sur la distribution instrumentale ; si la densité a priori est 0 alors forcément la densité a posteriori égale 0.

7.1.2. Monte-Carlo élémentaire

On génère n nombres aléatoires, θ, compris entre a et b et distribués suivant une densité de probabilité uniforme. Cette densité de probabilité est constante dans l'intervalle considéré, c'est-à-dire que toutes les valeurs comprises entre a et b sont équiprobables. L'échantillonnage utilisé est homogène. C'est-à-dire que l'on visite également toutes les régions comprises entre a et b. On qualifie cet échantillonnage de MC *élémentaire* (ou MC simple). Si la distribution a posteriori est fortement hétérogène, l'estimation par la méthode de MC élémentaire est particulièrement inefficace. En effet, on échantillonne trop les régions où la probabilité a posteriori est faible et pas assez celles où elle est élevée, ce qui conduit à une convergence excessivement lente du calcul de l'intégrale. Cette méthode est utilisée uniquement pour l'estimation des paramètres ayant des distributions a priori uniforme car ici le support de la distribution instrumentale est fini et ce support doit pouvoir couvrir celui de la distribution a priori (ça ne serait pas le cas pour une loi a priori normale dont le support est infini).

7.1.2.1. Algorithme et formalismes

I. *Définir une distribution instrumentale* $g(\theta)$. *Ici,* $g(\theta) = U(\theta_{min}, \theta_{max})$

II. *Pour n de 1 à p*

{Générer $\theta^{(n)}$ *en tirant une valeur de* $g(\theta)$

- *Si* $\pi(\theta^{(n)}) = 0$

  Générer $\theta^{(n)}$ *en tirant une valeur de* $g(\theta)$

- *Sinon*
  Calculer le poids

$$\omega(\theta^{(n)}) = \frac{P(Y|\theta^{(n)}) \times \pi(\theta_A^{(n)})}{g(\theta^{(n)})}$$

Or $g(\theta) = \pi(\theta_A)$ *d'où*

$$\omega(\theta^{(n)}) = P(Y|\theta^{(n)})$$

$$\omega(\theta^{(n)}) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}^{N/2}} \times e^{-\sum_{k=1}^{N}\left[Y_k - f(X_k, \theta^{(n)})\right]^2 / 2\sigma_\varepsilon^2}$$

}

III. *Normaliser les poids*

$$\omega_{\%}(\theta^{(n)}) = \frac{\omega(\theta^{(n)})}{\sum_{n=1}^{p} \omega(\theta^{(n)})}$$

L'échantillon $\left(\theta^{(n)}, \omega_{\%}(\theta^{(n)})\right)$ *de valeurs et pondérations associées est une approximation à la distribution a posteriori*

7.1.3. Monte-Carlo biaisé

Pour améliorer la convergence, on utilise une méthode légèrement différente. L'idée est de privilégier les régions où la fonction possède des valeurs élevées par rapport à celles où elle est proche de zéro. Plutôt que de générer des nombres aléatoires uniformément répartis, on tire ces nombres dans une distribution non uniforme qui échantillonne mieux les régions où l'intégrale est importante. Le biais introduit par le choix de cette distribution est ensuite corrigé en divisant de la probabilité a posteriori par la probabilité de l'échantillon dans la distribution instrumentale. Il en découle que la distribution obtenue ne dépend pas du choix de la distribution instrumentale. Cette méthode représente une amélioration par rapport à la méthode précédente, c'est une méthode de MC *biaisé* par opposition au MC élémentaire. On biaise par rapport à la distribution uniforme, mais le résultat n'est pas biaisé.

7.1.3.1. Algorithme et formalismes

I. *Définir une distribution instrumentale* $g(\theta)$. *Ici,* $g(\theta) = g(\theta_A) \times g(\theta_B)$

$$g(\theta_B) = \pi(\theta_B) = \prod_{i=1}^{14} \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\theta_{B_i}}}} \times e^{-\frac{\left(\theta_{B_i} - \hat{\mu}_{\theta_{B_i}}\right)^2}{2\hat{\sigma}^2_{\theta_{B_i}}}}$$

$$g(\theta_A) = \prod_{i=1}^{3} \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \times e^{-\frac{1}{2}(\theta_{A_i} - \hat{\theta}_{MV_A})^t \Sigma^{-1} (\theta_{A_i} - \hat{\theta}_{MV_A})}$$

II. *Pour n de 1 à p*

{Générer θ^n *en tirant une valeur de* $g(\theta)$

- *Si* $\pi\left(\theta^{(n)}\right) = 0$

  Générer $\theta^{(n)}$ *en tirant une valeur de* $g(\theta)$
- *Sinon*

  Calculer le poids

$$\omega\left(\theta^{(n)}\right) = \frac{P(Y|\theta^{(n)}) \times \pi(\theta_A^{(n)}) \times \pi(\theta_B^{(n)})}{g\left(\theta_A^{(n)}\right) \times g(\theta_B^{(n)})}$$

Or $g(\theta_B) = \pi(\theta_B)$ *d'où*

$$\omega\left(\theta^{(n)}\right) = \frac{P(Y|\theta^{(n)}) \times \pi(\theta_A^{(n)})}{g\left(\theta_A^{(n)}\right)}$$

$$\omega\left(\theta^{(n)}\right) = \frac{\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}^{N/2}} \times e^{-\sum_{k=1}^{N}\left[Y_k - f\left(X_k, \theta^{(n)}\right)\right]^2 / 2\sigma_\varepsilon^2} \times \pi(\theta_A^{(n)})}{\prod_{i=1}^{3} \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \times e^{-\frac{1}{2}\left(\theta_A^{(n)}{}_i - \hat{\theta}_{MV_A}\right)^t \Sigma^{-1} \left(\theta_A^{(n)}{}_i - \hat{\theta}_{MV_A}\right)}}$$

}

III. Normaliser les poids

$$\omega_{\%}\left(\theta^{(n)}\right) = \frac{\omega\left(\theta^{(n)}\right)}{\sum_{n=1}^{p} \omega(\theta^{(n)})}$$

L'échantillon $\left(\theta^{(n)}, \omega_{\%}\left(\theta^{(n)}\right)\right)$ *de valeurs et pondérations associées est une approximation à la distribution a posteriori*

7.2. Intérêts et limites

Le principal intérêt de ce type de méthodes repose sur le caractère indépendant des échantillons générés. Cette indépendance va considérablement simplifier l'établissement de statistiques a posteriori car il n'y a pas de corrélations entre ces échantillons. La limite de ce type d'algorithmes est que, dans la pratique, il nécessite d'avoir quelques informations sur la distribution a posteriori (au moins forme et mode) pour être efficace.

Cette limite explique notre choix de distribution instrumentale dans le cas biaisé ; Pour obtenir de plus d'informations sur les paramètres nous avons réalisé une estimation par maximum de vraisemblance (méthode *Nelder-Mead*) sur les 3 paramètres uniforme a priori. Ainsi nous disposons d'une information sur les corrélations, les variances et les modes de ces paramètres. Nous avons donc utilisée cette information pour définir la distribution instrumentale. Cette tentative n'a pas amélioré significativement l'efficacité de l'algorithme et nous a amené à nous tourner vers d'autres types approches plus prometteuses.

8. Monte-Carlo par Chaînes de Markov (MCMC)

8.1. Modalités générales

Comme décrit ci-dessus, les méthodes de Monte Carlo permettent de générer des échantillons de la 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 iid 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).

Une chaîne de Markov est une suite de variables aléatoires qui peuvent être considérées comme évoluant au cours du temps, avec une probabilité de transition dépendant uniquement de l'ensemble particulier dans lequel la chaîne se trouve à un moment donné. Il est donc naturel et, en fait, mathématiquement plus juste de définir une telle chaîne par son noyau de transition qui est donc la fonction déterminant le passage d'un ensemble à un autre.

8.2. Loi de proposition et taux de rejet

Pour construire une chaîne de Markov, les méthodes MCMC requièrent la définition d'un noyau de transition ou loi de proposition. C'est cette loi qui détermine le mouvement de la chaîne à partir d'un endroit donné. En théorie, la loi de proposition peut être quelconque à conditions de permettre l'exploration de l'espace des paramètres et d'être « facilement simulable ». En pratique, les lois de propositions les plus rencontrées, et que nous utiliseront ici, sont des lois normales centrées sur la valeur courante de la chaîne et avec une matrice de variance-covariance à déterminer en fonction des connaissances a priori sur les paramètres ainsi que sur leurs relations.

La matrice de variance-covariance de la loi de proposition joue ici un rôle crucial ; c'est elle qui détermine l'amplitude des pas de la chaîne dans l'espace des paramètres et donc la capacité de la chaîne à parcourir cette espace. C'est aussi cette matrice qui va déterminer une des composante de l'efficacité de l'algorithme ; la fraction des cas où on rejette une proposition (logiquement appelée taux de rejet).

Ce taux de rejet doit être optimal pour obtenir une bonne approximation de la distribution a posteriori. Si ce taux est trop grand la chaîne reste bloquée, pour plusieurs itérations, à une même valeur et le nombre d'itérations nécessaires pour avoir une bonne approximation à la distribution a posteriori serait trop grand pour être raisonnablement envisageable. Si ce taux est trop faible, c'est souvent parce que la chaîne se déplace très peu à chaque pas, et encore une fois il faudrait un trop grand nombre d'itérations pour avoir une bonne approximation à la distribution a posteriori. Il a été suggéré qu'un bon taux de d'acceptation (=1-taux de rejet) est d'environ ¼ pour plusieurs dimensions et de ½ pour dimensions 1 ou 2 (Robert and Casella, 2004).

8.3. Convergence

Une fois que l'on a la chaîne de valeurs, certains problèmes subsistent. D'abord, le début de la chaîne est fortement influencé par le point de départ. On élimine en général donc cette première partie. Combien de valeurs faut-il éliminer ? Une possibilité est alors de regarder les valeurs produites et de décider visuellement où la chaîne semble avoir atteint une distribution d'équilibre.

