Analyse statistique des risques agro-environnementaux

Études de cas

Springer

Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

David Makowski
Hervé Monod

Analyse statistique des risques agro-environnementaux
Études de cas

Springer

David Makowski
Directeur de Recherche INRA
UMR 211 INRA AgroParisTech
BP 01
78850 Thiverval-Grignon

Hervé Monod
Directeur de Recherche INRA
Unité MIAJ (UR341)
78352 Jouy-en-Josas Cedex

ISBN-13 : 978-2-8178-0250-3 Springer Paris Berlin Heidelberg New York

Imprimé en France

Springer-Verlag France est membre du groupe Springer Science + Business Media

Maquette de couverture : Jean-François Montmarché

Collection
Statistique et probabilités appliquées
dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
Suisse
yadolah.dodge@unine.ch

Comité éditorial :

Dans la même collection :

- *Statistique. La théorie et ses applications*
  Michel Lejeune, avril 2004

- *Optimisation appliquée*
  Yadolah Dodge, octobre 2004

- *Le choix bayésien. Principes et pratique*
  Christian P. Robert, novembre 2005

- *Régression. Théorie et applications*
  Pierre-André Cornillon, Éric Matzner-Løber, janvier 2007

- *Le raisonnement bayésien. Modélisation et inférence*
  Éric Parent, Jacques Bernier, juillet 2007

- *Premiers pas en simulation*
  Yadolah Dodge, Giuseppe Melfi, juin 2008

- *Génétique statistique*
  Stephan Morgenthaler, juillet 2008

- *Maîtriser l'aléatoire. Exercices résolus de probabilités et statistique, 2e édition*
  Eva Cantoni, Philippe Huber, Elvezio Ronchetti, septembre 2009

- *Pratique du calcul bayésien*
  Jean-Jacques Boreux, Éric Parent, décembre 2009

- *Statistique. La théorie et ses applications, 2e édition*
  Michel Lejeune, septembre 2010

- *Le logiciel R*
  Pierre Lafaye de Micheaux, Rémy Drouilhet, Benoît Liquet, novembre 2010

- *Probabilités et processus stochastiques*
  Yves Caumel, avril 2011

Avant-propos

Les ingénieurs et scientifiques sont souvent sollicités pour réaliser des analyses de risque dans le cadre d'expertises pour des agences gouvernementales. Ce type d'analyse comporte généralement trois étapes — l'évaluation du risque, sa gestion et la communication des résultats — qui nécessitent la mise en œuvre de méthodes statistiques et de modèles mathématiques. Dans cet ouvrage, nous nous intéressons à des risques d'un type particulier : les risques associés aux activités agricoles. Ces risques concernent à la fois les produits agricoles et l'environnement, et touchent aussi bien l'agriculteur que le consommateur ou le citoyen.

Cet ouvrage n'a pas la prétention d'être exhaustif mais il constitue une introduction aux principaux types de modèle et aux principales méthodes statistiques utiles pour l'analyse des risques agro-environnementaux. Le chapitre 1 est un chapitre introductif qui souligne le rôle que peuvent jouer les modèles et les méthodes statistiques pour l'analyse des risques agro-environnementaux. Le chapitre 2 présente des notions de base en modélisation et statistique. Le chapitre 3 décrit plusieurs types de modèle utiles pour l'évaluation du risque. Le chapitre 4 décrit des méthodes pour optimiser des décisions et gérer les risques, et le chapitre 5 présente une série de techniques pour l'analyse et la communication de l'incertitude.

Chaque type de modèle et chaque méthode statistique sont illustrés par une ou plusieurs applications. Les programmes informatiques utilisés pour développer les modèles et appliquer les méthodes statistiques sont présentés et commentés en détail. Ils ont tous été réalisés avec des logiciels librement téléchargeables. Les chapitres 2 à 5 incluent également des exercices.

Nous espérons que notre livre sera utile au plus grand nombre et qu'il encouragera les ingénieurs et scientifiques à s'intéresser à la modélisation et à la statistique. Dans ce but, nous avons souhaité que les applications décrites concernent des risques variés : risque de pollution de l'eau par les nitrates, invasion par des espèces nuisibles, diminution de la biodiversité, flux de gènes d'une culture OGM vers une culture non OGM, pertes de rendement et de qualité, émission de gaz à effet de serre, etc. Ces applications ont été réalisées au cours des 10 dernières années en collaboration avec des agronomes, biologistes et écologues appartenant à divers organismes, notamment à l'INRA et à AgroParisTech.

Nous les remercions très chaleureusement. Sans eux, cet ouvrage n'existerait pas.

Nous remercions tout particulièrement, par ordre alphabétique : Katarzyna Adamczyk, Bruno Andrieu, Frédérique Angevin, Aude Barbottin, Nicolas Beaudoin, Liliane Bel, Claire Cadet, Marion Casagrande, Bruno Chauvel, Jean-Jacques Daudin, Jean-Baptiste Denis, Thierry Doré, Sabah Ennaïfar, Robert Faivre, Benoît Gabrielle, Arnaud Gauffreteau, Martine Guérif, Laurence Guichard, Jonathan Hillier, Marie-Hélène Jeuffroy, Jim W. Jones, Anne Lacroix, Matieyendou Lamboni, François Laurent, Marc Lavielle, Marianne Le Bail, Chantal Loyce, Philippe Lucas, Alexandra Maltas, Antoine Messéan, Jean-Marc Meynard, Marie Morfin, Nicolas Munier-Jolain, Cédric Naud, Eric Parent, Annette Penaud, Aurore Philibert, Sophie Primot, Lorène Prost, Marie Taverne, Muriel Tichit, Laurent Ruck, Muriel Valantin-Morison, Valérie Viaud, Daniel Wallach.

Nous remercions également Yadolah Dodge, Stephan Morgenthaler et Charles Ruelle pour avoir soutenu et accompagné notre projet d'ouvrage dès ses premières versions.

Nous exprimons notre reconnaissance tout particulièrement à Stephan, dont les relectures détaillées et les nombreux conseils ont permis d'améliorer très fortement la qualité de l'ouvrage. [1]

David Makowski et Hervé Monod
INRA France

1. Cet ouvrage a été réalisé avec le style LaTeX « iris » développé par Nicolas Puech, avec l'aide et les suggestions de Patrik Fuhrer, Michel Lejeune, Bruno Petazzoni, Laurent Decreusefond et Philippe Martins.

Sommaire

Chapitre 1

Introduction

Les risques associés aux activités agricoles touchent aussi bien l'agriculteur que le consommateur ou le citoyen. Leurs conséquences sont diverses : pertes de rendement et de qualité, développement de maladie, pollution de l'eau par les nitrates, disparition d'espèces d'intérêt écologique, invasion par des espèces nuisibles, flux de gènes, érosion et bien d'autres encore. Comme ces exemples le montrent, ces risques concernent à la fois la production agricole, l'environnement et l'alimentation. Le rôle essentiel de l'agriculture dans l'économie et les préoccupations actuelles de nos sociétés pour l'alimentation et l'environnement rendent donc la question des risques agro-environnementaux incontournable.

De façon formelle, un risque peut être défini comme un événement futur qui, s'il se réalise, aura des conséquences perçues comme néfastes. Dans cet ouvrage, nous proposons d'appliquer aux risques agro-environnementaux la démarche standard d'analyse du risque, qui comporte trois étapes (Chevassus-au-Louis, 2007). La première étape est *l'évaluation du risque*. Elle consiste à quantifier la probabilité et l'impact de l'événement, en précisant le contexte de l'analyse. La deuxième est *la gestion du risque*, c'est-à-dire l'identification de techniques permettant de réduire le risque. La dernière est la *communication sur le risque* auprès des décideurs, des administrations et du grand public. La démarche d'analyse du risque permet aux scientifiques, ingénieurs et experts d'évaluer les probabilités et les impacts d'un risque donné, de proposer des règles de décision pour le maîtriser et, enfin, de communiquer sur ce risque et sur les mesures envisagées pour le contrôler, en tenant compte de l'incertitude associée aux résultats. Les spécialistes de l'analyse des risques sont amenés à utiliser différents types d'information dans leur travail : des données expérimentales, des avis d'experts, mais aussi des modèles mathématiques plus ou moins complexes (Bernier *et al.*, 2000 ; Vose, 2000).

L'analyse des risques agro-environnementaux présente plusieurs caractéristiques spécifiques. Une première caractéristique est que l'occurrence d'un risque agro-environnemental et son impact peuvent généralement être décrits de diverses manières, nécessitant la définition de plusieurs critères pour en mesurer

la gravité. Considérons par exemple le risque lié à l'occurrence d'une maladie fongique dans les cultures de blé d'une région donnée. Dans ce cas, l'étape d'évaluation du risque consiste à estimer d'une part la probabilité d'occurrence de cette maladie dans la région considérée et, d'autre part, l'impact de cette maladie sur la culture de blé et son environnement. Or plusieurs choix sont possibles pour mesurer l'occurrence de la maladie et son impact. L'occurrence peut être mesurée par un niveau d'incidence (pourcentage de plantes malades), un niveau de sévérité des symptômes (*e.g.*, pourcentage de surface de feuille nécrosée) ou une combinaison des deux. L'impact de la maladie peut également être mesuré de plusieurs manières : perte de rendement, diminution de la teneur en protéines des grains et donc de la qualité de la récolte de blé, perte économique, impact des traitements chimiques sur la qualité de l'eau et sur le risque d'apparition de souche résistante, etc.

Une deuxième caractéristique de l'analyse des risques agro-environnementaux est la nécessité de prendre en compte un ou plusieurs facteurs du milieu pour estimer la probabilité d'occurrence d'un risque, pour estimer son impact et pour évaluer l'efficacité de mesures permettant de gérer ce risque. Les facteurs du milieu correspondent à des caractéristiques du sol, du climat, de la culture ou à des techniques culturales. Ils ont une forte influence sur l'occurrence d'un risque et sur son impact. En général, il est donc nécessaire de tenir compte d'un ou plusieurs facteurs du milieu lors de l'analyse.

Enfin, une troisième caractéristique importante est que les informations disponibles sont souvent de nature et de qualité très diverses : données issues de mesures obtenues en parcelles expérimentales, données issues de mesures réalisées en parcelles d'agriculteurs, avis d'expert, simulations issues de modèles mathématiques plus ou moins complexes. L'hétérogénéité de ces informations rend souvent leur synthèse difficile.

La modélisation et les méthodes statistiques constituent des outils puissants pour les trois étapes de l'analyse du risque : l'évaluation, la gestion, la communication. Elles permettent d'analyser et de synthétiser l'ensemble des informations disponibles — données, avis d'expert, simulations — afin d'évaluer la probabilité d'occurrence et l'impact potentiel d'un risque donné. Elles peuvent également servir à optimiser des règles de décision en vue de réduire la probabilité d'occurrence ou l'impact d'un risque. Enfin, certaines méthodes statistiques sont utiles pour déterminer les niveaux d'incertitude associés aux estimations réalisées à partir des données, des avis d'expert et des modèles. Ces incertitudes peuvent alors être communiquées auprès des non-spécialistes.

L'objectif de cet ouvrage est de présenter des méthodes pour aider les agronomes, les biologistes et les spécialistes de l'environnement à analyser les risques agro-environnementaux. Nous souhaitons tout particulièrement sensibiliser le lecteur à l'intérêt de la démarche statistique pour :

– modéliser les risques en fonction de facteurs environnementaux et anthropiques, en s'appuyant sur des modèles adaptés à diverses situations (modèle linéaire, modèle linéaire généralisé, modèle non linéaire, modèle

hiérarchique, régression quantile) ;
- optimiser des décisions ou des règles de décision pour mieux gérer les risques, en intégrant des variables décisionnelles dans les modèles (optimisation de seuils de décision, optimisation par simulation, analyses ROC) ;
- gérer et communiquer les incertitudes associées aux modèles et aux stratégies proposées (estimation et description de distributions de probabilité, assimilation de données, analyses d'incertitude et de sensibilité).

La modélisation, la conception et l'évaluation de règles de décision, et l'analyse d'incertitude sont traitées successivement dans trois chapitres qui peuvent être consultés indépendamment en fonction des centres d'intérêt du lecteur. Chaque chapitre présente une série de méthodes permettant de traiter le problème considéré, des programmes informatiques pour appliquer ces méthodes, ainsi qu'une série d'exemples illustrant leur utilisation dans le cadre de l'analyse des risques agro-environnementaux. Les cas concrets présentés dans ce livre concernent essentiellement les grandes cultures, mais les techniques proposées peuvent être utilisées pour traiter une large gamme de problèmes liés aux activités agricoles.

Nous présentons dans cet ouvrage des méthodes issues aussi bien de la statistique fréquentiste que de la statistique bayésienne. En statistique fréquentiste, les paramètres des modèles sont définis comme des valeurs fixées mais inconnues, que l'on estime à partir de données observées. En statistique bayésienne, les paramètres sont décrits par des variables aléatoires dont la loi de probabilité *a priori* reflète l'état des connaissances avant l'analyse des données. La loi *a posteriori* résulte de la combinaison de cette loi *a priori* et de l'information apportée par les données. Dans bien des cas, un problème donné peut être traité avec les deux approches, fréquentiste et bayésienne. D'un point de vue pratique, les méthodes fréquentistes ont longtemps été plus faciles à mettre en œuvre, mais le développement récent de logiciels permettant d'appliquer des techniques bayésiennes a changé la donne. Les progrès réalisés en statistique et en informatique font que, dans beaucoup de situations, il est dorénavant aussi facile d'utiliser les méthodes bayésiennes que les méthodes fréquentistes. Nous présentons le plus souvent possible les deux approches et encourageons le lecteur à les comparer sur ses propres bases de données.

L'objectif de ce livre n'est pas de décrire de manière exhaustive toutes les formes de modélisation et toutes les méthodes statistiques existantes, mais plutôt de fournir au lecteur des notions de base et de nombreux exemples d'application qui lui permettront de maîtriser rapidement les méthodes les plus simples, avant d'aller plus loin. Nous avons ainsi volontairement limité l'usage de notations mathématiques complexes de façon à ne pas décourager les lecteurs peu expérimentés. Les concepts théoriques ne sont que brièvement traités et nous encourageons les lecteurs intéressés par ces concepts à se reporter aux ouvrages qui leur sont explicitement dédiés. Parmi les nombreux ouvrages à considérer, citons en particulier Daudin, Robin et Vuillet (1999) et Lejeune (2010) pour une introduction à la statistique inférentielle, et Saporta (2006) pour une présentation complète et pédagogique de nombreuses méthodes sta-

tistiques. Azaïs et Bardet (2006) offrent un traitement approfondi du modèle linéaire et de ses extensions. Robert (2006), Parent et Bernier (2007), Boreux *et al.* (2010) présentent les méthodes bayésiennes de façon détaillée. Des références bibliographiques plus spécifiques sont proposées dans chaque chapitre.

Les méthodes présentées dans les chapitres suivants peuvent être mises en œuvre avec des logiciels gratuits et téléchargeables depuis internet. La plupart de nos applications ont ainsi été réalisées avec les logiciels R et WinBUGS. Le logiciel R est téléchargeable depuis le site

`http://cran.r-project.org/`.

Son utilisation est décrite, par exemple, par Lafaye de Micheaux *et al.* (2011). Ce logiciel inclut un langage de programmation (Chambers, 2008) et des algorithmes permettant d'appliquer des méthodes statistiques fréquentistes et bayésiennes. Il possède également de nombreuses fonctionnalités graphiques. Le second logiciel est WinBUGS (Lunn *et al.*, 2000), téléchargeable depuis le site

`http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`.

WinBUGS est un logiciel exclusivement bayésien qui inclut un langage de modélisation spécifique. Les programmes R et WinBUGS utilisés dans nos exemples sont présentés au fil du texte.

Chapitre 2

Notions de base

2.1 Variables aléatoires et lois de probabilité

Le concept de probabilité offre un cadre et des outils très puissants pour modéliser des variables aléatoires, et il est à la base de toute approche statistique. Ce chapitre rappelle les principales notions et définitions utiles pour la suite.

2.1.1 Variable aléatoire

Une variable aléatoire est une fonction définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Les résultats d'un lancer de dés, d'un tirage à pile ou face ou d'un tiercé sont des variables aléatoires dont on n'observe les valeurs précises qu'à la fin de l'action. En sciences expérimentales, les variables aléatoires servent à décrire des mesures prises sur un échantillon, par exemple un échantillon de plantes. Une mesure est en effet réalisée dans des conditions particulières (climat, choix des échantillons, etc.) susceptibles de changer et non complètement contrôlées. La variabilité de ces conditions et leurs effets sur le résultat de la mesure sont donc modélisés à l'aide de variables aléatoires.

Le résultat de l'observation d'une variable aléatoire est appelée sa *réalisation*. Nous noterons dans la suite X la variable aléatoire et x sa réalisation éventuelle. L'ensemble des réalisations possibles x est appelé le *support* de X. Selon la nature du support, on distingue deux grands types de variables aléatoires, les variables *discrètes* et les variables *continues*. Les variables discrètes ont pour support un ensemble fini ou dénombrable, par exemple $\{0, 1, 2, 3, 4\}$ ou $\mathbb{N}$ (entiers naturels), tandis que les variables continues ont pour support un intervalle de l'ensemble des nombres réels $\mathbb{R}$.

2.1.2 Variables aléatoires discrètes

Pour une variable aléatoire discrète X de support $\{x_1, x_2, \ldots\}$, toutes les propriétés probabilistes sont déterminées par les probabilités $\pi_i = P(X = x_i)$ qu'une réalisation de X prenne la valeur x_i. Les probabilités π_i sont toutes positives et leur somme est égale à un : $\pi_1 + \pi_2 + \cdots = 1$, ce qui exprime le fait que toutes les valeurs possibles sont listées. Dans des situations telles que le lancer de dé, la probabilité π_i peut s'interpréter comme la fréquence théorique de la réalisation x_i si l'on répétait l'expérience (le lancer) indéfiniment. Dans d'autres cas, elle est définie hors de toute référence fréquentiste, comme une mesure objective ou subjective des chances que la valeur inconnue de X soit ou devienne égale à x_i.