Une deuxième question concerne la longueur de la chaîne. Quelle longueur faut-il pour avoir une bonne approximation à la distribution a posteriori ? Encore une fois, un examen visuel des valeurs de la chaîne peut indiquer si la chaîne semble avoir convergé vers une distribution stable. Une autre possibilité, plus rigoureuse, et de générer au moins deux chaînes, avec des points de départ différents, puis (après avoir jeté le début de chaque chaîne) de comparer la variance inter- et intra-chaîne. Les deux variances devraient être comparables si les chaînes ont convergé vers la même distribution stable. C'est l'approche développé par Gelman et Rubin (1992) pour diagnostiquer la convergence d'une chaîne de Markov.

8.4. Metropolis-Hastings

8.4.1. Modalités générales

On considère 2 cas, celui avec variance résiduelle connue ($\sigma_\varepsilon^2 = \sigma_{MV}^2$) et le cas où la variance est inconnue. Dans le premier cas, l'algorithme construit une chaîne de valeurs $\theta_1,\ \theta_2, \dots, \theta_N$ et cette chaîne, après avoir éliminé les premières valeurs qui sont influencées par le point de départ, est une approximation à la distribution a posteriori $P(\theta|Y)$.

8.4.2. Algorithme et formalismes

I. Définir une distribution de proposition $g\left(\theta^{*(n+1)}, \theta^{(n)}\right)$. *Ici on travaille avec* $g\left(\theta^{*(n+1)}, \theta^{(n)}\right) = N\left(\theta^{(n)}, \Sigma\right)$

II. Choisir la valeur de départ de la chaîne $\left(\theta^{(1)}\right)$ *et calculer*

$$P\left(\theta^{(1)}|Y\right) \propto P\left(Y|\theta^{(1)}\right) \times \pi\left(\theta^{(1)}\right)$$

III. Pour n de 1 à p {

A. Générer $\theta^{*(n+1)}$ *en tirant une valeur de* $g\left(\theta^{*(n+1)}, \theta^{(n)}\right)$ *et calculer*

$$P\left(\theta^{*(n+1)}|Y\right) \propto P\left(Y|\theta^{*(n+1)}\right) \times \pi\left(\theta^{*(n+1)}\right)$$

B. Calculer le rapport

$$\alpha = \frac{P\left(\theta^{*(n+1)}|Y\right)}{P(\theta^{(n)}|Y)} \times \frac{g\left(\theta^{(n)}|\theta^{*(n+1)}\right)}{g\left(\theta^{*(n+1)}|\theta^{(n)}\right)}$$

Notre loi de proposition étant symétrique on a $g\left(\theta^{*(n+1)}\middle|\theta^{(n)}\right) = g\left(\theta^{(n)}|\theta^{*(n+1)}\right)$ *d'où*

$$\alpha = \frac{P\left(\theta^{*(n+1)}|Y\right)}{P(\theta^{(n)}|Y)}$$

Or

$$P(\theta|Y) \propto P(Y|\theta) \times \pi(\theta)$$

D'où

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

C. Si $\alpha \geq 1$ *:* $\theta^{(n+1)} = \theta^{*(n+1)}$

D. Si $\alpha < 1$ *: On tire une variable* $u \sim U(0,1)$

- Si $u > \alpha$ *alors* $\theta^{(n+1)} = \theta^{(n)}$
- Si $u < \alpha$ *alors* $\theta^{(n+1)} = \theta^{*(n+1)}$

}

L'échantillon de valeurs $\left(\theta^{(n)}\right)$ *est une approximation à la distribution a posteriori des paramètres*

8.5. Metropolis-Hastings within Gibbs

8.5.1. Modalités générales

Nous considérons maintenant le cas où la variance résiduelle est inconnue. L'algorithme construit une chaîne $(\theta_1, {\sigma_\varepsilon^2}_1)$, $(\theta_2, {\sigma_\varepsilon^2}_2), \dots, (\theta_N, {\sigma_\varepsilon^2}_N)$. L'algorithme est basé sur le fait qu'à chaque point de la chaîne, on peut d'abord générer une nouvelle valeur de θ puis une nouvelle valeur de σ_ε^2. Pour θ_{n+1}, il s'agit de générer une valeur de la distribution $P(\theta|Y, \sigma_\varepsilon^2)$. On le fera exactement comme dans le cas précédent avec σ_ε^2 connu. Puis pour ${\sigma_\varepsilon^2}_{n+1}$ il faut une valeur de la distribution $P(\sigma_\varepsilon^2|Y, \theta_{n+1})$. Mais cette distribution est connue :

$$P(\sigma_\varepsilon^2|Y,\theta) = P(Y|\sigma_\varepsilon^2,\theta)\pi(\sigma_\varepsilon^2)$$

$$P(\sigma_\varepsilon^2|Y,\theta) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}^{N/2}} \times e^{-\sum_{k=1}^{N}[Y_k - f(X_k,\theta)]^2/2\sigma_\varepsilon^2} \times \frac{1}{\sigma_\varepsilon^2}$$

On pose alors $\tau = \frac{1}{\sigma_\varepsilon^2}$ pour obtenir

$$P(\tau|Y,\theta) = C\tau^{N/2+1} e^{-\tau\sum_{k=1}^{N}[Y_k - f(X_k,\theta)]^2/2}$$

$$P(\tau|Y,\theta) \sim \Gamma\left(shape = \frac{N}{2} + 2, \qquad scale = \sum_{k=1}^{N} 2/\left[Y_k - f\left(X_k, \theta^{(n+1)}\right)\right]^2\right)$$

La distribution a posteriori est une distribution gamma. On peut tirer directement la variance résiduelle dans cette distribution. Cette façon d'échantillonner des distributions conditionnelles s'appelle l'échantillonneur de Gibbs (Gelman et al., 2004). Dans notre cas, on utilise la méthode MCMC pour tirer la valeur de θ_{n+1}. Cette variante de la méthode s'appelle « Metropolis Hastings within Gibbs ».

8.5.2. Algorithme et formalismes

I. Définir une distribution de proposition $g\left(\theta^{*(n+1)}, \theta^{(n)}\right)$

Ici on travaille avec $g\left(\theta^{*(n+1)}, \theta^{(n)}\right) = N\left(\theta^{(n)}, \Sigma\right)$

II. Choisir la valeur de départ de la chaîne $\left(\theta^{(1)}, \sigma_\varepsilon^{2(1)}\right)$ *et calculer*

$$P\left(\theta^{(1)}|Y, \sigma_\varepsilon^{2(1)}\right) \propto P\left(Y|\theta^{(1)}, \sigma_\varepsilon^{2(1)}\right) \times \pi\left(\theta^{(1)}\right)$$

III. Pour n de 1 à p {

A. Générer $\theta^{*(n+1)}$ *en tirant une valeur de* $g\left(\theta^{*(n+1)}, \theta^{(n)}\right)$ *et calculer*

$$P\left(\theta^{*(n+1)}|Y, \sigma_\varepsilon^{2(n)}\right) \propto P\left(Y|\theta^{*(n+1)}, \sigma_\varepsilon^{2(n)}\right) \times \pi\left(\theta^{*(n+1)}\right)$$

B. Calculer le rapport

$$\alpha = \frac{P\left(\theta^{*(n+1)}|Y,\sigma_\varepsilon^{2(n)}\right)}{P\left(\theta^{(n)}|Y,\sigma_\varepsilon^{2(n)}\right)} \times \frac{g\left(\theta^{(n)}|\theta^{*(n+1)}\right)}{g\left(\theta^{*(n+1)}|\theta^{(n)}\right)}$$

- *Si* $\alpha \geq 1$ *:*

a. $\theta^{(n+1)} = \theta^{*(n+1)}$

b. *Tirer* $\frac{1}{\sigma_\varepsilon^{2(n+1)}} \sim \Gamma\left(shape = \frac{N}{2} + 2,\ scale = \sum_{k=1}^{N} 2/\left[Y_k - f\left(X_k, \theta^{(n+1)}\right)\right]^2\right)$

c. Calculer

$$P\left(\theta^{(n+1)}|Y,\sigma_\varepsilon^{2(n+1)}\right) \propto P\left(Y|\theta^{(n+1)},\sigma_\varepsilon^{2(n+1)}\right) \times \pi\left(\theta^{(n+1)}\right)$$

- *Si* $\alpha < 1$ *: On tire une variable* $u \sim U(0,1)$

  - *Si* $u > \alpha$ *alors* $\theta^{(n+1)} = \theta^{(n)}$

    a. Tirer $\frac{1}{\sigma_\varepsilon^{2(n+1)}} \sim \Gamma\left(shape = \frac{N}{2} + 2,\ scale = \sum_{k=1}^{N} 2/\left[Y_k - f\left(X_k, \theta^{(n+1)}\right)\right]^2\right)$

    b. Calculer

$$P\left(\theta^{(n+1)}|Y,\sigma_\varepsilon^{2(n+1)}\right) = P\left(\theta^{(n)}|Y,\sigma_\varepsilon^{2(n+1)}\right) \propto P\left(Y|\theta^{(n)},\sigma_\varepsilon^{2(n+1)}\right) \times \pi\left(\theta^{(n)}\right)$$

  - *Si* $u < \alpha$ *alors* $\theta^{(n+1)} = \theta^{*(n+1)}$

    a. Tirer $\frac{1}{\sigma_\varepsilon^{2(n+1)}} \sim \Gamma\left(shape = \frac{N}{2} + 2,\ scale = \sum_{k=1}^{N} 2/\left[Y_k - f\left(X_k, \theta^{(n+1)}\right)\right]^2\right)$

    b. Calculer

$$P\left(\theta^{(n+1)}|Y,\sigma_\varepsilon^{2(n+1)}\right) \propto P\left(Y|\theta^{(n+1)},\sigma_\varepsilon^{2(n+1)}\right) \times \pi\left(\theta^{(n+1)}\right)$$

}

L'échantillon de valeurs $\left(\theta^{(n)}, \sigma_\varepsilon^{2(n)}\right)$ *est une approximation à la distribution a posteriori des paramètres et de la variance résiduelle.*

9. Choix de l'algorithme et modalités spécifiques

Les approches de type MCMC se révèlent, pour notre cas, bien plus performantes que les approches importance sampling. L'algorithme de Metropolis-Hastings donne des résultats satisfaisant et, couplé à l'échantillonneur de Gibbs, il nous permet d'obtenir une approximation à la distribution de la variance résiduelle du modèle. L'algorithme retenu pour la suite de notre étude est donc celui de Metropolis-Hastings within Gibbs.