La fonction $p(x) = P(X = x)$ définie sur le support de X est appelée la *loi de probabilité* de la variable aléatoire discrète X. Certaines lois de probabilité discrètes sont bien connues. La variable aléatoire la plus simple ne prend que deux valeurs, 0 ou 1 par exemple. Elle suit la *loi de Bernoulli* qui est définie par

$$P(X = 1) = \pi \text{ et } P(X = 0) = 1 - \pi$$

où π est la probabilité de succès, comprise entre 0 et 1. La *loi binomiale* de paramètre π et d'effectif n s'applique, par exemple au nombre de piles obtenu après n lancers à pile ou face. Plus généralement, elle décrit les probabilités du nombre de succès parmi n réalisations indépendantes d'une variable qui suit une loi de Bernoulli de probabilité π. La *loi de Poisson*, définie sur l'ensemble des entiers naturels $\mathbb{N}$, est souvent utilisée pour modéliser des données de dénombrement, telles que le nombre d'insectes dans une zone cultivée. Sa définition fait intervenir un paramètre d'intensité noté λ.

Le programme R ci-dessous permet de calculer et représenter la loi binomiale d'une part, la loi de Poisson d'autre part (figure 2.1). Les probabilités de la loi binomiale sont obtenues par la fonction `dbinom()`. Son premier argument est le vecteur des valeurs dont on veut calculer la probabilité, et les arguments suivants sont la probabilité π et l'effectif n. Les probabilités de la loi de Poisson sont obtenues par la fonction `dpois()`, dont le premier argument est le vecteur des valeurs dont on veut calculer la probabilité et le second argument est le paramètre λ.

```
#Code R

par(mfrow=c(1,2)) # divise la fenetre graphique en deux parties

# 1- Loi binomiale

probasBinomiale <- dbinom(x=0:10, prob=0.5, size=10)
plot(0:10, probasBinomiale, xlab="x", ylab="p(x)")

# 2- Loi de Poisson (graphique limite a x=0, ... , 15)
```

```
# de parametre lambda=1
plot(0:15, dpois(x=0:15, lambda=1), xlab="x", ylab="p(x)")

# de parametre lambda=5
points(0:15, dpois(x=0:15, lambda=5), pch=2)
```

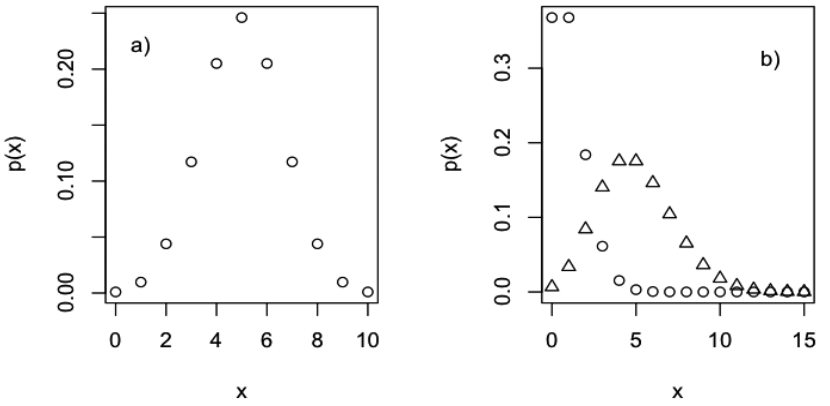

Figure 2.1 – Lois de probabilité de deux variables aléatoires discrètes : (a) loi binomiale de probabilité $\pi = 0.5$ et d'effectif $N = 10$; (b) lois de Poisson de paramètre $\lambda = 1$ (ronds) et $\lambda = 5$ (triangles).

2.1.3 Variables aléatoires continues

Les variables aléatoires continues ont pour support un intervalle continu et les réalisations possibles sont donc en nombre infini et non dénombrables. Elles peuvent être non bornées ou bornées, par exemple restreintes aux valeurs positives ou restreintes à des valeurs comprises entre une borne inférieure et une borne supérieure.

Si X est une variable aléatoire continue, la probabilité que X soit inférieure ou égale à une valeur donnée x est décrite par la *fonction de répartition*, définie par $F(x) = P(X \leq x)$. La fonction de répartition contient de façon concise toute l'information sur la loi de probabilité. En effet, elle permet de calculer la probabilité que X soit comprise dans un intervalle donné. Pour ce type de variable aléatoire, la probabilité que X prenne exactement une valeur réelle a ou b du support est mathématiquement nulle lorsque la fonction de répartition est continue. On a alors

$$F(b) - F(a) = P\left(a < X \leq b\right) = P\left(a < X < b\right) = P\left(a \leq X \leq b\right).$$

Pour la plupart des lois continues, la fonction de répartition est dérivable et on en déduit une autre fonction, appelée *densité de probabilité*. La densité $f(x)$ est une fonction non négative définie sur le support de la variable aléatoire. À partir de la densité, on peut retrouver $P(a < X \leq b)$ par intégration :

$$P\left(a < X \leq b\right) = \int_a^b f(x)dx.$$

C'est la fonction de densité $f(x)$ qui reflète le mieux le poids associé par la loi de probabilité aux différents points de support d'une variable aléatoire continue. L'usage que nous suivrons par la suite est de désigner par « densité » et de noter $p_X(x)$, ou plus simplement $p(x)$, soit la loi de probabilité $P(X = x)$ si X est discrète, soit la fonction de densité $f(x)$ si X est continue. Notez que dans les deux cas, la somme ou l'intégrale de $p(x)$ sur le support de X est égale à un.

La *loi uniforme* sur l'intervalle $[a, b]$ est une loi de probabilité continue définie par une densité de probabilité constante lorsque $a \leq X \leq b$ et nulle sinon. Les valeurs des bornes a et b sont les seuls paramètres de la loi uniforme. La loi de probabilité la plus utilisée pour une variable aléatoire continue est bien connue pour la forme en cloche de sa fonction de densité. Il s'agit de la *loi gaussienne* ou *loi normale*, caractérisée par sa moyenne μ et sa variance σ^2 (ou son écart type σ). Elle bénéficie de nombreuses propriétés intéressantes, qui expliquent en grande partie son succès. Plusieurs lois classiques sont dérivées de la loi normale : si X suit une loi normale, alors $\log(X)$ suit une *loi log-normale* ; si $X_1, \ldots, X_n$ sont indépendantes et suivent la loi normale de moyenne $\mu = 0$ et de variance σ^2=1, alors $X_1^2 + \ldots + X_n^2$ suit la *loi du Khi-deux* centrée à n degrés de liberté. La variable aléatoire $X/\sqrt{S/n}$, où X suit une loi normale centrée réduite et S suit une loi du Khi-deux à n degrés de liberté, indépendante de X, suit *une loi de Student* à n degrés de liberté. La variable aléatoire S_1/S_2, où S_1 et S_2 suivent des lois du Khi-deux indépendantes ayant les mêmes degrés de liberté, suit une *loi de Fisher*.

Les lois de probabilité que nous venons de citer se rencontrent fréquemment en statistique, mais il en existe encore bien d'autres : *loi beta, loi gamma, loi exponentielle,* etc. Les propriétés de ces lois sont décrites dans de nombreux ouvrages ou sur internet. Pour un grand nombre d'entre elles, le logiciel R permet de calculer la fonction de répartition et la fonction de densité. Le programme R ci-dessous calcule les fonctions de répartition et les densités de la loi normale et de *la loi de Cauchy*. Contrairement à la loi normale, la loi de Cauchy accorde des probabilités importantes à des valeurs extrêmes de très grande amplitude.

Le programme ci-dessous génère également un échantillon représentatif de ces lois puis en trace les représentations graphiques (figure 2.2). La génération de nombres aléatoires selon des lois de distribution précises est extrêmement utile en statistique. Elle permet en effet de conduire des simulations informatiques pour analyser le comportement d'un phénomène à travers un modèle ou pour étudier les propriétés d'un estimateur (voir Dodge et Melfi, 2008 ; Robert et Casella, 2011). Nous utiliserons la simulation à plusieurs reprises par la suite.

```
# Code R

par(mfrow=c(3,2))  # pour diviser la fenetre graphique en
                   # 3 lignes et 2 colonnes

set.seed(123)  # pour controler le germe du generateur
               # de nombres aleatoires

help(Normal)  # aide sur les fonctions sur la loi normale
help(Cauchy)  # aide sur les fonctions sur la loi de Cauchy

# Fonctions de repartition

x <- seq(-10,10,length=100)

plot(x, pnorm(x,mean=0,sd=1), type="l", xlab="x",ylab="F(x)",
                main="Fonction de repartition")
lines(x, pcauchy(x, location=0, scale=1), type="l", lty=2)

# Fonctions de densite

plot(x,dnorm(x, mean=0, sd=1), type="l", xlab="x", ylab="f(x)",
                main="Densite")
lines(x, dcauchy(x, location=0, scale=1), type="l", lty=2)

# Echantillonnage

sampleNorm <- rnorm(50, mean=0, sd=1)
sampleCauchy <- rcauchy(50, location=0, scale=1)

plot(sampleNorm, xlab="Indice de repetition",ylab="x",
                 main="Loi normale")
abline(h=0) # trace un trait horizontal d'ordonnee 0

plot(sampleCauchy, xlab="Indice de repetition",ylab="x",
                   main="Loi de Cauchy")
abline(h=0)

hist(sampleNorm,   xlab="x", main="")
hist(sampleCauchy, xlab="x", main="")
```

Les fonctions de répartition de la loi normale centrée-réduite ($\mu = 0, \sigma^2 = 1$) et de la loi de Cauchy standard sont calculées avec les fonctions R `pnorm()` et `pcauchy()`. Les fonctions de densité sont calculées avec les fonctions `dnorm()` et `dcauchy()`. Pour chacune des deux lois, un échantillon aléatoire de taille 50

est simulé avec les fonctions `rnorm()` et `rcauchy()`. Il est ensuite représenté dans un diagramme de dispersion puis sous la forme d'un histogramme. Les différents graphiques montrent le comportement très différent des deux lois pour les queues de distribution et la répartition des valeurs extrêmes.

2.1.4 Valeurs caractéristiques d'une variable aléatoire

Les lois de probabilité de variables aléatoires X quantitatives sont caractérisées par certaines valeurs typiques de la distribution, notamment l'*espérance*, la *médiane* et le *mode*. L'espérance de X est la valeur moyenne de cette variable, pondérée par la densité $p(x)$. Elle est définie par $E(X) = \sum_i x_i P(X = x_i)$ pour une variable X discrète et par $E(X) = \int xf(x)\mathrm{d}x$ pour une variable X continue. La médiane de X est la valeur qui coupe au mieux le support en deux sous-ensembles de même probabilité 0.5. Le mode est la valeur la plus probable de X, c'est-à-dire celle qui maximise $p(x)$.

Une autre caractéristique importante d'une loi de probabilité est sa variance. Il s'agit d'une mesure de la dispersion de la variable aléatoire X définie par $\mathrm{Var}(X) = E\Big(\big(X - E(X)\big)^2\Big)$. L'écart type de X est égal par définition à la racine carrée de la variance de X : $\sigma(X) = \sqrt{\mathrm{Var}\,(X)}$.

Le tableau 2.1 présente les paramètres, la densité, l'espérance et la variance de quelques lois usuelles.

Loi	Support	Densité	Espérance	Variance
Bernoulli(π)	$x \in \{0, 1\}$	$\pi^x(1-\pi)^{1-x}$	π	π(1-π)
Binomiale(n, π)	$x \in \{0, 1, ..., n\}$	$\binom{n}{x}\pi^x(1-\pi)^{1-x}$	$n\pi$	$n\pi(1-\pi)$
Poisson(λ)	$x \in \mathbb{N}$	$\lambda^x \exp(-\lambda)/x!$	λ	λ
Normale(μ, σ)	$x \in \mathbb{R}$	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2

Tableau 2.1 – Fonctions densité, espérance et variance de quelques lois classiques.

La fonction quantile $q(u)$ de la variable aléatoire X, définie pour tout u appartenant à $]0, 1[$, est l'inverse de la fonction de répartition $F(x)$. Plus précisément, le quantile de niveau u d'une variable aléatoire X, pour $0 < u < 1$, est défini comme la plus petite valeur $x = q(u)$ telle que $F(x) \geq u$. En particulier, la médiane M, définie comme le quantile de niveau 0.5, vérifie $P(X \leq M) \geq 0.5$ et $P(X \geq M) \geq 0.5$. C'est une mesure du comportement moyen de la variable aléatoire X, moins sensible aux valeurs extrêmes que la moyenne. Les termes « quartile » et « décile » sont utilisés pour désigner des quantiles associés à des niveaux u multiples de 0.25 ou 0.10 respectivement. Les quantiles sont utilisés dans les tests statistiques pour définir des valeurs seuils permettant de trancher entre deux hypothèses. Ils permettent également de quantifier les valeurs que risque d'atteindre une variable aléatoire avec une probabilité donnée.

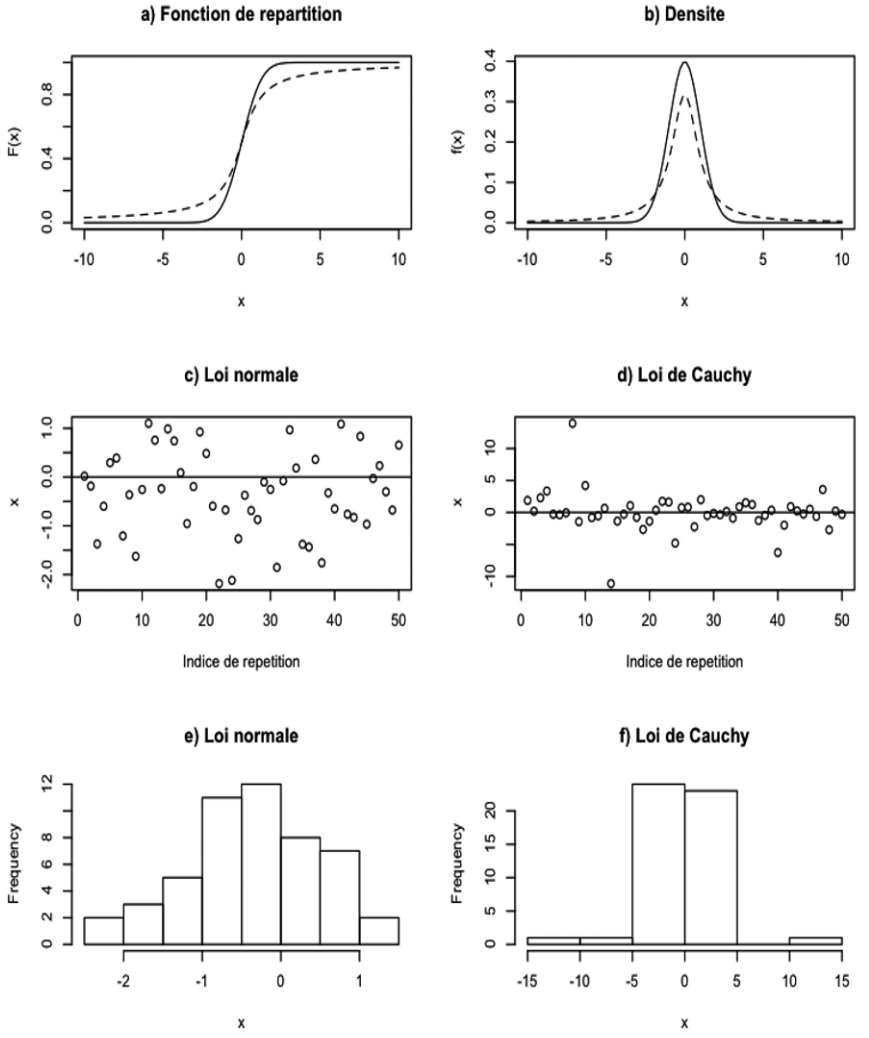

Figure 2.2 – Lois de probabilité de variables aléatoires continues : (a) fonctions de répartition de la loi normale de moyenne nulle et d'écart type 1 (trait continu) et de la loi de Cauchy centrée en 0 et de paramètre d'échelle 1 (tirets) ; (b) fonctions de densité des deux mêmes lois ; (c) et (d) graphiques de dispersion de 50 réalisations dans la loi normale et dans la loi de Cauchy, respectivement ; (e) et (f) histogrammes des 50 réalisations obtenues.

Leur usage en régression sera discuté dans la section 3.5. Sous R, les quantiles peuvent être calculés par des fonctions telles que `qnorm()` pour la loi normale ou plus généralement `q`*nomLoi*`()` pour une loi intitulée *nomLoi* sous R.

2.1.5 Dépendance entre variables aléatoires

La figure 2.3 présente 100 réalisations de deux variables aléatoires X et Y. Ces valeurs ont été générées par simulation mais elles pourraient être, par exemple, des mesures de la masse et de la taille sur un échantillon de 100 plantes. Sur cet exemple, on constate une liaison entre les réalisations de Y et celles de X. Si la réalisation de X est proche de zéro, celle de Y l'est aussi. Si la réalisation de X est proche de 20, celle de Y est également proche de 20. Le code R ayant permis de générer la figure 2.3 est présenté ci-dessous. Il utilise la fonction `runif()` pour générer les valeurs de X selon la loi uniforme, puis une expression dépendant de X et de la loi normale pour générer les valeurs de Y.

```
#Code R

set.seed(4217)  # pour controler le germe du generateur de
                # nombres aleatoires

x <- runif(100,min=0,max=30)
y <- 5*sqrt(x) + sqrt(x)*rnorm(100)

par(mfrow=c(1,3))  # pour diviser la fenetre graphique en
                   # 3 colonnes

hist(x)
hist(y)
plot(x,y)
```

Cet exemple illustre les liaisons ou *relations de dépendance* qui peuvent exister entre deux variables aléatoires. Il montre la nécessité de définir des lois de probabilité pour décrire de telles variables de façon conjointe. En fait, plusieurs lois de probabilité sont associées à un couple de variables aléatoires X et Y :

- la *loi conjointe* du couple (X, Y), dont la densité $p_{XY}(x, y)$, ou plus simplement $p(x, y)$, s'applique aux couples de valeurs (x, y) ; le diagramme de dispersion de la figure 2.3 représente un échantillon issu de la loi conjointe de X et Y ;
- les *lois marginales*, qui correspondent aux probabilités de X et Y considérées individuellement ; leurs densités se déduisent de celle de la loi conjointe par $p(x) = \int p(x, y)\mathrm{d}y$ et $p(y) = \int p(x, y)\mathrm{d}x$; les histogrammes de la figure 2.3 en représentent les approximations basées sur l'échantillon de taille 100 ;

– les *lois conditionnelles*, qui correspondent aux probabilités de X sous certaines contraintes sur Y ou vice-versa ; par exemple, $p(x \mid Y > 15)$ désigne la densité de la loi de X conditionnelle à $Y > 15$, soit mathématiquement

$$p(x \mid Y > 15) = \frac{\int_{15}^{+\infty} p(x, y)\mathrm{d}y}{\int_{-\infty}^{+\infty}\int_{15}^{+\infty} p(x, y)\mathrm{d}y\mathrm{d}x};$$

de façon générale, on désigne par $p(x \mid y)$ la loi de X conditionnelle à l'événement $Y = y$.

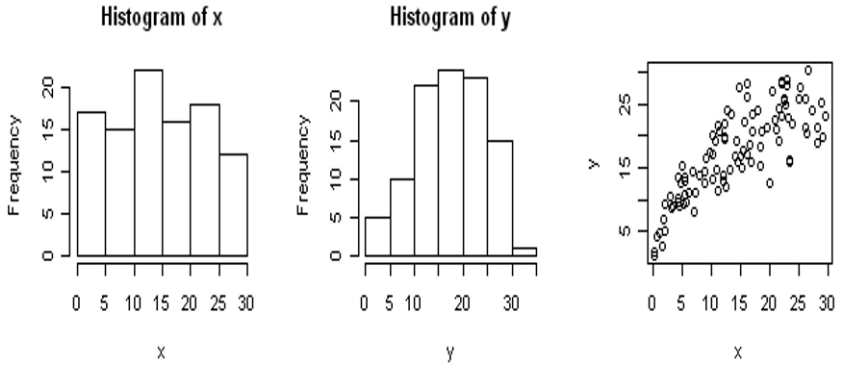

Figure 2.3 – Histogrammes et diagramme de dispersion d'un échantillon de taille 100 de deux variables aléatoires X et Y dépendantes.

La covariance et la corrélation sont utilisées pour mesurer le lien entre deux variables aléatoires X et Y. La covariance de X et Y est définie par

$$\mathrm{Cov}(X, Y) = E\Big(\big(X - E(X)\big)\big(Y - E(Y)\big)\Big)$$

et la corrélation linéaire de X et Y se déduit de la covariance à l'aide de la relation suivante :

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

Deux variables aléatoires sont indépendantes s'il n'y a aucune forme de liaison probabiliste entre elles (la densité de probabilité de la loi conjointe est égale au produit des densités des lois marginales). Elles ont alors une corrélation linéaire nulle, mais une corrélation nulle ne suffit pas à démontrer l'indépendance, qui est une propriété plus forte. Dans la figure 2.3 par exemple, la corrélation linéaire correspond à l'inclinaison moyenne du nuage de points. Cependant, la liaison entre X et Y ne se réduit pas à cette inclinaison moyenne. En effet, la relation entre les valeurs de X et Y est non linéaire. De plus, la variabilité de X est fortement liée à la valeur prise par Y et réciproquement.

Deux propriétés des variables aléatoires multivariées méritent d'être soulignées :

- la fonction de densité conjointe de variables aléatoires indépendantes est égale au produit des fonctions de densité des lois marginales : on a $p(x, y) = p(x)p(y)$ si et seulement si X et Y sont indépendantes ; dans ce cas, on a également $p(x \mid y) = p(x)$ et $p(y \mid x) = p(y)$;
- la loi multinormale est la généralisation multidimensionnelle de la loi normale. Comme la loi normale dans le cas univarié, elle vérifie de nombreuses propriétés intéressantes ; en particulier, les lois conditionnelles $p(x \mid y)$ de variables multinormales sont elles-mêmes normales (ou multinormales dans un cadre plus général).

Toutes ces notions se généralisent à plus de deux variables. On peut ainsi définir des lois conjointes sur des vecteurs (ou des matrices) de variables aléatoires, également appelées lois multidimensionnelles ou multivariées. L'espérance d'une loi multivariée est le vecteur composé des espérances marginales. La variance d'une loi multivariée est décrite par sa matrice de variance-covariance, dont la diagonale contient les variances marginales et les éléments non diagonaux contiennent les covariances entre variables distinctes.

2.2 La notion de modèle en statistique

2.2.1 Description

Par « modèle », nous entendons dans cet ouvrage un modèle mathématique, défini comme une représentation mathématique du fonctionnement d'un système. Parmi la diversité des modèles mathématiques existants (*e.g.*, Pavé, 1994), les modèles statistiques tiennent une place particulière. Une de leurs caractéristiques importantes est qu'ils incluent à la fois des éléments observables (les variables mesurées) et des éléments non observables (les paramètres et parfois certaines variables dites « cachées »). Par ailleurs, certains de ces éléments sont des variables aléatoires définies par des lois de probabilité. Les modèles statistiques sont ainsi des modèles stochastiques.

Illustrons ces différents aspects avec le modèle de régression linéaire d'ordonnée nulle à l'origine, défini par :

$$Y = \alpha X + \varepsilon \tag{2.1}$$

Dans cette expression, les deux variables observées sont Y, qui désigne la variable réponse à modéliser, et X, qui désigne l'unique variable d'entrée. Le modèle relie Y et X par une fonction linéaire incluant un coefficient α inconnu. Ce coefficient est un paramètre du modèle. Le terme $\varepsilon = Y - \alpha X$ représente l'écart entre la réponse observée Y et la valeur calculée par le modèle en fonction de la variable d'entrée. Ce terme est appelé erreur du modèle. Il est souvent

interprété comme une erreur de mesure de la variable réponse. En statistique, l'erreur ε est définie comme une variable aléatoire. Une loi de probabilité lui est associée, par exemple une loi gaussienne d'espérance nulle et de variance σ^2 inconnue. Comme Y dépend de ε, Y est également une variable aléatoire. Dans le modèle classique de régression linéaire, X n'est pas considérée comme une variable aléatoire.

Pour résumer, le modèle comprend une variable réponse (ou variable observée) aléatoire Y, une variable explicative fixe et connue X, une variable d'erreur aléatoire ε et deux paramètres : α et σ^2.

Le statut précis des paramètres diffère entre les approches fréquentiste et bayésienne. Dans le cadre de la *statistique fréquentiste,* un paramètre est simplement défini comme une constante inconnue qui doit être estimée à partir d'une série de mesures de Y. Dans le cadre de la *statistique bayésienne,* il lui est associé une variable aléatoire dont la distribution *a priori* reflète l'état de connaissance avant l'exploitation des données. Pour que le modèle soit complet, il faut donc aussi spécifier les lois *a priori* des paramètres. Nous reparlerons de ces deux approches dans le paragraphe 2.3.1 mais il est d'abord utile de définir la notion de vraisemblance.

2.2.2 Fonction de vraisemblance d'un modèle statistique

Le modèle statistique détermine la fonction de densité de probabilité de la variable observée Y conditionnellement aux valeurs des paramètres, notée $p(y \mid \theta)$, où y désigne une réalisation de Y et θ le vecteur des paramètres. Appliquée aux données observées, cette fonction joue un rôle important dans la phase d'inférence. En statistique fréquentiste, on l'utilise non pas comme une fonction de y mais comme une fonction des paramètres, avec y et les variables d'entrée fixées à leurs valeurs observées. Pour distinguer les deux usages, on appelle cette fonction des paramètres du modèle la fonction de vraisemblance (*e.g.*, Saporta, 2006).

Dans l'exemple du modèle de régression (2.1), les observations sont des paires de valeurs (x_i, y_i), avec $i = 1, \ldots, N$ où N est le nombre d'observations. La fonction de vraisemblance d'une variable aléatoire Y qui suit une loi normale de moyenne μ et d'écart type σ est définie par $V_0(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$, où y représente une observation, c'est-à-dire une réalisation de Y. La vraisemblance associée à N observations $y_1, \ldots, y_N$ indépendantes est le produit des vraisemblances associées à chaque observation. La vraisemblance globale dans l'exemple de la régression linéaire est donc

$$V(\alpha, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^N} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \alpha x_i)^2\right).$$

Le membre de droite désigne aussi $p(\mathbf{y} \mid \alpha, \sigma)$, c'est-à-dire la densité du vecteur

d'observations $\mathbf{y} = (y_1, \ldots, y_N)$ conditionnelle aux paramètres. Dans le cadre bayésien, c'est cette interprétation qui est privilégiée.

2.3 Inférence statistique

2.3.1 Approche fréquentiste et approche bayésienne

L'inférence statistique désigne l'élaboration d'informations précises sur les liens entre les variables modélisées (valeurs des paramètres du modèle, par exemple), à partir de la connaissance très partielle qu'en offrent les données disponibles. Comme nous l'avons déjà évoqué, il existe deux principales approches en statistique inférentielle, l'approche fréquentiste (souvent appelée « classique ») et l'approche bayésienne. Les deux approches ont de nombreux points communs. Dans les deux cas, on utilise des modèles pour décrire la relation entre des variables réponses, des variables d'entrée et des paramètres. Dans les deux cas, les paramètres sont inconnus et on souhaite en préciser la valeur à l'aide d'observations. Dans les deux cas enfin, les observations sont supposées issues d'un processus aléatoire dépendant des paramètres et elles apportent donc sur eux une information qu'il s'agit d'exploiter. Mais les deux approches diffèrent sur le statut précis accordé aux paramètres et sur la façon d'exploiter les données pour en préciser les valeurs. Ainsi, les notions de test statistique et d'intervalle de confiance sont des notions issues de la statistique fréquentiste qui n'ont pas de réels équivalents en statistique bayésienne.

Les partisans des deux approches se sont longtemps opposés dans des débats parfois très vigoureux, chacun mettant en cause la légitimité de l'approche adverse. Aujourd'hui, de nombreux statisticiens adoptent un point de vue pragmatique et considèrent que les deux approches font partie de la boîte à outils du praticien de la statistique. C'est le point de vue adopté dans cet ouvrage.

Selon l'approche fréquentiste, les paramètres des modèles ont de vraies valeurs et on doit raisonner conditionnellement à ces vraies valeurs, même si elles sont inconnues. Ainsi, dans l'exemple du modèle (2.1) de régression, ε et Y sont aléatoires, mais α est fixe. Les propriétés des estimateurs et des tests sont étudiées en imaginant leur comportement lorsqu'on répète les observations avec des paramètres fixés à leurs vraies valeurs.

Dans l'approche bayésienne (*e.g.* Robert, 2006 ; Parent et Bernier, 2007 ; Carlin et Louis, 2008 ; Boreux *et al.*, 2010), les paramètres aussi bien que les observations sont représentés par des variables aléatoires. Dans l'exemple du modèle de régression d'équation (2.1), les termes ε, Y, α sont donc tous aléatoires. On ne raisonne pas en fonction d'une vraie valeur des paramètres, mais en fonction de distributions de probabilité qui décrivent le niveau d'incertitude dans les valeurs de ces paramètres et qui synthétisent l'information apportée par l'expertise du modélisateur (à travers la distribution *a priori*)et par les observations (par l'intermédiaire de la vraisemblance). Dans l'exemple, la distribution *a priori*, de densité $p(\alpha, \sigma)$, résume l'état de connaissance dans les

valeurs des paramètres avant l'acquisition des données. Cette distribution *a priori* doit être définie par le modélisateur indépendamment des données disponibles, à partir d'articles scientifiques, de pré-études ou de dire d'experts, par exemple, ou par défaut en ayant recours à des lois de probabilité dites non informatives.La distribution *a posteriori*, de densité notée $p(\alpha, \sigma \mid \mathbf{y})$, résulte de la combinaison de la distribution *a priori* et des données expérimentales acquises pour estimer les paramètres. Elle est reliée à la distribution *a priori* et à la vraisemblance par le théorème de Bayes :

$$p(\alpha, \sigma \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \alpha, \sigma)\, p(\alpha, \sigma)}{p(\mathbf{y})}.$$

Très souvent, on utilise des distributions *a priori* indépendantes, ce qui entraîne que $p(\alpha, \sigma) = p(\alpha)p(\sigma)$. Par contre, les distributions *a posteriori* des paramètres sont généralement non indépendantes.

2.3.2 Estimateur

Pour estimer les valeurs des paramètres (α, σ dans l'exemple de régression linéaire) ou réaliser des tests sur ces valeurs, on utilise des fonctions des observations. Une fonction reliant les observations aux valeurs estimées du ou des paramètres du modèle est appelée un *estimateur*. L'estimateur du maximum de *vraisemblance* est l'un des estimateurs les plus classiques en statistique fréquentiste. Lorsqu'il existe, il est égal par définition à la valeur du paramètre qui maximise la fonction de vraisemblance. Dans l'exemple de régression, l'estimateur du maximum de vraisemblance des paramètres (α, σ) est égal à la paire de valeurs qui maximise $V(\alpha, \sigma) = p(\mathbf{y} \mid \alpha, \sigma^2)$.

En statistique fréquentiste, on évalue la qualité d'un estimateur — qui est une variable aléatoire puisqu'il dépend des observations — en étudiant sa distribution de probabilité, c'est-à-dire le comportement qu'il aurait si l'on répétait la procédure d'observation et d'estimation un très grand nombre de fois. On s'intéresse en particulier aux propriétés suivantes :

- un estimateur est dit *sans biais* si son espérance (la moyenne de toutes les valeurs de l'estimateur obtenues en répétant la même procédure d'observation et d'estimation indéfiniment) est égale à la vraie valeur du paramètre, quelle que soit cette valeur ;
- la convergence est une autre qualité importante pour un estimateur ; un estimateur d'un paramètre est *convergent* si sa valeur tend vers la vraie valeur de ce paramètre lorsque le nombre d'observations tend vers l'infini ;
- la précision d'un estimateur est souvent mesurée par sa variance (ou son écart type) sur les répétitions virtuelles de la procédure d'observation et d'estimation ; plus la variance est grande, moins l'estimateur est précis ; une variance élevée (ou un écart type élevé) indique qu'une autre série de données expérimentales acquises dans les mêmes conditions aurait pu conduire à une valeur estimée très différente ;

– l'écart quadratique moyen d'un estimateur est défini comme l'espérance de l'écart au carré entre l'estimateur et la vraie valeur du paramètre ; il est égal à la somme de la variance et du biais au carré de l'estimateur.

2.3.3 Test statistique et intervalle de confiance

La valeur estimée d'un paramètre donne une information ponctuelle sur la valeur de ce paramètre. Cette information n'est pas suffisante, car son intérêt dépend fortement de la précision de l'estimateur. Il est donc important d'associer à tout estimateur son écart type, ou mieux encore, d'indiquer un *intervalle de confiance* pour le paramètre. L'intervalle de confiance d'un paramètre est un intervalle de valeurs plausibles de ce paramètre, calculé en fonction des données observées. Ses bornes sont des variables aléatoires puisqu'elles dépendent des observations. Elles sont calculées de façon que la vraie valeur du paramètre soit incluse dans l'intervalle avec une certaine probabilité appelée probabilité de recouvrement. Cette probabilité est généralement fixée à une valeur élevée (0.95 ou 0.99) afin que l'intervalle de confiance ait une forte probabilité d'inclure la vraie valeur. Il existe de nombreuses méthodes statistiques pour calculer des intervalles de confiance en tenant compte du modèle et des données disponibles.

Une autre notion importante en statistique est celle de *test statistique*. Un test statistique consiste à appliquer une règle de décision pour décider si une hypothèse doit être rejetée ou non. Cette hypothèse est généralement appelée *hypothèse nulle* et notée H_0 car elle s'écrit souvent sous la forme d'une égalité à zéro. La règle de décision est basée sur une fonction T des données observées appelée *statistique de test*, choisie en fonction de l'hypothèse à tester H_0. Le test statistique consiste à calculer la valeur prise par T pour le jeu de données disponibles, puis à déterminer la probabilité que T dépasse la valeur calculée sous l'hypothèse H_0. Si cette probabilité s'avère trop faible (e.g., plus petite que 0.05), H_0 est rejetée. Le test est une notion de statistique fréquentiste : la probabilité de rejeter l'hypothèse H_0 est en effet déterminée en imaginant que l'on répète indéfiniment la procédure dans les mêmes conditions expérimentales et avec des paramètres fixés à leurs valeurs sous H_0.

L'hypothèse testée doit être définie en fonction du problème pratique considéré. Dans une problématique de modélisation, H_0 revient typiquement à supposer qu'un ou plusieurs paramètres du modèle sont égaux à zéro. Par exemple dans le cas du modèle de régression et de l'équation (2.1), évaluer s'il existe une relation entre le poids et la taille revient à tester l'hypothèse « $H_0 : \alpha = 0$ ».

En décidant d'accepter ou de rejeter H_0, un test statistique peut conduire soit à une bonne décision, soit à une mauvaise décision. Deux types d'erreur sont possibles : rejeter à tort H_0 alors que l'hypothèse est vraie, ou accepter à tort H_0 alors que l'hypothèse est fausse. Une décision correcte est prise dans les deux cas complémentaires : lorsque H_0 est rejetée et que l'hypothèse est fausse, ou lorsque H_0 n'est pas rejetée et qu'elle est vraie. Les quatre cas de figure sont résumés dans le tableau 2.2.

Les deux types d'erreur de décision sont notés erreur de type I et erreur

Décision	**Réalité**	
	H_0 vraie	H_0 fausse
Rejet de H_0	Erreur de type I	Bonne décision
Pas de rejet de H_0	Bonne décision	Erreur de type II

Tableau 2.2 – Les quatre cas de figure pour un test statistique.

de type II respectivement. La probabilité de rejeter H_0 dans le cas où cette hypothèse est fausse est appelée *puissance* du test. La puissance est égale à 1 – β, où β est la probabilité d'une erreur de type II. La performance d'un test statistique se mesure par les probabilités d'erreurs de type I et II. Ces probabilités dépendent de plusieurs facteurs, notamment du plan d'expériences (nombre de mesures, précision des mesures, organisation de l'expérimentation) de l'hypothèse testée et du type de test réalisé. En général, la probabilité de l'erreur de type I est fixée par la personne qui réalise le test, souvent à 1% ou à 5%. La probabilité d'erreur de type II est par contre difficile à déterminer ; elle dépend du vrai état du système, qui est par nature inconnu, mais également de la qualité des expérimentations. De façon générale, un test basé sur des mesures imprécises conduit moins souvent à un rejet de H_0 qu'un test basé sur des mesures précises. Il a donc une probabilité d'erreur de type II élevée (*i.e.*, une puissance faible) si H_0 est fausse.

De nombreuses méthodes ont été développées par les statisticiens pour calculer des intervalles de confiance et réaliser des tests dans diverses situations que l'on rencontre en pratique. Pour des compléments sur les principes et sur les méthodes, voir par exemple Daudin *et al.* (2001), McCulloch et Searle (2001), Saporta (2006), Denis et Monod (2007).