Pour chaque situation du plan d'expérience, on fait partir deux chaînes de Markov avec des points de départs différents et suffisamment dispersés dans l'espace a priori des paramètres (plus ou moins 2 écart-type de la moyenne). On commence l'estimation avec une matrice de variance covariance diagonale avec, dans la diagonale, les variances a priori des paramètres (0 dans l'extra-diagonale). Le fait que cette matrice soit diagonale suppose l'indépendance a priori des paramètres car les covariances sont nulles. Après 20 000 itérations on fait un diagnostic de convergence un et on regarde le taux de rejet.

Si la convergence est atteinte assez rapidement on s'arrête là et on conserve les chaînes obtenues. Si la convergence est jugée trop longue (plus de 20 000 itérations) on calcule les variances et covariances des paramètres à partir des échantillons acceptés dans la précédente simulation et on refait une estimation avec cette nouvelle matrice de variance covariance en refaisant partir deux chaînes.

Pour améliorer l'efficacité de l'algorithme en terme de taux de rejet nous avons recours au « tuning » de la loi de proposition (Gelman et al., 2004). Cette méthode consiste à adapter de la matrice de variance covariance de la loi de proposition en fonction du taux de rejet des itérations précédentes. Toutes les 200 itérations, on calcule le taux de rejet des 200 itérations précédentes ; Si ce taux est supérieur à 80% on divise la matrice de variance covariance de la loi de proposition par 2, s'il est inférieur à 70% on la multiplie par 2. On rappelle que, pour notre cas, le taux de rejet optimal est d'environ 75%.

10. Estimation ponctuelle par maximum de vraisemblance

Pour s'assurer de ne pas être passé complètement à côté de vecteur de paramètres ayant de bonnes performances statistiques (qualité d'ajustement et de prédiction) et vérifier par la même occasion la « précision » de notre méthode par rapport à d'autres, nous avons réalisé, pour chaque modalité du facteur quantité de données du plan d'expérience, une estimation fréquentiste par maximum de vraisemblance sur 3 paramètres (les 3 jugés les plus incertains) les autres étant fixés à leur valeur nominale. On utilise l'algorithme de Nelder-Mead (Nelder et al., 1965)à partir de plusieurs points de départ. Les estimations obtenus permettront de comparer les estimateurs bayesien et fréquentiste en terme de qualité d'ajustement et de qualité de prédiction.

11. Evaluation de l'incertitude des prédictions

11.1. Le conseil d'Arvalis

Le modèle *Septo-LIS®* étant utilisé pour délivrer un conseil, nous nous intéressons ici à l'incertitude sur les prédictions de date de premier traitement. Les sorties « bruts » du modèle représentent l'évolution de la surface des lésions sur les 6 dernières feuilles du blé. Un seuil pour le déclenchement du traitement est établi par les chercheurs et ingénieurs d'Arvalis. La date de premier traitement est alors déterminée par la date à laquelle ce seuil est dépassé.

11.2. Incertitude due à la variabilité des paramètres

On dispose d'échantillons $(\theta_i, \sigma_{\varepsilon}^2{}_i)$ représentant la distribution a posteriori des paramètres et de la variance résiduelle du modèle. Pour évaluer l'incertitude des prédictions du modèle, on fait tourner le modèle avec les échantillons de la distribution a posteriori des paramètres. On obtient donc une distribution des prédictions mais la variabilité de cette distribution ne prend en compte que la variabilité des paramètres et pas la variabilité résiduelle du modèle. Autrement dit, si on considère la distribution obtenue ici comme distribution des prédictions du modèle, cela suppose que l'on considère que ce modèle ne commet pas d'erreur et que toute la variabilité des résultats s'explique par la variabilité des paramètres. Or l'hypothèse posée préalablement sur le modèle statistique définit clairement l'erreur ε et sa distribution $N(0, \sigma_{\varepsilon}^2)$.

11.3. Incertitude due à la variabilité résiduelle

Pour intégrer la variabilité due à la variance résiduelle dans la distribution des prédictions, on dispose de tous ce dont on a besoin ; une distribution de l'erreur qui dépend de la variance résiduelle et une approximation à la distribution de cette variance. On réalise donc toujours les prédictions avec la distribution des paramètres mais cette fois, pour chaque jeu de paramètres θ_i on tire une erreur aléatoire ε_i dans la loi de l'erreur $\varepsilon_i \sim N(0, \sigma_{\varepsilon}^2{}_i)$. On ajoute ensuite cette erreur aléatoire à toutes les prédictions pour un même jeu de paramètres.

11.4. Incertitude totale

L'incertitude totale dans les prédictions de *Septo-LIS®* est évaluée en combinant les deux principales sources de variabilité que nous venons de décrire ; la variabilité des paramètres et la variabilité résiduelle. Le fait de qualifier cette incertitude de « totale » implique que nous considérons les autres sources éventuelles de variabilité, telles que la structure des équations ou le choix des variables explicatives comme négligeables.

11.5. Correction des incohérences dues à l'utilisation de tirage aléatoire

Lorsque l'on tire une erreur aléatoire dans la loi de l'erreur, il est possible d'obtenir une erreur supérieure au seuil de déclenchement du traitement. Ce type de tirage conduit à la conclusion que la date de traitement optimale est le jour du semis ! En effet l'erreur commise par le modèle dans ce cas engendre un dépassement du seuil au premier jour de simulation qui correspond au premier jour de la culture et donc un traitement optimal ce jour ci. Or cette situation serait totalement incohérente d'un point de vue agronomique!

Cette conclusion n'ayant aucun fondement agronomique, nous pouvons légitimement attribuer la responsabilité de cette incohérence à l'utilisation de tirage aléatoire. Nous appliquons donc une correction pour les situations conduisant à celle-ci : Si l'erreur est supérieure au seuil de déclenchement du traitement, on considère la date à partir de laquelle le modèle donne une valeur de symptômes différente de zéro. Le fait d'appliquer cette correction suppose que nous considérons l'erreur du modèle comme négligeable tant que les valeurs calculées sont égales à zéro.

11.6. Algorithme et formalismes

On dispose d'échantillons $(\theta_i, \sigma^2_{\varepsilon\,i})$ $i \in \{1,2,\dots,p\}$ *et d'un seuil s pour le déclenchement du traitement*

Pour i de 1 à p

{

I. *Calculer* $Y_{ki}(t) = f(X_k(t), \theta_i)$

II. *Tirer* $\varepsilon_i \sim N(0, \sigma^2_{\varepsilon\,i})$

III. *Calculer* $\tilde{Y}_{ki}(t) = Y_{ki}(t) + \varepsilon_i$

IV. *Si* $\varepsilon_i > s$

Prendre $Date_i = \text{Min}\{t(\, Y_{ki}(t) > 0)\}$

Sinon
Prendre $Date_i = \text{Min}\{t(\, \tilde{Y}_{ki}(t) > s)\}$

}

L'échantillon de valeurs $(Date_i)$ *est une approximation à la distribution des dates de traitement.*

(Le code R de cet algorithme figure en annexe I)

12. Logiciel et ressources informatiques

Le code source du modèle nous ayant été fourni sous forme d'un script R, tous les algorithmes implémentés et toutes les études statistiques nécessaires à l'achèvement du projet ont été réalisés sur le logiciel R. Ce logiciel présente pour nous des avantages non-négligeables ; tout d'abord il est totalement gratuit et entièrement programmable. Il permet d'effectuer une gamme importante d'analyses statistiques qui sont pré-implémentées, dans la distribution de base pour les analyses classiques ou dans des librairies (ou packages) téléchargeable gratuitement pour les méthodes récentes. De plus, R offre un vaste éventail de méthodes graphiques pour la visualisation des données et résultats d'analyses. Enfin R est multi-plateforme et multi-OS c'est-à-dire qu'il peut fonctionner sur des ordinateurs utilisant des systèmes d'exploitations différents.

Malgré tous ces avantages, R possède tout de même un défaut assez handicapant pour nous ; la lenteur dans l'exécution de boucle de calcul et la manipulation de tableaux de grandes taille. Or les méthodes décrites précédemment font effectivement appel à des structures de répétition type boucles (ex : *Pour i de 1 à p*) et nécessitent de stocker un très grand nombre de données dans des tableaux. Ce défaut nous a amené à considérer plus sérieusement la question du temps de calcul et à l'intégrer à la méthode en tant que composante à optimiser.