2.3.4 Inférence bayésienne

En statistique bayésienne, l'inférence consiste avant tout à déterminer la distribution *a posteriori* des paramètres. Une fois que celle-ci est déterminée, on peut en déduire des estimateurs, des intervalles de confiance — plutôt appelés intervalles de crédibilité en statistique bayésienne — ou diverses autres quantités d'intérêt pratique.

Dans des cas bien définis, la distribution *a posteriori* peut être calculée analytiquement, mais, en pratique, cette distribution doit généralement être estimée par un algorithme mathématique. Les algorithmes itératifs de type *Markov Chain Monte Carlo* (MCMC) ont été développés dans ce but (Gilks *et al.*, 1996 ; Carlin et Louis, 2008 ; Robert, 2006 ; Boreux *et al.*, 2010 ; Robert et Casella, 2011). Leur principe est de générer une longue série de valeurs des paramètres (plusieurs milliers), qui peut être considérée comme un échantillon représentatif de la distribution *a posteriori*. Ces valeurs sont générées de façon

aléatoire à partir de valeurs initiales fournies par l'utilisateur, selon une chaîne de Markov. Les algorithmes MCMC peuvent être mis en œuvre avec des logiciels statistiques. Le logiciel WinBUGS est basé sur un de ces algorithmes, l'échantillonneur de Gibbs.

Un certain nombre de notions de la statistique fréquentiste ont été adaptées aux approches bayésiennes. Ainsi, *l'estimateur bayésien* d'un paramètre est une quantité représentant une valeur typique de sa distribution *a posteriori* : il s'agit généralement de l'espérance, de la médiane ou du mode de la distribution *a posteriori* de ce paramètre. *L'intervalle de crédibilité* d'un paramètre est l'adaptation bayésienne de l'intervalle de confiance : il s'agit d'un intervalle de valeurs représentant une probabilité fixée de la distribution *a posteriori* du paramètre. Concrètement, ces différentes valeurs sont calculées à partir de l'échantillon généré par l'algorithme MCMC.

2.4 Les quatre étapes de la modélisation

Qu'il soit fréquentiste ou bayésien, le développement d'un modèle comporte quatre étapes importantes (tableau 2.3) :

- la définition des variables d'entrée et de sortie ;
- le choix des équations et des lois de probabilité associées aux variables aléatoires ;
- l'estimation des paramètres ;
- l'évaluation de la qualité du/des modèles.

Il est souvent souhaitable de répéter ces quatre étapes plusieurs fois, car les choix initiaux réalisés aux étapes *i*, *ii* et *iii* peuvent s'avérer peu pertinents une fois l'évaluation réalisée. Le modélisateur peut alors être amené à choisir de nouvelles variables d'entrée, une autre équation, ou à estimer plus précisément les paramètres du modèle initial (voir Ennaïfar *et al.*, 2007 pour une illustration de cette approche en épidémiologie végétale).

2.4.1 Définition des variables

La définition des variables d'entrée et de sortie (étape *i*) est réalisée en fonction des objectifs du projet. Le choix de la ou des variables de sortie, appelées également variables réponses, dépend directement de la nature du risque qui doit être étudié. Par exemple, si l'objectif est de prévoir un risque de pollution de l'eau par les nitrates, la variable de sortie devra avoir un lien aussi direct que possible avec ce risque. Dans ce cas, le modélisateur peut définir la variable de sortie comme la teneur en nitrate de l'eau ou le reliquat d'azote minéral dans le sol à un stade clé (Lacroix *et al.*, 2005 ; Makowski *et al.*, 2001). Si l'objectif est de prévoir le risque d'infestation d'une parcelle agricole par les adventices, la variable de sortie peut être, par exemple, la biomasse d'adventices, leur densité ou le nombre de graines présentes dans le sol (Primot *et al.*, 2006). Le choix de la variable dépend également du type de données expérimentales disponibles.

Une variable de sortie pertinente sur le plan pratique peut s'avérer difficile à introduire dans un modèle si cette variable est peu ou pas disponible dans les bases de données.

Étape principale	Sous-étape
i. Définition des variables d'entrée et de sortie	Formulation des objectifs du modèle
	Analyse des données disponibles
	Définition de variables candidates
ii. Définition des équations, modélisation	Bibliographie et analyse des connaissances sur le fonctionnement du système
	Formulation des équations du modèle (en une ou plusieurs variantes)
	Description de l'information *a priori* sur les paramètres (cadre bayésien)
iii. Estimation des paramètres	Planification expérimentale et constitution d'une base de données
	Choix de la méthode d'estimation
	Application avec un logiciel
iv. Évaluation	Définition d'un ou plusieurs critères d'évaluation
	Planification expérimentale et constitution d'une base de données pour l'évaluation
	Estimation des critères d'évaluation pour chaque modèle candidat
	Choix d'un modèle, définition de nouveaux modèles candidats ou combinaison de plusieurs modèles

Tableau 2.3 – Une démarche de modélisation en quatre étapes.

Les variables d'entrée sont des variables qui sont supposées avoir un effet sur la ou les variables de sortie, ou au moins une relation avec ces variables. Elles doivent être choisies en fonction des connaissances disponibles sur le fonctionnement du système étudié. Elles doivent également être choisies en fonction du mode d'utilisation du modèle. Par exemple, si la variable de sortie doit être simulée à une certaine date, il est inutile de considérer des variables d'entrée qui

ne seront disponibles qu'après cette date. Par ailleurs, certaines variables d'entrée identifiées sur la base de connaissances théoriques peuvent s'avérer difficiles à introduire dans un modèle du fait d'un manque de données expérimentales. L'introduction de telles variables peut alors dégrader les performances du modèle car les paramètres associés seront mal estimés dans une telle situation.

Une bonne approche consiste souvent à définir plusieurs jeux de variables d'entrée candidates, à construire les modèles correspondants, puis à les évaluer lors de l'étape *iv* (Ennaifar *et al.*, 2007 ; Makowski *et al.*, 2001 ; Primot *et al.*, 2006).

2.4.2 Choix des équations

L'objectif de l'étape *ii* est de définir une ou plusieurs équations candidates pour relier les variables d'entrée aux variables de sortie. Différents types d'équation peuvent être définis incluant un nombre de paramètres plus ou moins grand. Le modélisateur est souvent tenté de définir une équation très complexe qui semble satisfaisante sur le plan biologique. Il est souvent plus pertinent de définir plusieurs équations avec des niveaux de complexité contrastés. Le choix final est alors fait lors de l'étape *iv* ou avec des techniques de sélection de modèles.

2.4.3 Estimation des paramètres

L'estimation des paramètres (étape *iii*) nécessite *a)* de constituer une base de données incluant des valeurs observées des variables d'entrée et de sortie ; *b)* de choisir une méthode d'estimation (moindres carrés, maximum de vraisemblance, etc.) ; *c)* d'appliquer la méthode choisie avec un logiciel. Si une méthode d'estimation bayésienne est utilisée, il est également nécessaire de décrire l'information disponible sur les valeurs des paramètres et d'en déduire des lois *a priori*, par exemple à partir de résultats déjà publiés dans le passé. La difficulté de l'étape *iii* dépend de la nature de l'équation (linéaire, non linéaire, etc.), du nombre de paramètres à estimer et de la base de données disponible.

Il existe une grande diversité de méthodes d'estimation. Celles-ci sont décrites dans les sections suivantes pour différentes familles de modèle. L'utilité d'un modèle dépend fortement de la précision des valeurs estimées des paramètres. La précision des valeurs estimées dépend elle-même de la méthode utilisée pour estimer les paramètres et surtout de la qualité de la base de données disponible (*e.g.*, Makowski et Wallach, 2001, 2002 ; Wallach *et al.*, 2006), que celle-ci soit issue du recueil de données existantes, d'une enquête ou d'un échantillonnage sur le terrain, ou encore d'une expérimentation spécifique. Il est donc important de s'assurer que les données utilisées pour l'estimation aient été collectées avec des méthodes de mesure adaptées, selon un plan d'expérience approprié.

2.4.4 Évaluation des modèles

L'étape *iv* consiste à évaluer la qualité du ou des modèles obtenus. L'objectif est de déterminer si les modèles considérés donnent une description suffisamment réaliste du système étudié. Cette étape d'évaluation peut se faire en partie à partir du jeu de données utilisé pour estimer les paramètres, mais elle est bien plus fiable lorsqu'elle peut s'effectuer à partir de jeux de données différents, réservés à l'évaluation de la qualité de prédiction du modèle. Des approches basées sur l'analyse des résidus et sur le calcul de critères quantitatifs (critère d'Akaïke et *Deviation Information Criterion*) sont présentées dans le chapitre 3. Lorsque les modèles sont utilisés pour optimiser des décisions, il est important d'évaluer la qualité des décisions prises grâce aux modèles. Cet aspect de l'évaluation sera traité dans le Chapitre 4.

L'étape d'évaluation nécessite : *a)* la définition d'un ou plusieurs critères d'évaluation ; *b)* la constitution d'une base de données ; *c)* l'estimation du critère avec les données pour chacun des modèles candidats ; *d)* le choix d'un modèle ou la définition de nouveaux modèles candidats (tableau 2.3). L'étape d'évaluation peut aboutir à un choix précis de modèle ou à une remise en cause du ou des modèles testés. D'autres modèles peuvent alors être définis pour un nouveau cycle de développement et d'évaluation, basés éventuellement sur de nouvelles variables d'entrée et de nouvelles équations. L'étape d'évaluation peut également remettre en cause la méthode utilisée pour estimer les paramètres.

Il arrive fréquemment que plusieurs modèles aient des performances très voisines pour un ou plusieurs critères d'évaluation. Dans ce cas, il est souvent plus pertinent de combiner les modèles plutôt que de choisir le meilleur (Burnham et Anderson, 2002). Cette approche revient à utiliser un *ensemble de modèles* plutôt qu'un modèle unique pour réaliser l'analyse du risque. Des exemples d'application de cette démarche seront présentés dans le chapitre 3. Plusieurs méthodes de combinaison de modèles peuvent être facilement appliquées avec les librairies `BMA` et `MMIX` du logiciel R.

2.4.5 Importance de la planification expérimentale

L'influence du *plan d'expérience* sur la précision des valeurs estimées des paramètres et sur la performance d'un modèle peut être illustrée avec le modèle de régression d'équation (2.1). Un estimateur classique du paramètre α de ce modèle est l'estimateur des moindres carrés défini par

$$\widehat{\alpha} = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} {x_i}^2} \tag{2.2}$$

avec y_i, $i = 1, \ldots, N$, les valeurs observées de Y. Ces observations ont été obtenues pour N valeurs de x, notées $x_1, \ldots, x_i, \ldots, x_N$. Lorsque les erreurs ε

du modèle (2.1) sont d'espérance nulle et de variance constante σ^2, la variance de l'estimateur (2.2) est

$$\mathrm{Var}(\widehat{\alpha}) = \frac{\sigma^2}{\sum_{i=1}^{N} {x_i}^2}. \tag{2.3}$$

L'équation (2.3) montre que, pour un nombre de mesures N fixé, la variance de l'estimateur est d'autant plus petite (et l'estimateur d'autant plus précis) que la somme des valeurs de x_i au carré est grande. Cela signifie que l'expérimentateur a intérêt à définir un plan d'expérience avec des valeurs de x_i grandes en valeurs absolues pour estimer précisément le paramètre du modèle (2.1).

D'une manière plus générale, la définition d'un plan d'expérience consiste à définir les conditions de réalisation de l'expérience destinée à collecter des données. Cette définition passe par l'identification des facteurs expérimentaux et de l'ensemble des combinaisons de niveaux des facteurs à tester. Dans le cadre de l'estimation des paramètres d'un modèle, on choisira un plan d'expériences permettant de maximiser la précision des estimations avec un coût expérimental minimal et une bonne robustesse par rapport aux hypothèses du modèle. Des principes rigoureux de planification expérimentale ont été définis depuis les travaux de Fisher et Yates dans les années 1920 et 1930, en particulier la répétition des traitements pour estimer la variabilité des observations et améliorer la précision, la prise en compte de facteurs blocs pour contrôler au mieux l'hétérogénéité des observations, la randomisation pour rendre les analyses plus robustes à des écarts à certaines hypothèses sur les erreurs. De nombreuses méthodes ont été définies pour optimiser les plans d'expérience pour différents types de modèle et des ouvrages spécialisés ont été consacrés à ce sujet (*e.g.*, Droesbeke *et al.*, 1997 ; Dean et Voss, 1999 ; Dagnelie, 2003 ; Bailey, 2008).

2.5 Exercices

Exercice 2.1. Supposons que six observations d'une variable aléatoire Y soient disponibles (0.14, 0.03, 4.69, 6.74, 8.98, 8.28) et que ces observations aient été obtenues pour six valeurs d'une variable X égales respectivement à 2, 7, 10, 14, 17 et 21. On veut relier Y à X à l'aide du modèle (2.1).
Utiliser les données pour estimer le paramètre α par la méthode des moindres carrés.

Exercice 2.2. Déterminer la distribution de probabilité de la variable $Z = X_1 + 3X_2$ sachant que X_1 et X_2 sont des variables aléatoires indépendantes gaussiennes d'espérances respectivement égales à 1 et 2, et de variances respectivement égales à 1 et 3.
Propriété : Si X_1 et X_2 sont deux variables aléatoires gaussiennes indépen-

dantes alors $aX_1 + bX_2$ suit une loi gaussienne telle que :

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2),$$
$$\mathrm{Var}(aX_1 + bX_2) = a^2 \mathrm{Var}(X_1) + b^2 \mathrm{Var}(X_2).$$

Exercice 2.3. Supposons que la variable aléatoire Y soit distribuée selon une loi gaussienne d'espérance μ et de variance 1. Supposons que trois observations indépendantes de Y soient disponibles et égales à : 2.10, 1.48, 3.09.
Calculer les vraisemblances de $\mu = 2$ et $\mu = 5$.
Quelle valeur de μ a la plus grande vraisemblance ?
Propriété : la densité de probabilité d'une loi gaussienne d'espérance μ et de variance σ^2 est égale à

$$g\,(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2 \right).$$

Exercice 2.4. Considérons un paramètre α de valeur inconnue, mais ne pouvant prendre que deux valeurs : $\alpha = 0.25$ et $\alpha = 0.75$. Une seule observation d'une variable aléatoire Y est disponible. On suppose que cette observation est égale à 5 et qu'elle est reliée à α par les probabilités conditionnelles suivantes : $P(y = 5 \mid \alpha = 0.25) = 0.25$ et $P(y = 5 \mid \alpha = 0.75) = 0.5$.
Quelle valeur de α a la plus grande vraisemblance ?
Supposons que les probabilités *a priori* des valeurs de α soient égales à $P(\alpha = 0.25) = P(\alpha = 0.75) = 0.5$.
Calculer les probabilités *a posteriori* pour les deux valeurs de α. Quelle valeur a la plus grande probabilité *a posteriori* ?
Même question si $P(\alpha = 0.25) = 0.75$ et $P(\alpha = 0.75) = 0.25$.

Exercice 2.5. Un champignon pathogène affecte gravement les cultures de blé en Amérique du Nord. Il est pour l'instant absent en Europe, mais la Commission européenne vous demande de déterminer si les conditions climatiques de l'Union européenne sont favorables au développement de ce champignon. Définir une démarche pour développer un modèle qui pourrait être utilisé pour répondre à la question de la Commission européenne.

Exercice 2.6. Un lot de 20 rats est nourri avec des grains issus d'une culture OGM pendant 90 jours. À l'issue de cette période, deux rats présentent une pathologie particulière. Sachant que la fréquence de cette pathologie est habituellement de 1% chez les rats nourris de manière traditionnelle, calculer la probabilité d'obtenir deux rats malades parmi 20 sous l'hypothèse que les grains OGM n'ont aucun effet sur la santé des rats. Acceptez-vous ou rejetez-vous cette hypothèse ?

Chapitre 3

Modèles statistiques et évaluation des risques

Ce chapitre présente plusieurs types de modèle permettant de prévoir l'occurrence d'un risque et d'évaluer son impact, en s'appuyant sur des données observées. Les modèles présentés ici permettent de simuler une variable ayant un intérêt pratique en fonction d'une ou plusieurs variables explicatives correspondant à des caractéristiques du milieu et/ou à des pratiques agricoles.

Le premier type de modèle considéré dans ce chapitre est le modèle linéaire, outil de base pour de nombreux problèmes appliqués (section 3.1). Nous présentons ensuite le modèle linéaire généralisé (section 3.2) ainsi que le modèle non linéaire (section 3.3). Les deux dernières sections du chapitre décrivent des variantes utiles pour traiter certains risques agro-environnementaux : modèle hiérarchique (section 3.4) et régression quantile (section 3.5).

Tous ces modèles constituent des outils puissants pour la première étape de l'analyse des risques, *l'évaluation des risques*. D'autres méthodes sont présentées dans les chapitres 4 et 5 pour les étapes de gestion des risques et de communication.

3.1 Modèle linéaire

3.1.1 Définition

Le modèle linéaire permet d'étudier la relation entre une variable réponse Y quantitative (*e.g.*, rendement d'une culture, quantité d'azote minéral du sol) et des variables d'entrée ou variables explicatives $X_1, ..., X_p$, quantitatives (*e.g.*, pluviométrie), qualitatives ordonnées (*e.g.*, grand, moyen, petit) ou qualitatives non ordonnées (*e.g.*, rouge, jaune, noir, blanc, ou des noms de variétés ou de régions). Il s'applique à un jeu de N données pour lesquelles on dispose à la fois des observations y_i, pour $i = 1, \ldots, N$ et des valeurs des variables explicatives

$x_{1,i}, \ldots, x_{p,i}$. La variable réponse est considérée comme aléatoire, car observée avec erreur, alors que les variables explicatives sont considérées connues de façon exacte.

De façon plus précise, le modèle linéaire (Azaïs et Bardet, 2006 ; Cornillon et Matzner-Lober, 2011) relie la variable réponse Y aux variables d'entrée $X_1, \ldots, X_p$ par l'équation suivante :

$$\begin{aligned} Y &= f(X_1, \ldots, X_p\,;\, \alpha_0, \ldots, \alpha_p) + \varepsilon \\ &= \alpha_0 + \sum_{j=1}^{p} \alpha_j X_j + \varepsilon, \end{aligned} \tag{3.1}$$

où $\alpha_0, \alpha_1, ..., \alpha_p$ sont des paramètres et ε est une variable aléatoire représentant l'erreur du modèle. L'erreur ε est supposée

- centrée, c'est-à-dire d'espérance nulle $E(\varepsilon) = 0$;
- *homoscédastique*, c'est-à-dire de même variance σ^2 pour toutes les observations ;
- et indépendante, ou en tout cas de corrélation nulle, entre observations.

Le plus souvent, on suppose de plus que l'erreur suit une loi de probabilité gaussienne ; on note alors $\varepsilon \sim N\left(0, \sigma^2\right)$.

La variance σ^2 correspondant à un paramètre additionnel du modèle, l'ensemble des paramètres du modèle (3.1) est $\alpha_0, \alpha_1, ..., \alpha_p, \sigma^2$. Comme les erreurs du modèle sont centrées et que les variables d'entrée $X_1, ..., X_p$ sont supposées fixées et connues, la variable aléatoire Y vérifie

$$\begin{cases} E(Y) = \alpha_0 + \displaystyle\sum_{j=1}^{p} \alpha_j X_j ; \\ \mathrm{Var}(Y) = \sigma^2. \end{cases}$$

Ce modèle est appelé « modèle linéaire » parce que la variable Y est reliée linéairement aux paramètres $\alpha_0, \alpha_1, \ldots, \alpha_p$. Notons qu'avec cette définition le modèle $Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \varepsilon$ est un modèle linéaire avec $X_1 = X$ et $X_2 = X^2$: le modèle linéaire n'est pas nécessairement linéaire par rapport aux variables d'entrée.

3.1.2 Généralité du modèle linéaire

On rencontre parfois l'expression *modèle linéaire général*, en particulier dans la bibliographie anglo-saxonne. Dans cette expression, le terme *général* illustre le fait que l'on peut intégrer et combiner dans le modèle linéaire des variables explicatives aussi bien quantitatives que qualitatives. Différents types de modèle linéaire ont d'ailleurs été définis selon la nature des variables d'entrée, qui ne sont que des cas particuliers du modèle linéaire :

- la régression linéaire simple (modèle incluant une seule variable d'entrée quantitative, défini par $E(Y) = \alpha_0 + \alpha_1 X$) ;

- la régression linéaire multiple (plusieurs variables d'entrée quantitatives) ;
- le modèle d'analyse de la variance (une ou plusieurs variables d'entrée qualitatives) ;
- le modèle d'analyse de la covariance (mélange de variables d'entrée quantitatives et qualitatives).

Ces modèles peuvent tous être décrits à l'aide de l'équation (3.1). Ils peuvent être limités à la somme des effets simples des différentes variables d'entrée ou bien inclure des puissances et produits de variables quantitatives (régression polynomiale) ou des effets d'interaction entre variables qualitatives.

Pour les variables d'entrée qualitatives, le modèle linéaire s'écrit en associant des variables d'entrée binaires (0-1) aux différentes modalités. Par exemple, un modèle linéaire incluant une variable d'entrée X qualitative à trois niveaux est défini par :

$$Y = \mu_j + \varepsilon,$$

où j représente une modalité de la variable X ($j = 0, 1, 2$) et μ_j représente l'effet de cette modalité j sur la variable réponse. Ce modèle s'écrit aussi

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon,$$

où X_1 et X_2 sont construites de la façon suivante :

$$\begin{aligned} X_1 = 0 \text{ et } X_2 = 0 \text{ si } X = 1,\\ X_1 = 1 \text{ et } X_2 = 0 \text{ si } X = 2,\\ X_1 = 0 \text{ et } X_2 = 1 \text{ si } X = 3. \end{aligned}$$

Ce type de modèle peut, par exemple, être utilisé pour comparer l'effet de trois variétés de blé sur le rendement. Dans ce cas, Y est alors le rendement observé, $\mu_0 = \alpha_0$ est l'espérance du rendement pour la première variété ($X_1 = X_2 = 0$), $\mu_1 = \alpha_0 + \alpha_1$ est l'espérance du rendement pour la deuxième ($X_1 = 1$ et $X_2 = 0$) et $\mu_2 = \alpha_0 + \alpha_2$ est l'espérance du rendement pour la troisième variété ($X_1 = 0$ et $X_2 = 1$).

Ce modèle peut également être utilisé pour étudier l'effet de plusieurs régimes alimentaires sur les rats de laboratoire, par exemple un régime basé sur une culture génétiquement modifiée ($X_1 = X_2 = 0$), un régime basé sur une variété non génétiquement modifiée mais ayant un fond génétique proche ($X_1 = 1$, $X_2 = 0$) et une variété commerciale standard ($X_1 = 0$, $X_2 = 1$). Dans ce cas, la variable Y correspond soit au poids d'un rat, soit à une variable physiologique (*e.g.*, teneur en sodium du sang). Ce modèle permet ainsi d'évaluer les risques toxicologiques liés à la consommation de produits dérivés de cultures génétiquement modifiées.

3.1.3 Estimation des paramètres

Les paramètres $\alpha_0, \alpha_1, \ldots, \alpha_p$ et σ^2 doivent être estimés à partir des données observées $(y_i, x_{1,i}, \ldots, x_{p,i})$, pour $i = 1, \ldots, N$. De nombreuses méthodes

d'estimation ont été développées dans cette optique, mais la méthode de loin la plus courante pour le modèle linéaire, dans le cadre fréquentiste, est celle desmoindres carrés. Elle consiste à utiliser comme estimateurs $\widehat{\alpha}_0, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_p$, les valeurs qui minimisent la somme des carrés résiduelle, définie par

$$\mathrm{SCR} = \sum_{i=1}^{N} \left(y_i - (\alpha_0 + \alpha_1 x_{1,i} + \ldots + \alpha_p x_{p,i})\right)^2 .$$

Si l'on suppose que les erreurs suivent des lois gaussiennes ($\varepsilon \sim N\left(0, \sigma^2\right)$) et qu'elles sont indépendantes entre les observations, la fonction de vraisemblance des paramètres est égale à

$$V(\alpha_0, \ldots, \alpha_p, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j x_{j,i}\right)^2\right).$$

On peut en déduire que la méthode des moindres carrés coïncide alors avec la *méthode du maximum de vraisemblance* pour l'estimation des paramètres α_j. L'estimateur du maximum de vraisemblance de la variance résiduelle est SCR/N. On lui préfère généralement $\widehat{\sigma^2} = \mathrm{SCR}/(N - p - 1)$, qui est un estimateur non biaisé de la variance.

Le résultat de la méthode des moindres carrés s'obtient par calcul matriciel. Il se présente sous la forme de valeurs ponctuelles des paramètres ou d'intervalles de confiance.

Dans le cadre bayésien, la différence essentielle avec l'approche classique est que les paramètres $\alpha_0, \alpha_1, \ldots, \alpha_p$, σ^2 du modèle défini par l'équation (3.1) sont représentés par des variables aléatoires auxquelles on associe une loi de probabilité *a priori* et une loi *a posteriori*. La loi *a priori* décrit les connaissances disponibles sur les valeurs des paramètres avant utilisation des données. Elle résume la connaissance initiale que possède le modélisateur sur les valeurs des paramètres. La loi *a posteriori* est calculée en combinant la vraisemblance et la loi *a priori*. La loi *a posteriori* résume l'ensemble des informations disponibles sur les paramètres : les connaissances disponibles avant utilisation des données et les données expérimentales.

En statistique bayésienne, l'estimation a pour objectif de déterminer la loi *a posteriori* des paramètres à partir de la loi *a priori* et de la fonction de vraisemblance. Une caractéristique importante des méthodes d'estimation bayésiennes est qu'elles permettent d'obtenir des distributions de valeurs des paramètres $\alpha_0, \alpha_1, \ldots, \alpha_p, \sigma^2$ et pas seulement des estimations ponctuelles.

3.1.4 Évaluation et limites du modèle linéaire

Le modèle linéaire est limité à des relations linéaires entre paramètres et variables réponses. Les hypothèses de base sur le terme d'erreur sont par ailleurs

des hypothèses fortes pas toujours réalistes. Ainsi, dans de nombreuses situations, il n'est pas réaliste de définir une valeur unique σ^2 pour les erreurs de toutes les observations, par exemple lorsque certaines observations sont moins précises que d'autres. Les erreurs peuvent également être corrélées plutôt qu'indépendantes. Il est alors nécessaire de faire d'autres hypothèses et de définir une loi de probabilité plus complexe incluant un nombre plus élevé de paramètres pour décrire la loi du terme d'erreur.

Lors d'une analyse par le modèle linéaire, il est indispensable de s'assurer que les hypothèses de base sont vérifiées, en particulier par l'étude des résidus. Les résidus sont les différences $r = y - \widehat{\alpha}_0 - \sum_{j=1}^{p} \widehat{\alpha}_j x_j$ entre les observations et les prédictions obtenues à partir des valeurs estimées des paramètres. On peut les interpréter comme des estimations $\widehat{\varepsilon}$ des erreurs et certains graphiques sur les résidus (voir les exemples ci-dessous) permettent de vérifier si les hypothèses sur le terme d'erreur du modèle sont justifiées. Si les hypothèses de base du modèle linéaire ne sont pas vérifiées, il est parfois possible d'opérer une transformation sur les variables réponse mais, dans la plupart des cas, il est préférable d'utiliser l'une des nombreuses extensions du modèle linéaire que nous verrons dans les sections suivantes.

3.1.5 Exemple : prédiction de la teneur en azote et de la teneur en protéines des grains de blé

La teneur en azote des grains de blé est utilisée pour calculer les valeurs d'indicateurs de risque de pollution de l'eau par les nitrates, notamment les indicateurs de type « bilan » (*e.g.* « dose d'engrais appliquée − teneur en azote des grains × rendement »). Ces indicateurs permettent d'identifier des situations à forts risques de pollution et de raisonner la mise en place de mesures visant à réduire la pollution par les nitrates, par exemple l'implantation de cultures pièges à nitrates (Makowski *et al.*, 2009 ; Meynard *et al.*, 2002 ; Yang *et al.*, 2007). La précision des valeurs fournies par ces indicateurs dépend de la précision de l'estimation des différents termes du bilan, notamment de la teneur en azote.

La teneur en protéines des grains de blé est proportionnelle à la teneur en azote des grains. Elle constitue un critère de qualité important pour les entreprises qui collectent et stockent les récoltes de blé. La teneur en protéines des grains détermine le type d'utilisation industrielle d'une récolte (panification, fabrication de biscuits, alimentation animale, etc.). Si les grains de blé ont une teneur en protéines trop faible, ils ne pourront pas être utilisés pour la panification, à moins de les mélanger avec des grains ayant une teneur plus élevée. Il est donc important pour les entreprises de collecte-stockage de pouvoir prédire, avant la récolte, la qualité du blé afin d'organiser le stockage des grains en silo et de passer des contrats (Le Bail et Makowski, 2004 ; Le Bail *et al.*, 2005).

Nous montrons ici comment la teneur en protéines des grains de blé peut

être prédite en utilisant des modèles linéaires incluant une ou deux variables explicatives mesurables avant la récolte. Nous nous focalisons ici sur la teneur en protéines mais les mêmes modèles peuvent être utilisés pour prédire la teneur en azote. Les quatre étapes présentées dans la section 2.4 (tableau 2.3) sont reprises ici une à une. Les variables et les équations sont définies lors des étapes *i* et *ii*. Les paramètres des modèles sont ensuite estimés (étape *iii*) à partir de données expérimentales en utilisant une méthode de la statistique classique puis une méthode bayésienne. Finalement, les modèles sont évalués lors de l'étape *iv*.

Étape *i* : définition des variables

La variable de sortie est la teneur en protéines mesurée à la récolte Y. Deux variables d'entrée sont considérées, X_1 et X_2. La variable X_1 est une mesure de transmittance réalisée sur un échantillon de feuilles de blé avec le chlorophyl meter Minolta (*SPAD*) selon la méthode proposée par Le Bail *et al.* (2005). La variable X_2 est une mesure de l'indice de nutrition azotée du blé (INN) à floraison, définie par Justes *et al.* (1994). Les deux types de mesure peuvent être réalisés avant la récolte et permettent donc d'anticiper le niveau de qualité d'une récolte de blé. La mesure SPAD est cependant nettement plus facile à réaliser en pratique que la mesure INN. La mesure SPAD ne nécessite en effet que des pincements des feuilles avec l'appareil de mesure alors que la mesure INN est basée sur des prélèvements de plantes et des mesures de biomasse et de teneur en azote réalisées en laboratoire.

Étape *ii* : définition des équations

Trois équations sont définies pour relier Y aux variables explicatives :

$$\begin{aligned} \mathrm{M}_1 &: Y = \alpha_0 + \alpha_1 X_1 + \varepsilon, \\ \mathrm{M}_2 &: Y = \alpha_0 + \alpha_2 X_2 + \varepsilon, \\ \mathrm{M}_3 &: Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon. \end{aligned}$$

Ces trois équations définissent trois modèles linéaires différents. Le premier permet de simuler la teneur en protéines en fonction de la mesure SPAD, le deuxième en fonction de l'INN et le troisième en fonction des deux variables explicatives. Nous supposons que les erreurs ε sont indépendantes et ont toutes la même loi, $\varepsilon \sim N\left(0, \sigma^2\right)$. Cela revient à supposer, par exemple pour le modèle M_1, que $Y \sim N\left(\alpha_0 + \alpha_1 X_1, \sigma^2\right)$. Les modèles M_1 et M_2 sont des cas particuliers du modèle M_3, correspondant à $\alpha_2 = 0$ et $\alpha_1 = 0$ respectivement : les modèles M_1 et M_2 sont *emboîtés* dans le modèle M_3. Notons que ces modèles peuvent être utilisés pour prédire la teneur en azote des grains plutôt que la teneur en protéines en divisant leur sortie par une constante de conversion (« teneur en protéines = 5.7 × teneur en azote », Grundy *et al.*, 1996).

Étape *iii* : estimation des paramètres

Données Quarante-trois expérimentations ont été réalisées en exploitations agricoles sur trois années (2004, 2005, 2006) et sur plusieurs sites de la Beauce (France). Elles sont utilisées pour estimer les paramètres des trois modèles. Chaque expérimentation est constituée d'une parcelle d'agriculteur sur laquelle l'INN, le SPAD et la teneur en protéines (%) ont été mesurés (figure 3.1). L'INN a été mesuré à partir de prélèvements de biomasse réalisés sur 15 placettes de 0.15 m^2 par parcelle. Le SPAD a été mesuré sur 15 feuilles de blé par parcelle. La teneur en azote des grains a été mesurée à partir de prélèvements de grains réalisés à la récolte sur 6 placettes de 0.15 m^2 par parcelle. Les données sont décrites en détail par Barbottin *et al.* (2008).

Deux méthodes sont comparées ci-dessous pour estimer les paramètres à partir des 43 données expérimentales, une méthode issue de la statistique classique (moindres carrés ordinaires) et une méthode bayésienne.

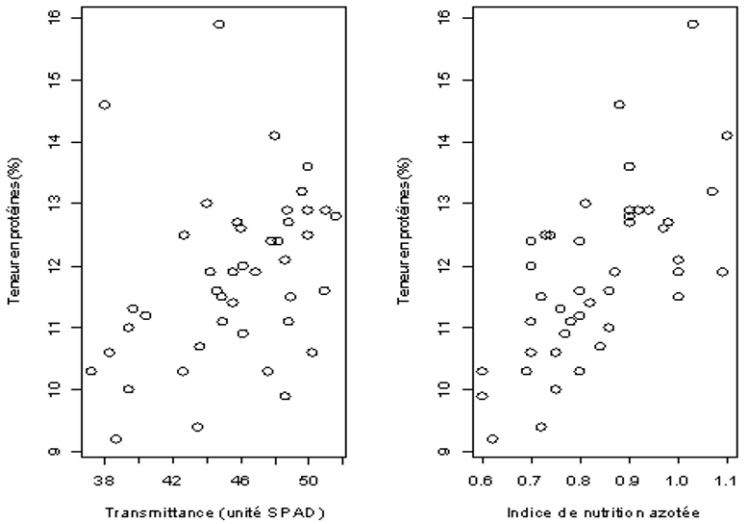

Figure 3.1 – Mesures de teneur en protéines, transmittance et indice de nutrition azotée obtenues sur 43 sites-années de Beauce.

Moindres carrés ordinaires Les paramètres $\alpha_0, \alpha_1, \alpha_2$ et σ^2 sont définis comme des nombres fixes inconnus. Le principe de la méthode est de calculer les valeurs de $\alpha_0, \alpha_1, \alpha_2$ qui minimisent la somme des carrés des écarts entre les mesures de Y et les valeurs de Y calculées par la partie fixe du modèle. Pour le premier des trois modèles, cela revient à calculer les valeurs de α_0 et α_1 qui minimisent $\sum_{i=1}^{43} (y_i - \alpha_0 - \alpha_1 x_{1i})^2$, avec y_i la i-ième mesure de teneur

en protéines et x_{1i} la i-ième mesure SPAD, pour $i = 1, \ldots, 43$. Ces valeurs sont notées $\widehat{\alpha}_0$ et $\widehat{\alpha}_1$ (le signe $\widehat{\ }$ désigne une valeur estimée).

On montre que

$$\widehat{\alpha}_1 = \frac{\sum_{i=1}^{43} (y_i - \bar{y})(x_{1i} - \bar{x}_1)}{\sum_{i=1}^{43} (x_{1i} - \bar{x}_1)^2} \quad \text{et} \quad \widehat{\alpha}_0 = \bar{y} - \widehat{\alpha}_1 \bar{x}_1,$$

avec $\bar{y}$ et $\bar{x}_1$ les moyennes des observations de Y et X_1 respectivement.

La variance des erreurs σ^2 est estimée par

$$\widehat{\sigma}^2 = \frac{1}{41} \sum_{i=1}^{43} (y_i - \widehat{\alpha}_0 - \widehat{\alpha}_1 x_{1i})^2,$$

appelée variance résiduelle.