Pour optimiser la vitesse d'exécution des scripts sans changer de logiciel, nous avons dû trouver une solution. La parallélisation des calculs (la figure 3 schématise ce processus) est apparu comme la plus innovante et efficace des solutions existantes

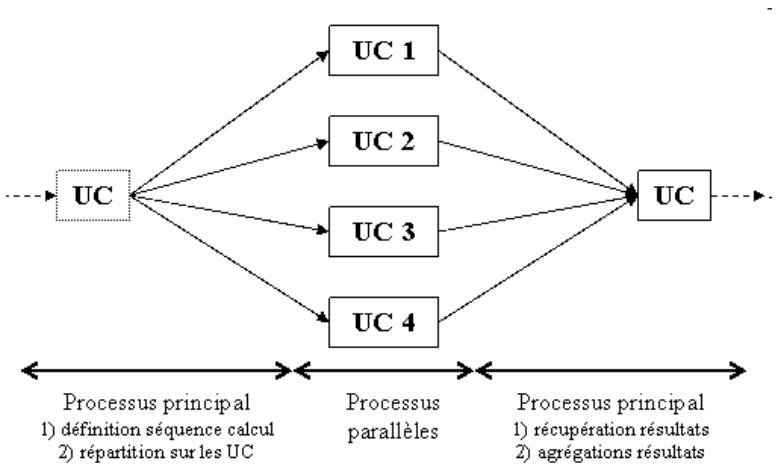


Figure 3 : Illustration du calcul parallèle sur 4 unités de calcul (UC)

La plupart des ordinateurs actuels sont équipés de processeur multi-cœur. Ce sont des processeurs qui présentent deux à quatre unités de calcul indépendantes. Ainsi ces ordinateurs sont capables de gérer différentes tâches simultanées de manière efficace. Cette apparition est liée à une limite technologique sur l'augmentation de la puissance des processeurs (fréquence) et on est passé de la course à la puissance des processeurs, à l'apparition de processeurs multi-cœurs ces dernières années. Néanmoins, cette puissance distribuée sur plusieurs cœurs reste très sous-utilisée, notamment lorsqu'on lance un processus calculatoire important (sous R par exemple).

Le calcul parallèle s'appui sur l'existence de plusieurs unités de calcul disponibles (N unités) que l'on peut mobiliser et avec lesquelles le processus principales peut communiquer. Le processus principal va définir une séquence de calcul, la diviser en N sous-séquence indépendantes et envoyer chacune sur les différentes unités de calcul. Ensuite, les N unités de calcul prennent la main et chacune traite de manière indépendante les calculs qui lui sont demandés et renvoie les résultats correspondant. Enfin, le processus reprend la main pour récupérer les N sous-séquences de résultats et les agréger.

III. RESULTATS ET DISCUSSIONS

1. Comparaison des estimateurs

Nous disposons de deux types d'estimateurs ; les estimateurs fréquentistes et les estimateurs bayesiens. L'estimateur fréquentiste est ici le maximum de vraisemblance. Pour les estimateurs bayesiens, nous avons sélectionné trois types d'estimateurs :

-Moyenne de la distribution a posteriori $E(\theta|Y)$

-Maximum de la distribution a posteriori $\text{Max}(\theta|Y)$

-Maximum de la vraisemblance $\text{Max}(Y|\theta)$

Pour la comparaison des performances des différents estimateurs, on utilise deux types de critères. Pour la qualité d'ajustement, on regarde la MSE (Mean Squared error) qui correspond à la moyenne de la somme des carrés des écarts entre valeurs mesurées et valeurs prédites (les valeurs mesurées étant celles utilisées pour l'estimation). Pour la qualité de prédiction, on regarde l'estimateur de MSEP (Mean Squared error of prediction). Ce critère se calcul exactement comme le précédent mais en utilisant de nouvelles données non utilisées pour l'estimation (jeu de validation).

- 1 site année

	MSE	*MSEP*
Fréquentiste MV	0.0473	0.0707
5% d'incertitude a priori sur 14 parametres		
Bayésien E(θ|y)	0.0507	0.0761
Bayésien Max(θ|y)	0.0537	0.0778
Bayésien Max(y|θ)	0.0453	0.0682
25% d'incertitude a priori sur 14 parametres		
Bayésien E(θ|y)	0.0577	0.0917
Bayésien Max(θ|y)	0.0687	0.0664
Bayésien Max(y|θ)	0.0319	0.0590

Tableau 1 : Qualité d'ajustement et de prédiction des différents estimateurs pour l'estimation utilisant 1 site année

- 8 sites année

	MSE	*MSEP*
Fréquentiste MV	0.0890	0.0634
5% d'incertitude a priori sur 14 parametres		
Bayésien E(θ|y)	0.0893	0.0598
Bayésien Max(θ|y)	0.0905	0.0629
Bayésien Max(y|θ)	0.0735	0.0586
25% d'incertitude a priori sur 14 parametres		
Bayésien E(θ|y)	0.0606	0.0594
Bayésien Max(θ|y)	0.0645	0.0567
Bayésien Max(y|θ)	0.0415	0.0894

Tableau 2 : Qualité d'ajustement et de prédiction des différents estimateurs pour l'estimation utilisant 8 sites année

Les estimations menées avec un seul site-année ne permettent pas réellement de tirer des conclusions générales ; en effet l'échantillon utilisé est très peu représentatif de la population cible. Néanmoins on remarque que les performances statistiques des estimateurs bayésiens et fréquentiste sont comparables.

La comparaison des deux modalités (5% et 25% d'incertitude a priori sur 14 paramètres) de l'estimation bayesienne sur 8 sites années (tableau 2) montre des différences appréciables : la modalité à 25% d'incertitude a priori donne un bien meilleur ajustement aux données utilisées pour l'estimation ce qui est légitime ; en augmentant la variance des lois a priori on donne moins de poids à ces lois et plus de poids aux données pour l'estimation. La modalité à 5% d'incertitude a priori conduit à un ajustement comparable à celui de l'estimation fréquentiste.

En terme de qualité de prédiction (ajustement des paramètres du modèle à de nouvelles données non utilisées pour l'estimation), nos deux modalités bayesiennes donnent de meilleurs résultats (critère MSEP plus faible) que l'estimation fréquentiste par maximum de vraisemblance, à l'exception de l'estimateur Max($Y|\theta$) sur la modalité 25%. Ceci est assez surprenant mais pas insensé ; en général le vecteur de paramètres qui s'ajuste le mieux au jeu d'entraînement a moins de chance de bien s'ajuster à de nouvelles données (jeu de validation). En revanche un vecteur de paramètres qui présente un ajustement moyen au jeu d'entraînement aura plus de chance de bien s'ajuster au jeu de validation et donc de bien prédire dans de nouvelles situations (Tremblay et Wallach, 2004 ; Makowski et al., 2006). Ce cas général ne s'applique ici qu'aux résultats fréquentiste et à l'estimateur bayesien Max($Y|\theta$) sur la modalité 25% d'incertitude : le meilleur vecteur sur le jeu d'entraînement est le moins bon sur le jeu de validation et vice et versa. Pour les résultats bayesiens, à l'exception de l'estimateur Max($Y|\theta$), le meilleur vecteur sur le jeu d'entraînement est aussi le meilleur pour la prédiction.

2. Distribution a posteriori

Pour illustrer les résultats de distribution a posteriori on se limitera ici à une condition du plan d'expérience (8 sites-année, 25% d'incertitude à priori sur 14 paramètres), à la variance résiduelle du modèle et à 6 paramètres :

-Coeff.mult : Coefficient d'auto contamination
-Echelle : Coefficient ramenant le pool d'inoculum à l'échelle des feuilles
-Prod.inoc : Production d'inoculum (spores)
-Vit.nec : Vitesse d'extension des lésions
-Appar.nec : Taille initiale des lésions
-Vit.expan : Vitesse d'expansion des feuilles

2.1. Pour les paramètres

Tous les résultats de l'estimation de distribution a posteriori de paramètres seront présentés, pour chaque paramètre, de la façon suivante :
-La trace des deux chaînes de Markov (valeurs retenues dans la chaîne en fonction des itérations),
-La trace de l'estimateur $E(\theta|Y)$ pour les deux chaînes (en fonction des itérations) avec la moyenne de la distribution a priori en pointillé,
-Les distributions a posteriori (en rouge) et a priori (en pointillé).

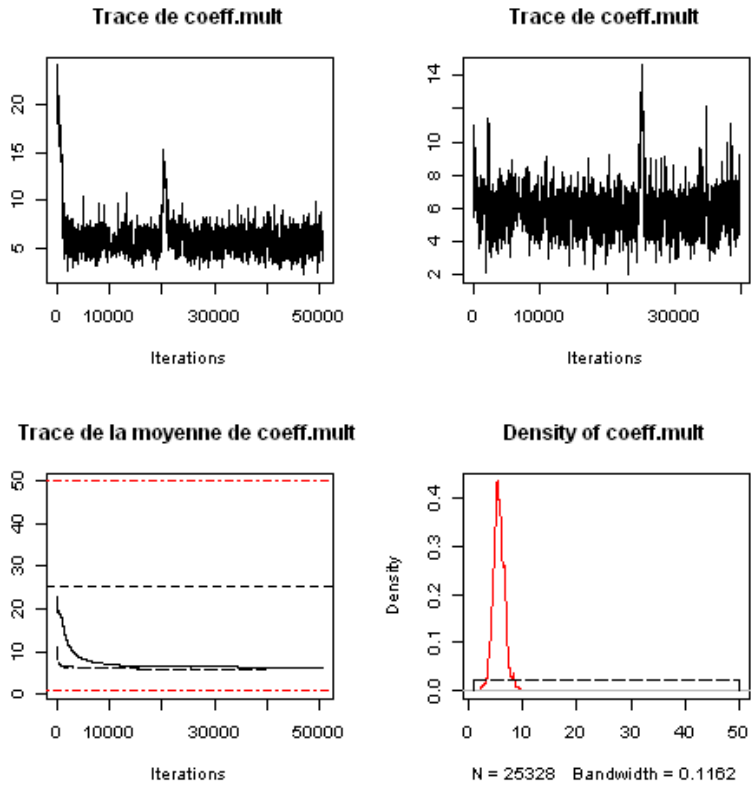


Figure 4: Résultats de l'estimation du paramètre coeff.mult

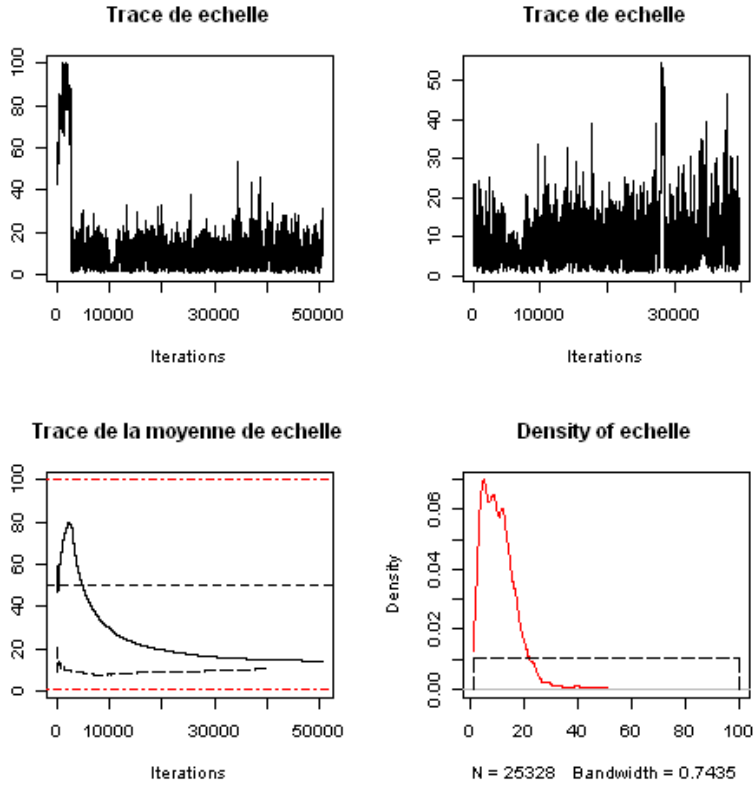


Figure 5: Résultats de l'estimation du paramètre echelle

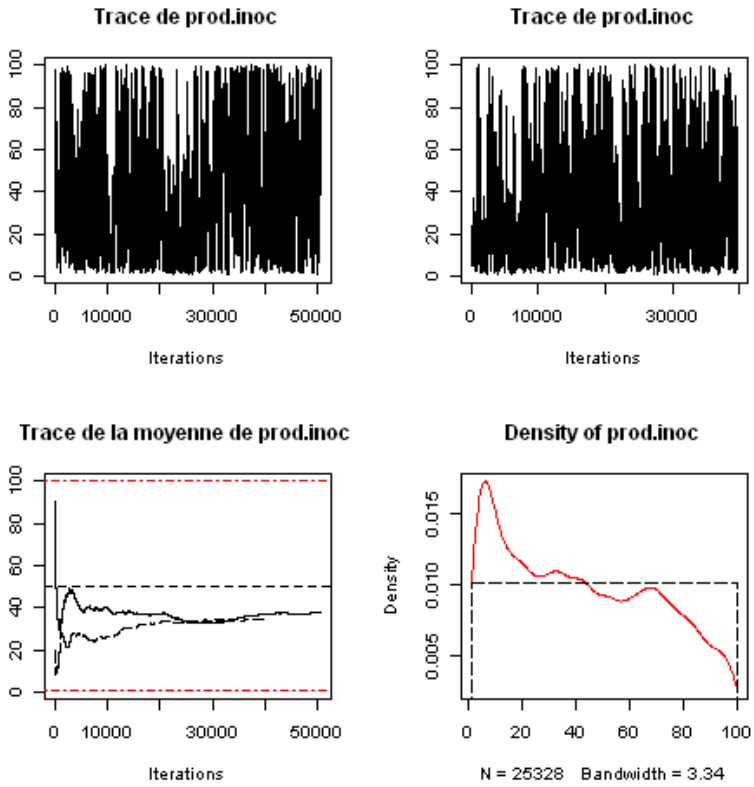


Figure 6: Résultats de l'estimation du paramètre prod.inoc

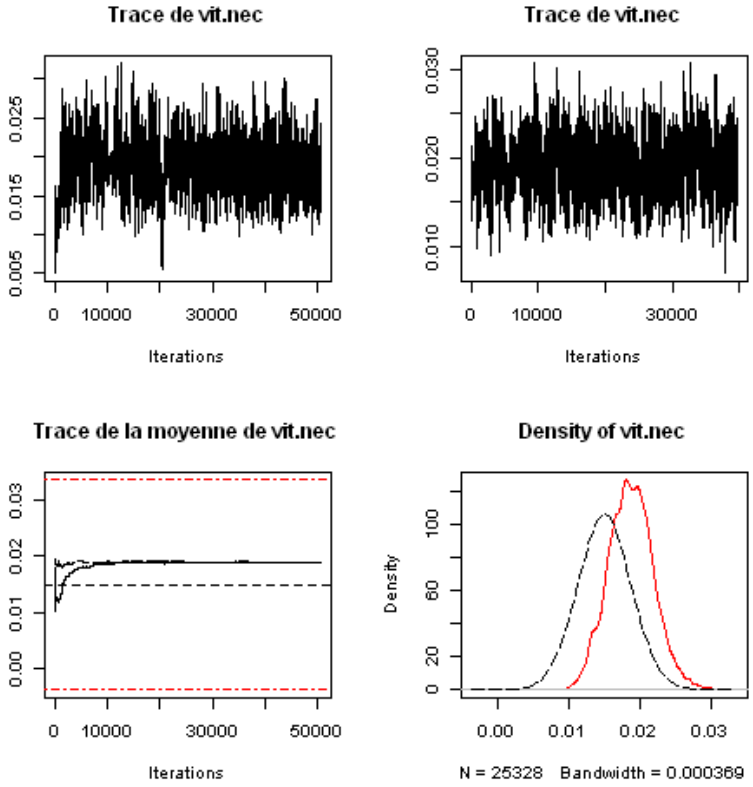


Figure 7: Résultats de l'estimation du paramètre vit.nec

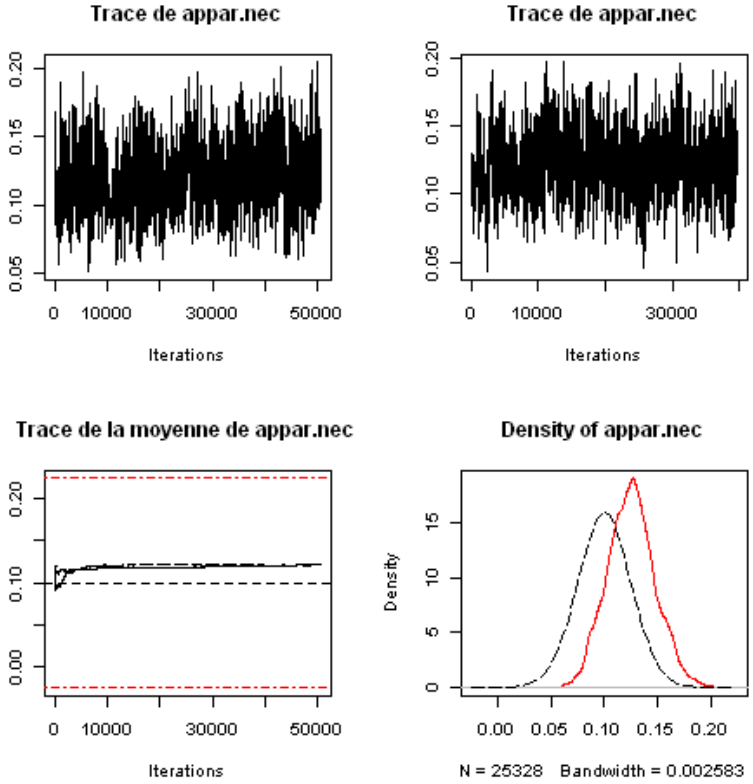


Figure 8: Résultats de l'estimation du paramètre appar.nec

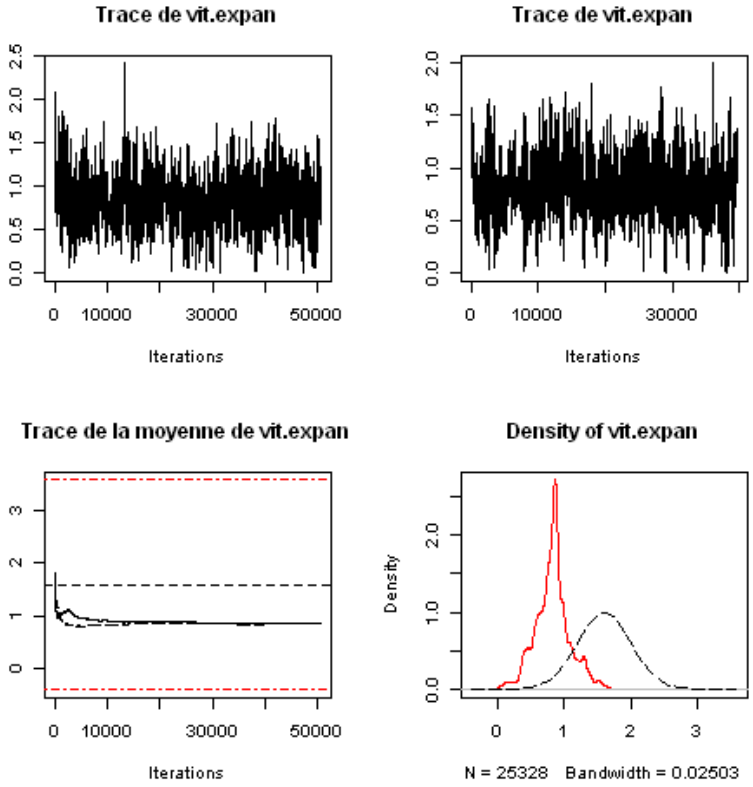


Figure 9: Résultats de l'estimation du paramètre vit.expan

L'examen visuel des traces des deux chaînes de Markov est très utile (et très simple) pour évaluer les caractéristiques intrinsèques de ces chaînes. Il permet en premier lieu de s'assurer, sans pour autant avoir valeur de diagnostic, d'une relative stabilité de la distribution. Par exemple le graphe en haut à gauche de la Figure 5 montre que la convergence vers une distribution stationnaire semble avoir été atteinte après environ 3000 itérations. En second lieu, cet examen permet d'évaluer le « mixage » de la chaîne c'est-à-dire sa capacité à se déplacer rapidement d'une région de la distribution à une autre. Si le mixage est bon, la chaîne doit pouvoir se déplacer dans la distribution d'une région éloignée des fortes probabilités à une autre en quelques itérations. Dans le cas contraire, la chaîne reste bloquée de longues périodes dans les mêmes régions et le nombre d'itérations nécessaires pour avoir une bonne approximation de la distribution a posteriori sera trop grand.