La méthode des moindres carrés peut être mise en œuvre avec le logiciel R en utilisant les instructions suivantes (Venables et Ripley, 2002) :

```
# Code R

### Stockage des donnees dans TAB a partir d'un fichier texte
### a l'aide de l'instruction read.table

TAB <- read.table("f:\\Projets\\Exemple1.txt", header=T, sep="\t")

### Moindres carres a l'aide de l'instruction lm

Mod.1 <- lm(Proteines ~ SPAD, data=TAB)
Mod.2 <- lm(Proteines ~ INN, data=TAB)
Mod.3 <- lm(Proteines ~ SPAD+INN, data=TAB)

### Presentation des resultats a l'aide de l'instruction summary

print(summary(Mod.1))
print(summary(Mod.2))
print(summary(Mod.3))
```

L'instruction `read.table` permet à R de lire un tableau de données écrit dans un fichier, par exemple un fichier texte `.txt`. Ici, le fichier contient trois colonnes numériques intitulées « Proteines », « SPAD » et « INN ». Les noms de colonne forment la première ligne du fichier et les mesures sont séparées par des tabulations. Les données lues par `read.table` sont stockées dans un tableau (appelé ici `TAB`) qui peut être manipulé dans R. L'instruction `lm(`*`linear model`*`)` permet d'estimer les paramètres de modèles linéaires incluant une ou

plusieurs variables explicatives à partir des données incluses dans `TAB`. L'utilisation de `lm` est illustrée dans le programme ci-dessus avec les trois modèles M_1, M_2 et M_3. On écrit le modèle en spécifiant la variable réponse à gauche du signe $\sim$ et le modèle proprement dit à droite du signe $\sim$ et en séparant les variables explicatives par le signe `+`. Les résultats sont extraits à l'aide de l'instruction `summary`. Ici, les deux variables explicatives sont quantitatives. Lorsque l'une des variables doit être considérée comme une variable qualitative, il faut utiliser l'instruction `factor` pour le déclarer.

Modèle	$\widehat{\alpha}_0$	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\sigma}$
1	6.69 (2.32) **	0.113 (0.051) *	-	1.31
2	6.09 (1.04) ***	-	6.90 (1.23) ***	1.05
3	3.92 (1.91) *	0.056 (0.042)	6.43 (1.27) ***	1.03

Tableau 3.1 – Valeurs des paramètres des trois modèles de teneur en protéines, estimées avec la méthode des moindres carrés ordinaires. Les valeurs entre parenthèses correspondent aux écarts types des estimateurs des paramètres. Les étoiles indiquent si la valeur estimée du paramètre est significativement différente de zéro avec une probabilité d'erreur de type I inférieure à 0.001 (***), 0.01 (**), 0.05 (*), 0.1 (.) selon le test de Student.

Le tableau 3.1 présente les valeurs estimées des paramètres des trois modèles ainsi que les écarts types des estimateurs, qui donnent une information sur la précision des valeurs estimées. Le tableau 3.1 présente également les résultats du test de Student appliqué à chaque paramètre. Dans ce test, l'hypothèse nulle est que le paramètre en question est égal à 0. Le résultat du test permet d'évaluer si les valeurs estimées sont significativement différentes de zéro ou si les effets peuvent être attribués aux erreurs d'observation et au hasard de l'échantillonnage. Les résultats montrent ici que $\widehat{\alpha}_2$ est significativement différent de zéro mais pas $\widehat{\alpha}_1$ dans le modèle M_3, lorsque l'on utilise un niveau de probabilité égal à 5% (*i.e.*, probabilité de rejeter l'hypothèse nulle à tort = 0.05).

Méthode d'estimation bayésienne Les modèles M_1, M_2 et M_3 peuvent également être analysés selon l'approche bayésienne. Dans le cadre bayésien, les paramètres $\alpha_0, \alpha_1, \alpha_2$ et σ^2 sont des variables aléatoires. Les relations entre les éléments aléatoires du modèle sont parfois présentées graphiquement, comme illustré sur la figure 3.2 pour le modèle M_1. Les flèches continues indiquent des relations stochastiques entre variables. Par exemple, dans le cas du modèle M_1, l'observation Y est reliée à $\alpha_0 + \alpha_1 X_1$ et à σ par une relation de nature stochastique :

$$Y|X_1, \alpha_0, \alpha_1, \sigma^2 \sim N\left(\alpha_0 + \alpha_1 X_1, \sigma^2\right).$$

Lors de la mise en œuvre d'une méthode d'estimation bayésienne, la première étape consiste à définir une loi *a priori* pour les paramètres. Cette loi

décrit l'état de connaissance initial dans les valeurs du paramètre, avant d'utiliser les données disponibles. Ici, nous supposons qu'aucune information *a priori* n'est disponible. Pourtant, même dans ce cas, il est nécessaire de définir des lois *a priori* pour pouvoir utiliser une méthode bayésienne. Il est alors judicieux de définir des lois de probabilité ayant de grandes variances de façon à représenter un faible niveau d'information sur les valeurs des paramètres.

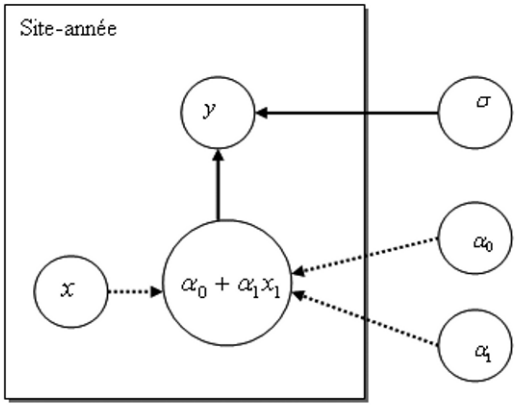

Figure 3.2 – Représentation graphique du modèle linéaire M_1 de teneur en protéines des grains de blé. Le rectangle représente l'unité expérimentale (ici un site-année). Les cercles correspondent à des variables aléatoires. Les flèches continues indiquent une relation stochastique. Les flèches en pointillés correspondent à des relations déterministes.

Ici, les lois *a priori* des paramètres $\alpha_0, \alpha_1, \alpha_2$ sont définies comme des lois gaussiennes indépendantes d'espérance nulle et de variance égale à 10^6, soit $\alpha_k \sim N\left(0, 10^6\right)$, pour $k = 0$, 1 ou 2. La valeur de la variance de la loi gaussienne a été fixée à une valeur suffisamment grande pour qu'on puisse supposer que la distribution *a priori* n'apporte que peu d'information sur les valeurs des paramètres. Avec cette loi *a priori*, la probabilité que les paramètres $\alpha_0, \alpha_1, \alpha_2$ prennent des valeurs très éloignées de zéro (1000, par exemple) n'est pas négligeable. La loi gaussienne, définie sur $]-\infty, +\infty[$, n'est pas adaptée au paramètre σ^2, qui est toujours positif puisqu'il représente la variance des erreurs. La loi *a priori* de σ^2 doit être définie uniquement sur des valeurs réelles positives en utilisant, par exemple, une loi Uniforme, une loi Gamma ou une loi Inverse Gamma. Ici, nous supposons que $1/\sigma^2 \sim \text{Gamma}\left(10^3, 10^3\right)$. En pratique, il est souvent utile de faire une analyse de sensibilité des résultats aux choix des lois *a priori* afin de déterminer l'influence de ces choix sur les conclusions de l'étude.

Une fois les lois *a priori* choisies, la vraisemblance des paramètres doit être définie. Pour le modèle linéaire, elle se déduit de la loi des erreurs, qui sont

supposées ici indépendantes et distribuées selon une loi gaussienne de variance constante. Pour M_1, la vraisemblance est égale à l'expression déjà rencontrée avec l'approche classique :

$$V(\alpha_0, \alpha_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{43/2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{43}(y_i - \alpha_0 - \alpha_1 x_{1i})^2\Big).$$

La dernière étape consiste à déterminer la loi *a posteriori* des paramètres en combinant la loi *a priori* et la vraisemblance grâce au théorème de Bayes. Le calcul exact de la loi *a posteriori* n'est pas évident mais la vitesse de calcul actuelle des ordinateurs fait qu'il est maintenant possible d'estimer la loi *a posteriori* en réalisant un grand nombre de simulations avec des algorithmes adaptés, notamment avec les algorithmes MCMC. Ces algorithmes peuvent être mis en œuvre avec divers logiciels, notamment avec WinBUGS comme nous le présentons maintenant.

Un programme WinBUGS comporte trois fichiers :
- un fichier incluant les données ;
- un fichier incluant les valeurs initiales des paramètres pour l'algorithme MCMC ;
- un fichier décrivant le modèle.

Les trois fichiers utilisés pour estimer les paramètres du modèle M_2 (reliant la teneur en protéines à l'INN) sont présentés ci-dessous.

```
# Code BUGS

# Fichier Exemple1inn.odc presentant les mesures de teneur en
# proteines (extrait), d'INN (extrait) et le nombre total de
# mesures disponibles, N=43.

list(
  y=c(11.9 , 10.3 ,...),
  x2=c(1 , 0.8 ,...),
  N=43)

# Fichier IniExemple1.odc presentant des valeurs initiales pour
# les trois parametres du modele M2. Le vecteur alpha contient
# les valeurs initiales de \alpha_0 et \alpha_1, prec correspond
# a la precision, c'est-a-dire a (1/\sigma^2).

list(alpha=c(0, 0), prec= 1.0)

# Fichier ModelExemple1inn.odc decrivant le modele.

model
```

```
# y = Teneur en proteines
# x2 = Indice de nutrition azotee
# alpha = parametres de la regression

{

# Vraisemblance #

for (i in 1:N) {
  y[i] ~ dnorm(mu[i], prec)
  mu[i] <- alpha[1] + alpha[2] * x2[i] }

# Distribution a priori #

for (j in 1:2) {
  alpha[j] ~ dnorm(0.0, 1.0E-6) }
  prec ~ dgamma(0.001, 0.001) }
```

Dans le fichier `ModelExemple1inn.odc`, la vraisemblance est définie dans une boucle `for(i in 1:N) { ... }`. Il n'est pas nécessaire de fournir l'expression analytique de la vraisemblance, mais seulement de décrire la distribution des observations conditionnellement à α_0, α_1 et σ^2 (ici par une loi gaussienne). Les lois *a priori* des trois paramètres sont ensuite définies dans le programme comme des lois gaussiennes et gamma. Notez que, dans WinBUGS, ce sont les niveaux de précision des variables aléatoires qui sont considérés, c'est-à-dire l'inverse de leurs variances.

Les paramètres du modèle M_2 peuvent être estimés à l'aide d'un algorithme MCMC en utilisant les trois fichiers ci-dessus. Deux techniques sont possibles :

- lancement de l'algorithme avec l'interface graphique du logiciel ;
- utilisation d'un fichier script listant toutes les opérations à réaliser.

La deuxième technique permet de modifier facilement la séquence des opérations, par exemple de changer de fichier de données ou de générer un plus grand nombre de valeurs de paramètres. Le fichier script utilisé pour estimer les paramètres du modèle M_2 est présenté ci-dessous.

```
#Code BUGS

display(log)
set.seed(1)

check(f:/Projets/ModelExemple1inn.odc)
data(f:/Projets/Exemple1inn.odc)
compile(1)
inits(1, f:/Projets/IniExemple1.odc)
```

```
set(alpha)
set(prec)
beg(5000)
update(25000)

coda(alpha)

stats(*)
```

Les deux premières instructions du fichier script permettent d'afficher la fenêtre des résultats et d'initialiser le générateur de nombres pseudo-aléatoires. L'instruction `check` permet de vérifier la syntaxe du modèle, l'instruction `data` permet de lire les données et `inits` de lire les valeurs initiales. L'instruction `set` est utilisée pour déterminer le type de résultat qui sera conservé, ici α_0, α_2 et l'inverse de la variance des erreurs (la précision).

Deux paramètres nécessaires au fonctionnement de l'algorithme MCMC sont ensuite définis par `beg` et `update` : le nombre total de valeurs de paramètres générées par WinBUGS (25 000 dans notre cas) et le nombre de ces valeurs qui ne sont pas conservées. Ici, les 5 000 premières valeurs générées par l'algorithme sont éliminées et les 20 000 suivantes sont conservées. Il est important d'éliminer les premières valeurs générées car les algorithmes MCMC ont besoin d'un certain nombre d'itérations avant de converger vers la distribution *a posteriori*. Cette période d'instabilité est appelée la période « de chauffe ». La détermination du nombre total de valeurs de paramètres à générer et de la durée de la période de chauffe est un problème délicat. Diverses méthodes ont été proposées pour traiter ce problème, notamment l'approche de Brooks et Gelman (1998). Le principe est de générer plusieurs séries de valeurs de paramètres (par exemple trois) et de voir si les valeurs de ces séries sont très différentes en comparant la variabilité entre les séries à la variabilité à l'intérieur des séries. Si les séries sont trop différentes, il faut augmenter leur longueur. Le temps de calcul est également un critère à prendre en compte. Les techniques de diagnostic de convergence sont illustrées de façon plus détaillée dans les sections 3.2.2 et 3.3.2.

L'instruction `coda(alpha)` permet à l'utilisateur de récupérer l'ensemble des 20 000 valeurs des paramètres `alpha[1]` et `alpha[2]` générées après la période de chauffe, sous la forme de fichiers pouvant être sauvegardés. L'instruction `stat(*)` permet de résumer les calculs sous la forme de moyennes, médianes, variances et quantiles, pour tous les paramètres. Les résultats obtenus pour les paramètres α_0 et α_2 du modèle M_2 sont présentés sur la figure 3.3. Ils peuvent être utilisés pour estimer la loi *a posteriori* des paramètres. Des représentations graphiques (figures 3.3cd) permettent de visualiser les distributions de valeurs générées pour les paramètres.

Ces distributions peuvent être également résumées par leurs moyennes et variances (ou écarts types) comme dans le tableau 3.2. Les moyennes *a posteriori* sont ici très proches des estimations obtenues par les moindres carrés

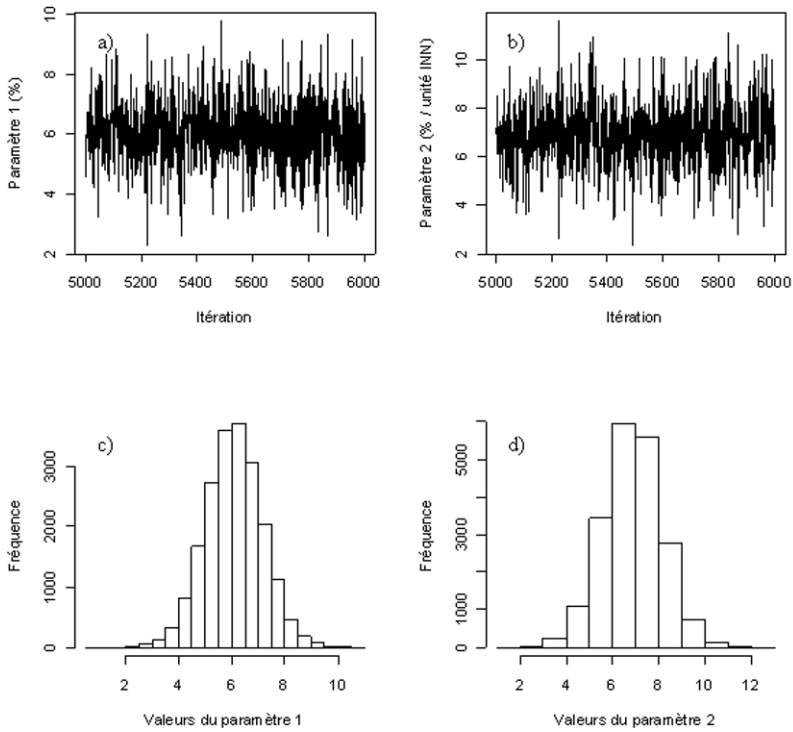

Figure 3.3 – Valeurs générées par l'échantillonneur de Gibbs pour les paramètres α_0 (paramètre 1) et α_2 (paramètre 2) du modèle M_2 de teneur en protéines. Les figures a) et b) présentent les 1 000 premières valeurs générées après une période de chauffe de 5 000 itérations. Les figures c) et d) présentent les histogrammes de 20 000 valeurs générées après la période de chauffe. Ces histogrammes constituent une approximation des distributions *a posteriori* des paramètres.

ordinaires (tableau 3.1). Cette grande similarité est due au fait que les lois *a priori* des paramètres ont été choisies de façon à apporter peu d'information.

Modèle	α_0		α_1		α_2		σ^2	
	Moy.	Et.	Moy.	Et.	Moy.	Et.	Moy.	Et.
M_1	6.71	2.38	0.113	0.052	-	-	1.81	0.42
M_2	6.09	1.07	-	-	6.89	1.27	1.15	0.27
M_3	3.93	1.96	0.056	0.043	6.42	1.30	1.13	0.27

Tableau 3.2 – Moyenne (Moy.) et écart type (Et.) des distributions *a posteriori* des paramètres de trois modèles linéaires pour la teneur en protéines du grain de blé. Calculs réalisés avec 20 000 itérations de l'échantillonneur de Gibbs après une période de chauffe de 5000 itérations.

Les écarts types des lois *a posteriori* présentés dans le tableau 3.2 n'ont pas la même signification que les écarts types des estimateurs des moindres carrés ordinaires présentés dans le tableau 3.1. Les écarts types des lois *a posteriori* sont en effet déterminés conditionnellement à la connaissance *a priori* et aux données disponibles. Ils quantifient un niveau d'incertitude dans les valeurs des paramètres qui tient compte à la fois des données utilisées pour l'estimation et du niveau de connaissance initial (lois *a priori*). Les écarts types des estimateurs des moindres carrés ordinaires, par contre, quantifient la variabilité des valeurs estimées qui résulterait d'un ré-échantillonnage des données.

Étape iv. Évaluation

La dernière étape consiste à évaluer les modèles obtenus après l'étape d'estimation des paramètres. L'objectif est de juger du réalisme des modèles, c'est-à-dire de déterminer si les équations utilisées et les hypothèses sur la distribution des erreurs donnent une description réaliste du système étudié. Lorsqu'un modèle est utilisé pour optimiser des décisions, il est également important d'évaluer la qualité des décisions prises grâce aux modèles. Ce deuxième aspect de l'évaluation sera traité dans le chapitre 4.

Étude des résidus Les analyses graphiques jouent un rôle important dans l'étape d'évaluation. Différentes analyses peuvent être réalisées, comme par exemple celles présentées dans la figure 3.4 pour les modèles M_1 et M_2 après estimation des paramètres par les moindres carrés ordinaires. L'analyse des résidus est particulièrement utile car elle permet d'évaluer le réalisme des hypothèses faites sur le modèle.

Une représentation graphique des résidus en fonction des variables explicatives permet de détecter l'existence d'un biais dans les prédictions des modèles, c'est-à-dire une surestimation ou une sous-estimation systématique pour une partie de la gamme de valeurs couverte par les variables d'entrée. L'existence d'un biais conduit souvent à une remise en cause des équations des modèles. La

figure 3.4cd ne révèle l'existence d'aucun biais et cette analyse ne remet donc pas en cause le caractère linéaire des modèles.