On peut aussi évaluer graphiquement la convergence en représentant la moyenne de la distribution des deux chaînes à chaque itération (Robert and Casella, 2004). Par exemple le graphe en bas à gauche de la Figure 5 montre qu'après 30 000 itérations, les moyennes des distributions des deux chaînes ne sont toujours pas confondues. Néanmoins on peut observer une tendance à la convergence vers la même moyenne. Pour les autres paramètres excepté prod.inoc, on peut observer des convergences très rapides (moins de 10 000 itérations) vers une même moyenne de la distribution.

Ici, toutes les chaînes semblent graphiquement avoir convergé et le mixage des chaînes est satisfaisant. On rappelle cependant que ce type d'examen n'a pas valeur de diagnostic et qu'il permet principalement de détecter les gros problèmes de convergence et de mixage (mais pas de les résoudre).

Pour diagnostiquer formellement la convergence de nos chaînes vers une distribution stable on réalise le test de *Gelman et Rubin* (Gelman et al., 1992) qui compare les variances inter et intra chaînes. La valeur retournée par le test correspond à un facteur de réduction potentiel de la largeur de l'intervalle crédible de la distribution. Si cette valeur est proche de 1 (en pratique inférieur à 1,1), les deux chaînes ont convergé vers la même distribution stable, sinon on peut garder la même loi de proposition et prolonger la simulation ou alors changer de loi de proposition et refaire partir deux chaînes. Le Tableau 3 présente le résultat du diagnostic pour tous les paramètres et la variance résiduelle du modèle.

Paramètres	*Psrf*	*Paramètres*	*Psrf*
coeff.mult	1.02	Pp	1.00
seuil.cumul	1.00	prod.inoc	1.00
echelle	1.04	vit.expan	1.00
transf.pluie	1.01	limite	1.00
transf.haut	1.01	min.lorin	1.01
vit.nec	1.00	Tlim.lorin	1.00
appar.nec	1.00	epuis.inoc	1.02
Tminjour	1.00	limiteF1	1.02
Tminveille	1.00	variance.erreur	1.00
Multivariate	1.07		

Tableau 3 : Statistique de Gelman et Rubin pour tous les paramètres et la variance résiduelle du modèle (Psrf= Potential scale reduction factor). La dernière ligne correspond au même diagnostic sur l'ensemble des paramètres

Le diagnostic de Gelman et Rubin nous indiques ici que nos chaînes ont bien convergé vers une même distribution stable et que la prolongation de la simulation ne permettrait pas de réduire la largeur de l'intervalle crédible des distributions recherchées. On pourra donc utiliser ces chaines en les considérants comme étant des approximations à la distribution a posteriori des paramètres.

Disposant de chaînes de valeurs représentant des approximations aux distributions des paramètres de notre modèle, nous pouvons établir les statistiques de cette distribution : Moyenne, variance et intervalle crédible à 95%. Le tableau 4 présente ces statistiques et permet de les comparer aux statistiques des distributions a priori.

	PRIOR			*POSTERIOR*		
	Mean	*Var*	*IC 95%*	*Mean*	*Var*	*IC 95%*
coeff.mult	25.5	200.083	[2.225 ; 48.775]	5.7373	1.0549	[3.615 ; 7.670]
seuil.cumul	20.698	26.776	[10.556; 30.840]	20.1577	27.282	[0.189 ; 30.504]
echelle	50.5	816.75	[3.475 ; 97.525]	10.6837	42.638	[1.015 ; 21.968]
transf.pluie	-0.28	0.0049	[-0.417; -0.143]	-0.3344	0.0031	[-0.449; -0.23]
transf.haut	0.160	0.0016	[0.082 ; 0.238]	0.1480	0.0013	[0.076 ; 0.215]
vit.nec	0.015	1,00E-05	[0.008 ; 0.022]	0.0187	9.61e-06	[0.013 ; 0.025]
appar.nec	0.10	0.00062	[0.051 ; 0.149]	0.1251	5,00E-04	[0.083 ; 0.168]
Tminjour	0.043	0.00012	[0.022 ; 0.065]	0.0423	1,00E-04	[0.021 ; 0.062]
Tminveille	6.756	2.85309	[3.446 ; 10.067]	6.7825	2.8696	[3.516 ; 10.018]
Pp	0.119	0.00089	[0.061 ; 0.178]	0.1163	0.001	[0.057 ; 0.178]
prod.inoc	50.5	816.75	[3.475 ; 97.525]	42.0501	795.091	[1.099 ; 90.444]
vit.expan	1.60	0.16	[0.816 ; 2.384]	0.8249	0.0673	[0.336 ; 1.384]
limite	9.0	5.0625	[4.590 ; 13.410]	7.0456	5.4925	[2.440 ; 11.618]
min.lorin	3.509	0.76984	[1.790 ; 5.229]	4.0763	0.7223	[2.387 ; 5.698]
Tlim.lorin	8.020	4.02032	[4.090 ; 11.950]	6.0706	2.4955	[3.158 ; 9.257]
epuis.inoc	7.685	3.69157	[3.920 ; 11.451]	6.8944	4.9292	[2.413 ; 11.114]
limiteF1	9.351	5.46584	[4.769 ; 13.934]	9.3879	5.2517	[4.895 ; 13.577]

Tableau 4 : Statistique des distributions a priori et a posteriori des 17 paramètres du modèle

L'analyse du tableau 4 montre que globalement les variances et les largeurs d'intervalle crédible à 95% des distributions a posteriori sont inférieures à ceux des distributions a priori. Cela montre que les données utilisées pour l'estimation apportent effectivement une information supplémentaire par rapport aux informations disponibles a priori. On peut néanmoins noter ici que, pour certains des paramètres (notamment Pp), l'estimation donne une distribution a posteriori quasiment identique a la distribution a priori.

Ce type de tableau est très utile pour résumer de manière synthétique les résultats d'une estimation bayesienne. Cependant il reste indissociable des graphes de distributions a posteriori car il ne permet pas à lui seul de rendre compte de certaines caractéristiques intrinsèque des distributions notamment sur leur forme ; Symétrie ou Asymétrie, Uni-, Bi-, Multimodale. Or, la méconnaissance de ces caractéristiques peut biaiser le calcul de l'intervalle crédible.

2.2. Pour la variance résiduelle

On utilise ici le même type de présentation appliquée aux paramètres pour présenter la distribution a posteriori de la variance résiduelle du modèle.

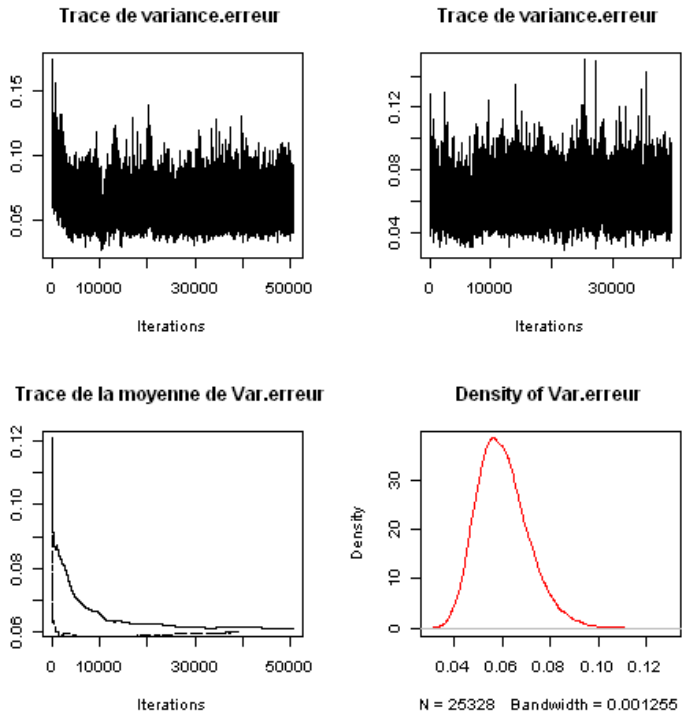


Figure 10 : Résultats de l'estimation de la variance résiduelle du modèle

Les deux graphes en haut de la Figure 10 montrent que la convergence vers une distribution stable est ici extrêmement rapide et que le mixage des chaînes est optimal voire parfait. Cette rapidité de convergence couplée à la quasi perfection du mixage peut s'expliquer par le fait que l'échantillonnage de la variance résiduelle n'est pas soumis à l'étape d'acceptation/rejet dans l'algorithme de *Metropolis-Hastings within Gibbs*. En effet une fois que l'on a accepté un vecteur de paramètres, on tire la variance résiduelle directement dans sa distribution a posteriori pour laquelle on dispose d'une expression analytique. Les bonnes performances de convergence et de mixage de la chaîne sont donc uniquement dues à la bonne qualité du générateur aléatoire utilisé ainsi qu'à la valeur de départ de la variance résiduelle.

3. Distribution et intervalle crédible des prédictions

Ici encore on se limitera à une condition du plan d'expérience (8 sites années et 25% de niveau d'incertitude a priori sur 14 paramètres) pour présenter la distribution des dates de premier traitement. A partir de l'échantillon de dates obtenu on effectue un lissage par un noyau Gaussien. La figure 11 présente ce lissage de la distribution de notre échantillon. On représente sur la même figure la distribution empirique, le mode et l'intervalle crédible à 95%.

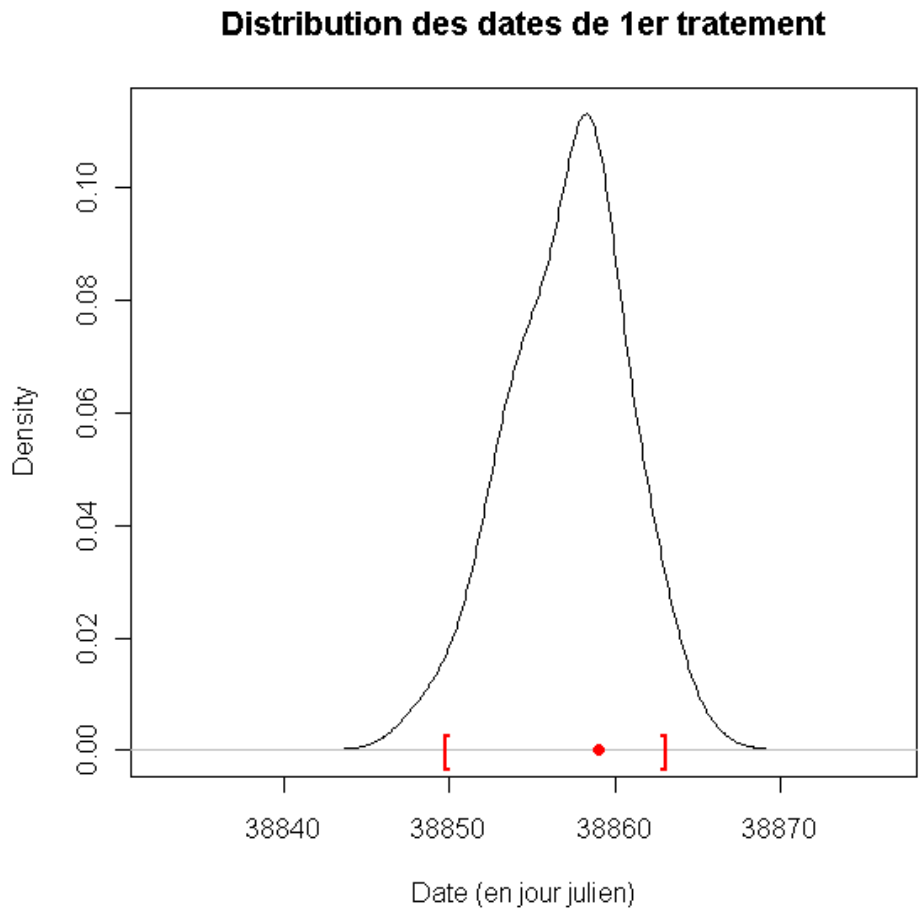


Figure 11 : Distribution des dates de premier traitement conseillées par *Septo-LIS®*. Le point indique la date la plus probable (mode de la distribution), les crochets délimitent l'intervalle crédible à 95%

Cette distribution permet donc effectivement de bien visualiser la dispersion des dates de 1er traitement possibles autour d'une date plus probable (représentée par le point rouge sur la figure 11). Notre méthode a donc permis d'intégrer de la stochasticité dans un modèle totalement déterministe à la base.

L'intervalle crédible à 95% de la distribution des dates de 1er traitement est [38850 ; 38863] avec un mode à 38859. Notre méthode de représentation de l'incertitude du modèle peut donc faire la différence avec un conseil qui ne se baserait que sur une valeur correspondant au mode de la distribution. Ici le conseiller pourra moduler ces préconisations en fonction de ses connaissances expertes sur son domaine géographiques et agronomiques. Par exemple, on peut observer une légère asymétrie dans la distribution ; Si l'agriculteur ne peut pas traiter au jour 38859 il sera préférable de lui conseiller de traiter un peu avant. En effet les probabilités de traitement sont plus élevées à gauche qu'à droite du mode.

Cette façon d'évaluer l'incertitude représente donc une évolution dans l'ergonomie et la souplesse d'utilisation des modèles car elle permet de moduler les préconisations non plus uniquement en fonction des réalités du terrain mais aussi à l'aide d'une « réalité virtuelle » facilement interprétable.

4. Critique de la méthode

La méthode d'estimation couplée a notre méthodologie d'évaluation d'incertitude fonctionne et permet effectivement de mieux rendre compte de l'imperfection et donc de l'incertitude du modèle. Cependant sa mise en place n'a pas été sans problème. Tout d'abord la définition des lois a priori pour les 17 paramètres n'a pas été évidente ; en effet, les experts et la littérature ont beaucoup de mal à s'accorder sur l'incertitude de 14 des 17 paramètres. Ceci explique notre choix de formaliser un plan expérience dont l'un des facteurs est le niveau d'incertitude a priori. Pourtant la distribution réelle d'un paramètre est unique et les différentes lois a priori donnent des résultats différents et dont l'analyse est difficilement généralisable, de plus cette loi peut introduire de la subjectivité dans l'estimation si elle est très mal choisie. D'autre part, du fait du temps de calcul considérable, nous n'avons pas pu finaliser les résultats de l'estimation sur le jeu de données complet (157 site-années). Or c'est avec ce jeu de données que l'on pourrait tirer de réelles conclusions. Pour palier à ces problèmes d'ordre technique et méthodologique, nous allons mettre en place une expérience de simulation décrite dans les perspectives.

5. Perspectives

L'étude présentée ici a montré qu'il était possible d'employer des méthodes bayesiennes pour approximer la distribution des paramètres d'un modèle mécaniste et d'utiliser ces distributions pour évaluer l'incertitude du modèle. C'était l'objectif principal du stage qui est donc atteint. A partir de là, il ya encore plusieurs possibilités pour améliorer la méthode. Nous avons notamment proposé d'améliorer le modèle statistique de l'erreur en se penchant sur les corrélations des erreurs entre site-année et entre feuilles d'un même site-année (volontairement négligées au départ) et, s'il y a lieu, d'inclure un ou deux niveau hiérarchique supplémentaire (correspondant aux sites-année ou aux feuilles) dans l'estimation.

Malgré les efforts déployés pour réduire le temps de calcul, il reste encore très important voire problématique (plus d'une semaine entre l'estimation des paramètres et le calcul de l'incertitude sur la date de traitement). Une des solutions envisageables serait de ne pas faire les calculs avec R et d'opter pour un logiciel (payant) optimisé pour les boucles.

De plus nous n'avons pas abandonné notre objectif de définir une démarche d'analyse applicable à d'autres modèles semblables. Nous allons donc mettre en place une expérience de simulation pour déterminer et analyser l'effet de la quantité de données et d'informations a priori sur les résultats de l'estimation des paramètres et l'évaluation de l'incertitude qui en découle. Pour ce faire, nous définirons un modèle linéaire pour lequel les paramètres seront tirés aléatoirement dans une distribution. On réalisera ensuite l'estimation bayesienne avec différentes distributions a priori pour les paramètres (dont celle dans laquelle on les a tirés) et différent jeux de données virtuels. L'idée principale est de pouvoir faire des calculs suffisamment rapides pour tester plusieurs conditions et vérifier si la « meilleure » loi a priori (la plus objective) est bien celle d'où proviennent les paramètres.

Un dernier point concerne l'utilisation du modèle. Les spécialistes qui mettent au point les règles de décisions ne se basent pas uniquement sur la prédiction des symptômes mais aussi sur celle du niveau d'infection, qui n'est pas observable. Pourtant le modèle à été calibré et évalué uniquement par rapport aux dégâts visibles. Que peut-on dire alors quant à la qualité de prédiction du niveau d'infection ? Une étude est en cours pour définir des indicateurs du niveau d'infection. Les méthodes de PCR quantitative donnent des indicateurs moléculaires qui semblent apporter un plus pour la calibration sur des variables non observables.

CONCLUSIONS

Nous avons montré dans cette étude comment mettre en œuvre une approche bayesienne pour l'estimation des paramètres d'un modèle mécaniste. Toutes les méthodes d'estimation ne se sont pas révélées aussi efficaces ; ainsi la méthode « importance sampling » s'est avérée totalement infructueuse tant qu'il n'y a pas de solides informations sur la forme et le mode de la distribution a posteriori. En revanche, les méthodes MCMC employées avec l'algorithme de Metropolis Hastings et avec l'échantillonneur de Gibbs ont donné des résultats très satisfaisants même si les temps de calcul demeurent problématiques. La méthode d'évaluation d'incertitude que nous proposons ensuite permet de mieux prendre en compte l'influence des différentes sources d'incertitudes identifiées (ici paramètres et variance résiduelle) dans la définition de l'incertitude sur les prédictions et donc dans les conseils d'Arvalis qui seront le support de la décision. Il en découle une évolution dans la flexibilité d'utilisation du modèle en offrant plus grande aisance d'interprétations de ces sorties pour les tous acteurs du conseil agricole.

Ce projet a duré 6 mois. C'est très court pour travailler avec un modèle complexe, en y appliquant une gamme de méthodes, parfois de façon assez originale. Néanmoins, on estime que ce travail apporte des améliorations ainsi que des informations nouvelles et utiles sur *Septo-LIS®* et qu'il contribue à définir une démarche d'analyse applicable pour une certaine classe de modèles mécanistes.

BIBLIOGRAPHIE

Audsley E, Milne A, Paveley N, 2005. A foliar disease model for use in wheat disease management decision support systems. *Annals of Applied Biology* **147**: 161-172.

Coakley, S. M., L. R. McDaniel, and G. Shaner. 1985. Model for predicting severity of Septoria tritici blotch on winter wheat. *Phytopathology* 75 (11):1245-1251.

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

Gelman A., J.B.carlin, H.S.Stern, and D.B.Rubin. 2004. Bayesian data analysis. Chapman and Hall/CRC, Boca Raton.