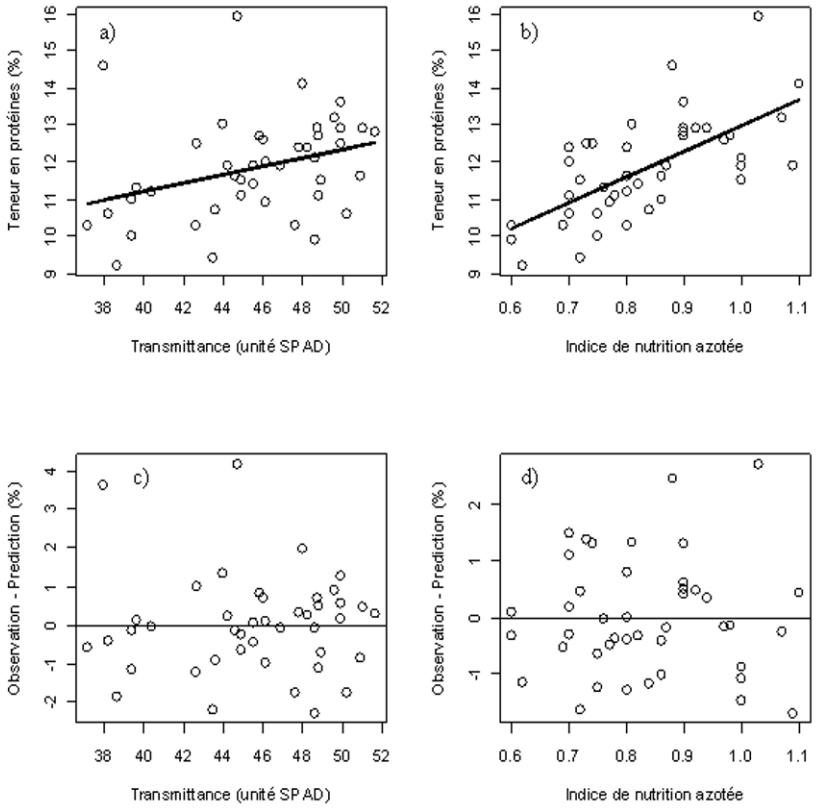

Figure 3.4 – Mesures de teneur en protéines en fonction de la transmittance (SPAD) (a) et de l'indice de nutrition azotée (INN) (b). Les droites présentent les valeurs prédites de teneurs en protéines obtenues avec les modèles M_1 et M_2 après estimation par les moindres carrés ordinaires. c) et d) présentent les résidus (observation – prédiction) en fonction des variables explicatives.

L'analyse des résidus peut également être utilisée pour vérifier la pertinence de l'hypothèse faite sur la variance des erreurs et sur leur indépendance. Dans notre exemple, nous avons supposé que la variance des erreurs est constante. La figure 3.4cd ne révèle aucune augmentation ou diminution de la gamme de variation des résidus, ce qui suggère que l'hypothèse d'une variance constante des erreurs (homoscédasticité) est réaliste.

La figure 3.4 a été obtenue avec le programme R suivant.

```
#Code R
```

```
par(mfrow = c(2,2))

plot(TAB$SPAD, TAB$TeneurPro,
                 xlab="Transmittance (unit\'e SPAD)",
                 ylab="Teneur en prot\'eines (%)", cex=1.4)
lines(TAB$SPAD, fitted(Mod.1), lwd=2)

plot(TAB$INN, TAB$TeneurPro,
                 xlab="Indice de nutrition azot\'ee",
                 ylab="Teneur en prot\'eines ({\%})", cex=1.4)
lines(TAB$INN, fitted(Mod.2), lwd=2)

plot(TAB$SPAD, residuals(Mod.1),
                 xlab="Transmittance (unit\'e SPAD)",
                 ylab="Observation - Prediction (%)", cex=1.4)
abline(0,0)

plot(TAB$INN, residuals(Mod.2),
                 xlab="Indice de nutrition azot\'ee",
                 ylab="Observation - Prediction (%)", cex=1.4)
abline(0,0)
```

L'instruction `par` permet de définir les caractéristiques de la fenêtre graphique de R. Ici, cette fenêtre a été découpée en quatre parties afin de faire quatre graphiques. Ces graphiques ont été réalisés avec l'instruction `plot`. Les deux premières instructions `plot` permettent de présenter les mesures de teneurs en protéines en fonction de SPAD et INN. L'instruction `lines` est utilisée pour superposer les valeurs ajustées `fitted(Mod.x)` aux données expérimentales. Les deux instructions `plot` suivantes permettent de présenter les résidus `residuals(Mod.x)`. Une ligne horizontale est ajoutée aux graphiques en utilisant l'instruction `abline`.

D'autres représentations graphiques que celles de la figure 3.4 peuvent être utiles, notamment la présentation de résidus normalisés ou studentisés (Venables et Ripley, 2002). Plusieurs fonctions R ont été développées dans ce but comme `plot.lm` et `qqnorm`. Des types particuliers de présentation graphique de résidus ont également été proposés pour les modèles bayésiens par Gelman *et al.* (2003).

L'analyse des résidus réalisée pour les modèles M_1, M_2 et M_3 ne remet pas en cause les hypothèses faites par ces modèles. Ce n'est pas toujours le cas. Bien souvent, l'analyse des résidus révèle l'existence de variances de l'erreur hétérogènes ou des erreurs corrélées. Dans ce type de situation, la méthode des moindres carrés ordinaires ne conduit pas à des estimateurs de variances minimales et d'autres méthodes doivent être envisagées. Ainsi, en statistique classique, une variance hétérogène (cas hétéroscédastique) peut être prise en

compte soit avec la méthode des moindres carrés pondérés, soit en paramétrant la variance des erreurs et en utilisant la méthode du maximum de vraisemblance. Ces méthodes peuvent être mises en œuvre avec la fonction `glm`de R (voir Hillier *et al.* 2005 pour un exemple d'application dans un contexte agronomique). Les corrélations des erreurs peuvent être prises en compte soit avec la méthode des moindres carrés généralisés, soit avec la méthode du maximum de vraisemblance. Nous reviendrons plus en détail sur la prise en compte des corrélations des erreurs dans la section 3.4.

Critères de qualité pour la sélection de modèles L'analyse des résidus constitue une étape importante de l'évaluation des modèles. Elles permet parfois de remettre en cause certaines hypothèses d'un modèle, mais elle n'est pas toujours suffisante pour en choisir un parmi plusieurs candidats. En pratique, il est souvent nécessaire d'avoir recours à des critères quantitatifs pour évaluer plusieurs modèles et choisir le plus adapté. De nombreux critères ont été proposés pour comparer les modèles linéaires (Miller, 2002). Le coefficient de détermination (R^2) est ainsi souvent calculé pour évaluer la qualité d'ajustement d'un ou plusieurs modèles à des données, mais son utilisation est sujette à caution. Le R^2 est défini par :

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

où N est le nombre de mesures disponibles et $\widehat{y}_i = \widehat{\alpha}_0 + \sum_{j=1}^{p} \widehat{\alpha}_j x_{ji}$ est la valeur prédite par le modèle, avec p le nombre de variables du modèle et $\widehat{\alpha}_0, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_p$ les valeurs estimées des paramètres. Le R^2 mesure la qualité d'ajustement d'un modèle aux données. Un ajustement parfait conduit à $R^2 = 1$. Si l'ajustement du modèle est équivalent à celui de la moyenne des mesures $\bar{y}$, le R^2 est nul. Ce critère est utile pour comparer des modèles ayant le même nombre de paramètres. Par contre, le R^2 conduira toujours à sélectionner le modèle incluant le plus grand nombre de paramètres lorsque les modèles candidats sont emboîtés, même si certains de ces modèles incluent des variables peu explicatives. Le R^2 n'est donc pas un critère pertinent pour comparer des modèles ayant différents niveaux de complexité.

L'évaluation de plusieurs modèles n'incluant pas le même nombre de paramètres peut être réalisée avec le critère d'information d'Akaike (AIC) (Akaike, 1973 ; Burnham et Anderson, 2002). L'AIC dépend de deux termes, le logarithme du maximum de vraisemblance et le nombre total de paramètres. Il est défini par :

$$\mathrm{AIC} = -2\log(\text{max.vraisemblance}) + 2q, \tag{3.2}$$

où q est le nombre de paramètres du modèle. L'AIC est une approximation de

l'espérance de la distance de Kullback-Leibler qui mesure un écart entre la distribution des valeurs simulées par le modèle et la distribution des observations.

Dans le cas d'un modèle linéaire défini par l'équation (3.1), avec erreurs indépendantes et de même loi gaussienne, l'AIC est défini à une constante près par :

$$\mathrm{AIC} = N \log\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right) + 2(p+1).$$

Le meilleur modèle selon le critère AIC est le modèle qui minimise l'AIC. La valeur de l'AIC sera d'autant plus faible que le modèle s'ajuste bien aux données (*i.e*, $\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2$ est faible) et inclut un petit nombre de paramètres (p faible). Le modèle sélectionné avec l'AIC ne sera donc pas nécessairement le modèle incluant le plus grand nombre de paramètres, contrairement à celui sélectionné avec le R^2. L'AIC est calculé par le logiciel R lors de l'appel de la fonction `summary`. La fonction `AIC()` peut également être utilisée pour obtenir la valeur de ce critère.

Plusieurs critères quantitatifs ont également été proposés pour comparer des modèles bayésiens (Gelman *et al.*, 2003), notamment le *Bayesian Information Criterion* (BIC) et le *Deviation Information Criterion* (DIC). Le DIC est calculé par WinBUGS. Comme l'AIC, il tient compte de la qualité d'ajustement du modèle aux données et de son niveau de complexité. En notant $\boldsymbol{\alpha}$ le vecteur des paramètres et $\mathbf{y}$ celui des observations, on définit tout d'abord la déviance par $D(\boldsymbol{\alpha}) = -2\log\big(p(\mathbf{y} \mid \boldsymbol{\alpha})\big)$. Notons que $p(\mathbf{y} \mid \boldsymbol{\alpha})$ est aussi la valeur de la vraisemblance $V(\boldsymbol{\alpha})$. Le DIC est défini par :

$$\mathrm{DIC} = \bar{D} + p_D \tag{3.3}$$

avec $\bar{D} = E_{\boldsymbol{\alpha}} D(\boldsymbol{\alpha})$ l'espérance de la déviance par rapport à la distribution *a posterori* des paramètres et p_D un terme de pénalisation. Le premier terme du DIC est une mesure moyenne de la qualité d'ajustement. Le deuxième terme de l'équation (3.3) est une mesure du niveau de complexité du modèle définie par $p_D = \bar{D} - D\left(\bar{\boldsymbol{\alpha}}\right)$, où $D\left(\bar{\boldsymbol{\alpha}}\right)$ correspond à la déviance obtenue lorsque les paramètres du modèle sont fixés à leur espérance *a posteriori*. Le terme p_D est lié au nombre de paramètres du modèle. En pratique, le premier terme du DIC est estimé par $\widehat{D}_{\mathrm{avg}} = \frac{1}{L}\sum_{l=1}^{L} D\left(\boldsymbol{\alpha}_l\right)$, où L est le nombre de valeurs de paramètres générées par MCMC et $D\left(\boldsymbol{\alpha}_l\right)$ est la déviance obtenue avec la l-ième valeur des paramètres générés par MCMC, $l = 1, \ldots, L$. Le deuxième terme peut être estimé à partir de la moyenne *a posteriori* des paramètres du modèle. Le DIC peut être calculé avec WinBUGS en modifiant la fin du fichier script de la façon suivante.

```
# Code BUGS

beg(5000)
```

```
update(5000)

dic.set()
update(20000)
stats(*)
dic.stats()
```

Ce programme initialise une variable avec l'instruction `dic.set()` après la période de chauffe. Les résultats basés sur les 20 000 itérations suivantes sont ensuite affichés à l'aide de l'instruction `dic.stats()`.

Les valeurs d'AIC, de DIC et de R^2 calculées avec R et WinBUGS sont présentées dans le tableau 3.3. Comme prévu, le modèle le plus complexe (M_3) a le R^2 le plus élevé. Les valeurs les plus satisfaisantes des critères AIC et DIC (c'est-à-dire les plus faibles) sont obtenues pour le modèle M_2, c'est-à-dire le modèle qui relie la teneur en protéines à l'INN. Le modèle le moins satisfaisant selon ces deux critères est le modèle M_1. Sur la base de ces résultats, c'est le modèle M_2 qui semble le plus pertinent. Les valeurs des paramètres estimées par les moindres carrés ordinaires et par MCMC étant très proches, le choix de la méthode d'estimation a ici peu de conséquences pratiques.

Modèle	AIC	DIC	R^2	W
M_1	149.22	149.41	0.11	$3\ 10^{-5}$
M_2	<u>129.74</u>	<u>129.93</u>	0.43	0.51
M_3	129.82	130.11	<u>0.46</u>	0.49

Tableau 3.3 – AIC, poids d'Akaike (W), DIC et R^2 des trois modèles linéaires de teneur en protéines du grain. L'AIC et le R^2 ont été calculés après avoir estimé les paramètres par les moindres carrés ordinaires. Le DIC a été calculé avec 20 000 itérations de l'échantillonneur de Gibbs après une période de chauffe de 5 000 itérations.

Mélange de modèles Le tableau 3.3 montre que les performances des modèles M_2 et M_3 sont très proches. La sélection du modèle M_2 peut donc paraître arbitraire. Une alternative consiste à mélanger les modèles candidats en tenant compte de leurs performances relatives. Différentes méthodes statistiques ont été proposées pour mélanger les modèles, fréquentistes ou bayésiennes (*e.g.*, Burnham et Anderson, 2002 ; Raftery *et al.*, 1997). Des études ont montré que l'utilisation d'une combinaison de modèles pouvait conduire à des prévisions plus précises que celles obtenues avec un seul modèle, même lorsque celui-ci est sélectionné avec une méthode statistique appropriée (Yuan and Yang, 2005).

Plusieurs méthodes de combinaison de modèles peuvent être mises en œuvre avec la librairie `MMIX` de R (disponible sur `www.r-project.org`). Nous présentons ici la méthode basée sur le critère AIC décrite en détail dans Burnham et Anderson (2002). Le principe est de calculer le poids d'Akaike W_k pour chaque modèle candidat k, $k = 1, \ldots, K$, puis de calculer la somme pondérée des

modèles. Le poids d'Akaike est compris entre zéro et un. Il est défini par :

$$W_k = \frac{\exp\Big(-0.5 \times (\mathrm{AIC}_k - \mathrm{AIC\,min})\Big)}{\sum_{k=1}^{K} \exp\Big(-0.5 \times (\mathrm{AIC}_k - \mathrm{AIC\,min})\Big)},$$

où AIC min est la valeur minimale d'AIC parmi les K modèles testés. Le poids d'Akaike d'un modèle est d'autant plus grand que son AIC est proche de celui du meilleur modèle, c'est-à-dire du modèle ayant le plus petit AIC. Les poids d'Akaike des trois modèles sont présentés dans le tableau 3.3. Le poids associé au modèle M_1 est très proche de zéro. Les poids des modèles M_2 et M_3 sont par contre proches de 0.5. La somme pondérée des trois modèles conduit au modèle suivant :

$$5.03 + 0.027x_1 + 6.66x_2.$$

Ce modèle correspond approximativement à la moyenne des modèles M_2 et M_3 car le poids de M_1 est négligeable. En pratique, cette approche peut être appliquée pour mélanger plusieurs dizaines voir plusieurs centaines de modèles (Prost *et al.*, 2008).

3.2 Modèle linéaire généralisé

3.2.1 Définition

Les modèles linéaires généralisés sont souvent utilisés pour analyser les observations catégorielles, par exemple des observations de présence/absence ou des comptages (*e.g.*, présence/absence d'une maladie, nombre d'insectes nuisibles dans une parcelle agricole). Comme le nom l'indique, ce type de modèle généralise le modèle linéaire. Dans un modèle linéaire généralisé (Agresti, 1990 ; McCulloch et Searle, 2001), il n'existe pas de relation linéaire directe entre Y et les variables d'entrée $X_1, ..., X_p$, mais il existe une transformation de l'espérance de la variable d'intérêt Y qui peut s'exprimer comme une combinaison linéaire des variables d'entrée. Cette transformation (par exemple $\log E(Y)$) s'appelle la *fonction de lien*, notée ici h.

Un modèle linéaire généralisé correspond à un modèle linéaire pour une fonction h de l'espérance μ de Y :

$$h(\mu) = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_j,$$

où $\mu = E(Y \mid X_1, ..., X_p)$ est l'espérance de Y conditionnellement aux variables d'entrée $X_1, ..., X_p$ et où $\alpha_0, \alpha_1, ..., \alpha_p$ sont des paramètres.

Comme dans un modèle linéaire, les variables d'entrée $X_1, ..., X_p$ peuvent être quantitatives ou qualitatives. Alors que dans un modèle linéaire, la variable aléatoire Y était une variable quantitative continue, le modèle linéaire généralisé peut être utilisé pour modéliser une variable Y discrète qui représente, par exemple, un nombre d'événements.

Régression logistique Parmi les différents modèles linéaires généralisés (Agresti, 1990), le modèle de régression logistique est particulièrement utile en agronomie pour modéliser des variables binaires Y ($y = 0$ ou $y = 1$). Avec ce modèle, la loi de Y est une loi de Bernoulli de paramètre μ, où $\mu = E(Y \mid X_1, ..., X_p)$ est égal à la probabilité que $y = 1$. La fonction de lien utilisée par ce modèle est la fonction *logit* définie par $h(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$. Avec cette fonction de lien, on a

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_j,$$

ce qui conduit à

$$\mu = \frac{\exp\left(\alpha_0 + \sum_{j=1}^{p} \alpha_j X_j\right)}{1 + \exp\left(\alpha_0 + \sum_{j=1}^{p} \alpha_j X_j\right)}.$$

Cette expression est comprise entre zéro et un. La fonction logit permet ainsi de modéliser l'effet des variables $X_1, ..., X_p$ sur la probabilité $\mu = P(y = 1)$ de façon que cette probabilité reste bien comprise entre zéro et un.