Geman S., Geman D., 1984. Stochastic relaxation, Gibbs distribution, and the Bayesian Restoration of Images. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 6 721-741

Gladders, P., Paveley, N. D., Barrie, I. A., Hardwick, N. V., Hims, M. J., Langton, S., and Taylor, M. C. 2001. Agronomic and meteorological factors affecting the severity of leaf blotch caused by *Mycosphaerella graminicola* in commercial wheat crops in England. *Annals of Applied Biology* 138: 301-311.

Hue C., Tremblay M., Wallach D., 2008. A Bayesian approach to crop model calibration under unknown error covariance. *Journal of Agricultural, Biological, and Environmental Statistics* 13 355-365

Jorgensen L.N., M. Jahn, W. Clark, D. Antichi, T. Goral, H. Schepers, P. Lucas, B. Rolland, D. Gouache, L. Hornok, 2008. Endure Wheat Case Report

Makowski D., Denis J.B., Ruck L., Penaud A., 2008. A Bayesian approach to asses the accuracy of a diagnostic test based on plant disease measurement. *Crop Protection* 27 1187-1193

Makowski D., J.Hillier, D.Wallach, B.Andrieu, and M.H.Jeuffroy. 2006. Parameter estimation for crop models. p. 101-150. *In* D.Wallach, D.Makowski, and J.W.Jones (ed.) Working with dynamic crop models. Elsevier, Elsevier.

Maufras J. Y., Maumene C., Couleaud G., Bousquet N. 2006. Comment résister à la résistance? *Perspectives Agricoles* 328, 30-36.

Myers R.H. 2007. Classical and modern regression with applications. PWS-Kent, Boston.

Nelder J.A. and Mead R. 1965. A simplex algorithm for function minimization. *Computer Journal* 7, 148-154.

Pietravalle, S., Shaw M. W., Parker S. R., van den Bosch F. 2003. Modeling of relationships between weather and Septoria tritici epidemics on winter wheat: a critical approach. *Phytopathology* 93 (10):1329-1339.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Rinella M.J., Luschei E.C., 2007 Hierarchical Bayesian methods estimate invasive weed impacts at pertinent spatial scales. *Biol Invasions* 545-558

Robert,C.P., and G.Casella. 2004. Monte Carlo statistical methods. Springer, New York.

Schabenberger,O., and F.J.Pierce. 2002. Contemporary statistical Models for the plant and soil sciences. CRC Press, Boca Raton.

Shaw M.W., 1987. Assessment of upward movement of rain splash using a fluorescent tracer method and ist application tot he epidemiology of cereal pathogens. *Plant Patholoy*, 36, 201-213

Shaw M.W., Royle D.J., 1993. Factor determining the severity of epidemics of *Mycosphaerella graminicola* (*Septoria tritici*) on winter wheat in the UK. *Plant Patholoy*, 42, 882-899

te Beest D. E., Shaw M. W., Pietravalle S., van den Bosch F. 2009. A predictive model for early-warning of Septoria leaf blotch on winter wheat. *European Journal of Plant Pathology* 124 (3):413-425.

Tremblay M., Wallach D., 2004. Comparison of parameter estimation methods for crop models. *Agronomie* 24 351-365

Wallach,D., B.Goffinet, J.E.Bergez, P.Debaeke, D.Leenhardt, and J.-N.Aubertot. 2001. Parameter estimation for crop models: a new approach and application to a corn model. *Agronomy Journal* 93:757-766.

Martyn Plummer, Nicky Best, Kate Cowles and Karen Vines (2009). coda: Output analysis and diagnostics for MCMC. R package version 0.13-4. http://CRAN.R-project.org/package=coda

Deepayan Sarkar (2010). lattice: Lattice Graphics. R package version 0.18-3. http://CRAN.R-project.org/package=lattice

Alan Genz and Adelchi Azzalini (2009). mnormt: The multivariate normal and t distributions. R package version 1.3-3. http://CRAN.R-project.org/package=mnormt

ANNEXES

Fonction R implémentée pour le calcul de la distribution des dates de 1er traitement

```
#INPUT:
######
#-pheno = data frame des données phénotypiques
#-DirOutput = chemin du dossier où se trouvera le fichiers .csv
#-k = data frame contenant la chaine de markov des paramètres
#-seuil = Valeurs numérique seuil pour le déclenchement du 1er traitement

#OUTPUT:
#######
#Le fichier "nomcsv2" contient pour chaque itération :
#-La date de premier traitement
#-La variance des erreurs pour le vecteur de paramètres utilisé
#-L'erreur tirée aléatoirement dans N(0,racine de variance des erreurs)

   Ecophyto<-function(pheno,k,DirOutput,nomcsv2,seuil)
{
####################
#   Itération 1    #
####################
#Sélection du 1er vecteur de paramètres
 parametres<-as.numeric(k[1,1:17])
 names(parametres)<-c("coeff.mult","seuil.cumul","echelle","transf.pluie",
 "transf.haut","vit.nec","appar.nec","Tminjour","Tminveille","Pp",
 "prod.inoc","vit.expan","limite","min.lorin","Tlim.lorin","epuis.inoc"
 ,"limiteF1")
#Sélection de la variance associée à ce vecteur de paramètres
 var.erreur<-k[1,c("variance.erreur")]

#Lancement du modèle avec le vecteur de paramètre sélectionné
 Data<-modele.date(pheno,parametres)

#Sélection des colonnes qui nous intéresse
 Data<-Data[,c('Date','moyenne')]

#Transformation des données (asin(sqrt()))
 Data[,"moyenne"]<-asin(sqrt((Data[,"moyenne"])/100))

#Tirage aléatoire d'une erreur
 epsilon<-rnorm(1,0,sqrt(var.erreur))

#Translation de la logistique par epsilon
 Data[,"moyenne.aleatoire"]<-  Data[,"moyenne"]+ epsilon

#Sélection de la date où Sévérité > seuil
  if (epsilon >= seuil )
    {
    Date.traitement<-min(Data[,'Date'][Data[,'moyenne']>0])
    }else
    {
    Date.traitement<-min(Data[,'Date'][Data[,'moyenne.aleatoire']>=seuil])
    }
#Rassemblement des resultat dans un vecteur récapitulatif
 result<-c(Date.traitement,var.erreur,epsilon)


#Création d'un fichier .csv contenant les titres des colonnes qui nous
intéressent
```

```
 nomcol<-c('Date.traitement','Var.erreur','epsilon')
 write(nomcol,paste(DirOutput,nomcsv2,sep=""),ncolumns=length(nomcol))

#A la fin de chaque boucle, on rajoute les résultats à la suite les uns des
autres
#Exportation du résultat de la 1ère itération
 write(result,paste(DirOutput,nomcsv2,sep=""),
 append=T,ncolumns=length(result))

############################
#  LE RESTE DE LA BOUCLE   #
############################
for (i in 2:length(k[,1]))
  {
  #Sélection du ième vecteur de paramètres
  parametres<-as.numeric(k[i,1:17])
  names(parametres)<-c("coeff.mult","seuil.cumul","echelle","transf.pluie",
  "transf.haut","vit.nec","appar.nec","Tminjour","Tminveille","Pp",
  "prod.inoc","vit.expan","limite","min.lorin","Tlim.lorin","epuis.inoc"
  ,"limiteF1")

  #Sélection de la variance associée à ce vecteur de paramètres
  var.erreur<-k[i,c("variance.erreur")]

  #Sélection du i-1ème vecteur de paramètres
  para<-as.numeric(k[i-1,c(names(parametres))])
  names(para)<-c("coeff.mult","seuil.cumul","echelle","transf.pluie",
  "transf.haut","vit.nec","appar.nec","Tminjour","Tminveille","Pp",
  "prod.inoc","vit.expan","limite","min.lorin","Tlim.lorin","epuis.inoc"
  ,"limiteF1")

    #Raccourci: Si teta[i]=teta[i-1]
    #=>  Pas  besoin  de  refaire  tourner  le  modèle,  seul  la  variance
    #résiduelle change
    if (all(parametres==para))
      {
        #Tirage aléatoire d'une erreur
        epsilon<-rnorm(1,0,sqrt(var.erreur))

        #Translation de la logistique par epsilon
        Data[,"moyenne.aleatoire"]<- Data[,"moyenne"]+ epsilon

        #Sélection de la date où Sévérité > seuil
        if (epsilon >= seuil )
        {
        Date.traitement<-min(Data[,'Date'][Data[,'moyenne']>0])
        }else
        {
        Date.traitement<-
min(Data[,'Date'][Data[,'moyenne.aleatoire']>=seuil])
        }
        #Rassemblement des resultat dans un vecteur récapitulatif
        result<-c(Date.traitement,var.erreur,epsilon)

        #Exportation du résultat de la ième itération
        write(result,paste(DirOutput,nomcsv2,sep=""),
        append=T,ncolumns=length(result))
      }else
      {
        #Lancement du modèle avec le vecteur de paramètre sélectionné
        Data<-modele.date(pheno,parametres)
```

```
        #Sélection des colonnes qui nous intéresse
        Data<-Data[,c('Date','moyenne')]

        #Transformation des données (asin(sqrt()))
        Data[,"moyenne"]<-asin(sqrt((Data[,"moyenne"])/100))

        #Tirage aléatoire d'une erreur
        epsilon<-rnorm(1,0,sqrt(var.erreur))

        #Translation de la logistique par epsilon
        Data[,"moyenne.aleatoire"]<-  Data[,"moyenne"]+ epsilon

        #Sélection de la date où Sévérité > seuil
        if (epsilon >= seuil )
        {
        Date.traitement<-min(Data[,'Date'][Data[,'moyenne']>0])
        }else
        {
        Date.traitement<-
min(Data[,'Date'][Data[,'moyenne.aleatoire']>=seuil])
        }
        #Rassemblement des resultat dans un vecteur récapitulatif
        result<-c(Date.traitement,var.erreur,epsilon)

        #Exportation du résultat de la ième itération
        write(result,paste(DirOutput,nomcsv2,sep=""),
        append=T,ncolumns=length(result))
      }
  }
print("THAT'S ALL FOLKS")
}
```