Dans le cas de la régression logistique et d'observations binaires 0-1, la vraisemblance peut s'écrire de la façon suivante. On suppose que l'on a effectué n_i observations de la variable binaire Y dans N situations i caractérisées par les valeurs $x_{i1}, ..., x_{ip}$ prises par les variables explicatives. Soit y_{ik} la k-ième valeur observée de Y dans la i-ième situation, pour $k = 1, \ldots, n_i$ et soit $n_i^+ = \sum_{k=1}^{n_i} y_{ik}$ le nombre d'observations y_{ik} égales à un. La vraisemblance, définie à partir de la loi binomiale, est proportionnelle à

$$V(\alpha_0, \ldots, \alpha_p) = \prod_{i=1}^{N} \mu_i^{n_i^+} (1 - \mu_i)^{n_i - n_i^+},$$
$$\text{où} \quad \mu_i = \frac{\exp\left(\alpha_0 + \sum_{j=1}^{p} \alpha_j x_{ij}\right)}{1 + \exp\left(\alpha_0 + \sum_{j=1}^{p} \alpha_j x_{ij}\right)} \tag{3.4}$$

représente la probabilité que $y = 1$ dans la situation i, pour $i = 1, \ldots, N$. Les paramètres $\alpha_0, \alpha_1, ..., \alpha_p$ peuvent être estimés par la méthode du maximum de vraisemblance qui, dans le cas de la régression logistique, consiste à trouver les valeurs de $\alpha_0, \alpha_1, ..., \alpha_p$ maximisant (3.4).

En statistique bayésienne, les paramètres du modèle $(\alpha_0, \alpha_1, ..., \alpha_p)$ correspondent à des variables aléatoires et l'estimation a pour objectif de déterminer la loi *a posteriori* des paramètres, c'est-à-dire la distribution de $\alpha_0, \alpha_1, ..., \alpha_p$ conditionnellement aux données disponibles. Les calculs sont réalisés à partir de la loi *a priori* définie par l'utilisateur et d'une fonction de vraisemblance calculée à partir des données comme, par exemple, la fonction définie par l'équation (3.4) dans le cas de la régression logistique.

Modèle Poisson log-linéaire D'autres fonctions de lien que la fonction logit sont utiles, notamment la fonction log qui est souvent utilisée en combinaison avec une loi de Poisson dans le cadre du modèle Poisson log-linéaire. Ce modèle permet d'analyser des comptages mesurant le nombre d'occurrences d'un événement. Il est défini par

$$\log(\mu) = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_j \quad \text{et} \quad P(Y = m \mid \mu) = \exp(-\mu)\frac{\mu^m}{m!},$$

où m est un nombre entier représentant la valeur observée du comptage et $\mu = E(Y)$ est le paramètre d'intensité de la loi de Poisson.

3.2.2 Exemple : présence/absence d'oiseaux dans une prairie

Certaines espèces d'oiseaux constituent des indicateurs de biodiversité. Leur présence peut également présenter un atout touristique important. La présence

ou l'absence d'oiseaux dans les prairies sont déterminées par plusieurs facteurs, notamment par les caractéristiques de leur habitat comme l'intensité du pâturage ou la surface en eau (Milsom *et al.*, 2000 ; Tichit *et al.*, 2005 ; Makowski *et al.*, 2009). Nous nous intéressons ici à la modélisation de l'effet d'une variable caractérisant la prairie, la hauteur d'herbe, sur la présence d'une espèce d'oiseau dans les prairies du Marais poitevin en France. L'espèce d'oiseau considérée ici est le chevalier gambette (*Tringa totanus*), un échassier vivant dans les prairies humides et dans les marais. Un modèle logistique est développé pour calculer la probabilité de présence de cet oiseau en fonction de la hauteur d'herbe.

Étape *i.* Définition des variables

La variable d'intérêt est la présence ou l'absence de l'oiseau dans une prairie. Cette variable peut être définie comme une variable binaire Y ayant deux réalisations possibles, $y = 1$ si l'oiseau est présent dans la prairie et $y = 0$ sinon. Le modèle ne comporte qu'une variable d'entrée, X, qui est la hauteur moyenne d'herbe de la prairie (en cm). La variable X est quantitative continue.

Étape *ii.* Définition des équations

La variable Y est supposée suivre une loi de Bernoulli de paramètre μ, $Y \sim \text{Bern}(\mu)$. Ceci revient à dire que $y = 1$ avec la probabilité μ. Cette probabilité est reliée à la variable X à l'aide d'un modèle linéaire généralisé de type régression logistique. Son expression est définie par :

$$\text{logit}(\mu) = \alpha_0 + \alpha_1 X$$

soit

$$\mu = E(Y|X) = \frac{\exp(\alpha_0 + \alpha_1 X)}{1 + \exp(\alpha_0 + \alpha_1 X)}, \tag{3.5}$$

où μ est la probabilité que l'oiseau soit présent lorsque la hauteur d'herbe est égale à X et α_0, α_1 sont deux paramètres à estimer.

Étape *iii.* Estimation des paramètres

Données La base de données est constituée de 424 sites-années correspondant à des prairies localisées dans le Marais poitevin en France. Une mesure moyenne de hauteur d'herbe (moyenne de 50 mesures par prairie, soit environ 20 mesures par hectare) et une mesure de présence/absence de chevalier gambette ont été réalisées sur chaque site-année. Les mesures de présence/absence ont été réalisées au stade « incubation » de l'oiseau selon la méthode de Milsom *et al.* (2000) (observation de chaque prairie à la jumelle binoculaire ou au télescope tous les 10 jours à une distance généralement supérieure à 150 m).

Les données disponibles sont décrites dans le tableau 3.4. La présence d'oiseau a été constatée sur 51 sites-années parmi les 424 disponibles. Les mesures

État y	Hauteur d'herbe x (cm)						
	Effectif	Min.	1er quartile	Médiane	Moyenne	3e quartile	Max.
0	373	5.9	19.00	27.74	28.33	36.80	57.20
1	51	4.9	15.35	23.80	24.66	34.60	43.75

Tableau 3.4 – Description des hauteurs d'herbe (x) en présence ($y = 1$) ou en absence ($y = 0$) d'oiseaux dans une prairie.

de hauteur d'herbe montrent que la hauteur avait tendance à être plus faible dans les situations où l'oiseau était présent.

Les valeurs présentées dans le tableau 3.4 ont été obtenues avec le programme R suivant :

```
# Code R

TAB <- read.table("f:\\Projets\\Exemple2\\CGIhabitat.txt",
                  header=T)
HH   <- TAB$HmoyM3
Pres <- TAB$CG_incub

summary(HH[Pres==0])
summary(HH[Pres==1])

length(HH[Pres==0])
length(HH[Pres==1])
```

L'instruction `read.table` permet de lire le fichier `CGIhabitat.txt` et de placer son contenu dans un objet R appelé `TAB`. Les deux lignes suivantes permettent de renommer les variables initiales en utilisant des noms plus courts, `HH` (hauteur d'herbe) et `Pres` (présence/absence d'oiseau). Les instructions `summary` et `length` sont utilisées pour résumer les valeurs du vecteur `HH` et calculer la longueur de ce vecteur, c'est-à-dire le nombre de mesures disponibles. Le vecteur est découpé en deux morceaux : une partie correspond aux situations où l'oiseau est absent (`HH[Pres==0]`) et une partie correspond aux situations où l'oiseau est présent (`HH[Pres==1]`).

Maximum de vraisemblance Les valeurs des paramètres sont estimées par les valeurs de α_0 et α_1 qui maximisent la vraisemblance. Celle-ci est, dans notre cas, proportionnelle à

$$\prod_{i=1}^{424} \mu_i^{y_i} \left(1-\mu_i\right)^{1-y_i} \quad \text{avec} \quad \mu_i = \frac{\exp\left(\alpha_0 + \alpha_1 x_i\right)}{1+\exp\left(\alpha_0 + \alpha_1 x_i\right)},$$

où x_i est la i-ième mesure de hauteur d'herbe et y_i est la i-ième observation de présence/absence. Cette expression de la vraisemblance est obtenue en supposant que Y_i suit une loi de Bernoulli de paramètre μ_i, $Y_i \sim \text{Bern}(\mu_i)$, $i = 1, \ldots, 424$. La maximisation de la vraisemblance peut être réalisée à l'aide des instructions R suivantes.

```
# Code R

Data <- data.frame(HH,Pres)

Mod.Log <- glm(Pres ~ HH, data=Data, family=binomial)
summary(Mod.Log)
```

La première ligne du programme ci-dessus crée un tableau de données R à partir des vecteurs `HH` et `Pres` créés précédemment. L'instruction `glm` est ensuite utilisée pour appliquer la méthode du maximum de vraisemblance et estimer les paramètres. Le type `logit` est indiqué par défaut par `family=binomial`. Les résultats fournis par `summary` donnent $\widehat{\alpha}_0 = -1.16$ et $\widehat{\alpha}_1 = -0.031$. Les écarts types des estimateurs sont respectivement égaux à 0.39 et 0.01 pour $\widehat{\alpha}_0$ et $\widehat{\alpha}_1$. L'instruction `summary` permet également d'obtenir les résultats d'un test sur les paramètres. Ceux-ci sont significativement différents de zéro au niveau 5%. Le signe négatif de $\widehat{\alpha}_1$ indique que la probabilité de présence est une fonction décroissante de la hauteur d'herbe, ce qui est cohérent avec les données (tableau 3.4).

Une fois les paramètres estimés, le modèle peut être utilisé pour prédire la probabilité de présence pour toute valeur x de hauteur d'herbe dans la gamme des valeurs représentées dans les données. Cette probabilité est égale à

$$\frac{\exp(-1.16 - 0.031x)}{1 + \exp(-1.16 - 0.031x)}.$$

La courbe reliant la probabilité de présence d'oiseau à la hauteur d'herbe est présentée sur la figure 3.5. Plus la hauteur d'herbe est élevée, plus la probabilité de présence est faible. Cette probabilité ne dépasse cependant jamais 0.25.

Méthode d'estimation bayésienne Dans le cadre bayésien, les paramètres α_0 et α_1 sont des variables aléatoires. La version bayésienne du modèle logistique est présentée sur la figure 3.6. L'estimation des paramètres est réalisée en trois étapes, comme dans le cas du modèle linéaire :

- définition des distributions *a priori* des paramètres ;
- définition de la vraisemblance ;
- génération d'un échantillon de valeurs de paramètres à l'aide d'une méthode MCMC.

Les distributions *a priori* des paramètres sont définies ici comme des lois gaussiennes indépendantes d'espérance nulle et de variance égale à 10^6, $\alpha_j \sim N\left(0, 10^6\right)$, $j = 0, 1$. La variance de la loi gaussienne a été fixée à une valeur

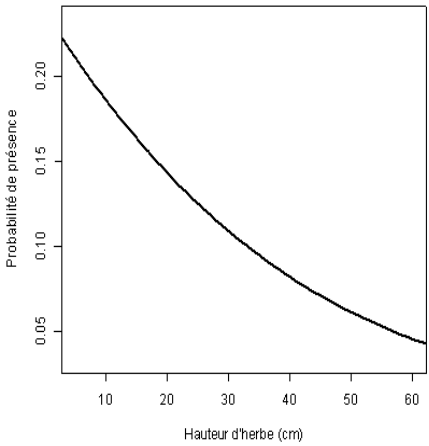

Figure 3.5 – Probabilité de présence d'oiseau en fonction de la hauteur d'herbe calculée à l'aide d'un modèle logistique. Paramètres estimés par maximum de vraisemblance.

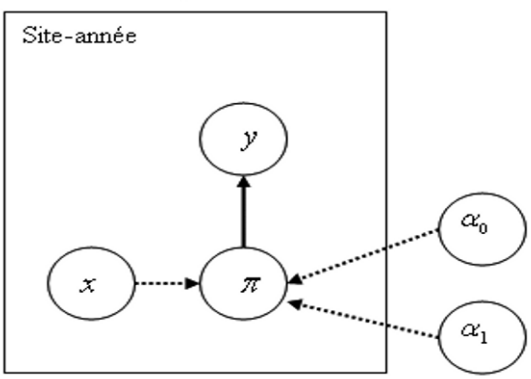

Figure 3.6 – Représentation graphique du modèle logistique sur la probabilité de présence d'oiseaux. Le rectangle représente l'unité expérimentale (ici une prairie). Les cercles correspondent à des variables aléatoires. Les flèches continues indiquent une relation stochastique. Les flèches en pointillés correspondent à des relations déterministes.

suffisamment grande pour qu'on puisse supposer que la distribution *a priori* n'apporte que peu d'information sur les valeurs des paramètres.

La deuxième étape consiste à définir la vraisemblance, c'est-à-dire la probabilité d'observer les données conditionnellement aux paramètres. La vraisemblance est identique à celle utilisée pour l'approche classique fréquentiste. Elle est basée sur la loi de Bernoulli, $Y_i \sim \text{Bern}(\mu_i)$, $i = 1, \ldots, 424$.

Finalement, le logiciel WinBUGS a été utilisé pour générer un échantillon de valeurs des deux paramètres du modèle (3.5) à l'aide de l'échantillonneur de Gibbs. Comme dans l'exemple présenté dans la section 3.1.5, la méthode d'estimation a été appliquée en utilisant trois fichiers WinBUGS, un fichier qui contient les données, un fichier qui inclut les valeurs initiales des paramètres et un fichier décrivant le modèle. Ces fichiers sont présentés ci-dessous.

Fichier de données (extraits) :

```
# Code BUGS

list(
hh=c(48.8 , 15.7 ,...),
pres=c(0 , 0 ,...),
N=424)
```

Fichier d'initialisation des paramètres :

```
list(alpha=c(0,0))
```

Fichier décrivant le modèle :

```
# Code BUGS

model

# pres= 1 si oiseau (chevalier gambette incubation)
#       present, 0 sinon,
# hh=hauteur d'herbe en cm
# alpha=parametres, theta=proba de presence

{
 for (i in 1:N) {
   pres[i] ~ dbern(theta[i])
   logit(theta[i]) <- mu[i]
   mu[i] <- alpha[1] + alpha[2]*hh[i]
   }

 ## Prior ##
 for (j in 1:2) {
```

```
   alpha[j]  ~   dnorm(0.0,1.0E-6)
   }
}
```

L'estimation a été réalisée avec le fichier script suivant :

```
# Code BUGS

display(log)
set.seed(1)

check(f:/Projets/Exemple2/ModelExemple2.odc)
data(f:/Projets/Exemple2/DataExemple2.odc)
compile(1)
inits(1,f:/Projets/Exemple2/IniExemple2.odc)

set(alpha)

beg(5000)
thin.samples(15)
update(5000)

dic.set()
update(20000)

coda(alpha)
stats(*)
dic.stats()
```

Ce fichier script est très proche de celui utilisé pour le modèle linéaire présenté dans la section 3.1.5. Une période de chauffe de 5 000 itérations est spécifiée, puis 20 000 valeurs de paramètres sont générées par l'échantillonneur de Gibbs. La durée de la période de chauffe et le nombre total de valeurs générées ont été déterminés par essai-erreur, en lançant plusieurs chaînes. Seul le fichier script final est présenté ici. Notez qu'une instruction a été ajoutée par rapport au fichier script de la section 3.1.5, l'instruction `thin-samples`. Cette instruction permet de ne garder qu'une fraction des valeurs générées par l'algorithme, ici une sur quinze. Cette procédure est utile lorsque des auto-corrélations sont détectées entre les valeurs successives générées par l'algorithme. Les tests essais-erreurs préliminaires ont révélé l'existence de telles corrélations. Le fait de ne garder qu'une valeur de paramètre sur quinze réduit les auto-corrélations à un niveau quasi nul.

Les valeurs générées ont été utilisées pour déterminer la distribution *a posteriori* des paramètres α_0 et α_1. Ces distributions sont présentées graphiquement sur la figure 3.7ab. La moyenne et l'écart type des valeurs générées de α_0 (figure 3.7a) sont égaux à -1.17 et 0.39 respectivement. Pour α_1, la moyenne

et l'écart type (figure 3.7b) sont égaux à -0.032 et 0.015 respectivement. Les valeurs moyennes obtenues avec la méthode bayésienne sont très proches des valeurs estimées par maximum de vraisemblance ($\widehat{\alpha}_0 = -1.16$ et $\widehat{\alpha}_1 = -0.031$). Les distributions *a posteriori* des paramètres peuvent être utilisées pour générer la distribution *a posteriori* de la probabilité de présence d'oiseau en fonction de la hauteur d'herbe. Celle-ci est présentée graphiquement sur la figure 3.7c. Chaque courbe de cette figure a été obtenue avec une paire de valeurs de α_0 et α_1 issues des distributions présentées sur les figures 3.7a et 3.7b. La courbe moyenne est présentée sur la figure 3.7d.

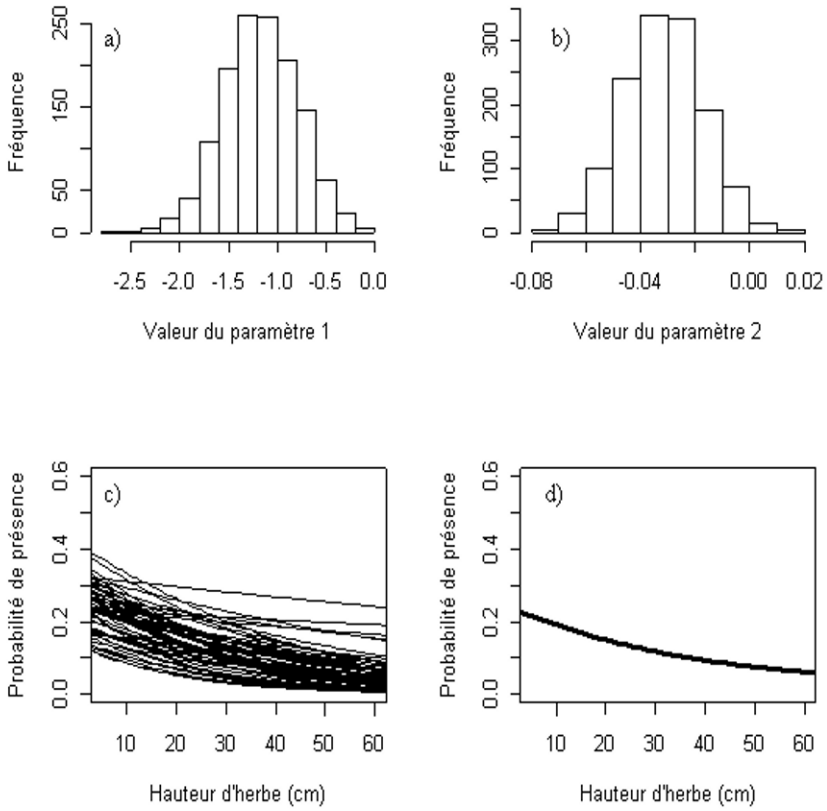

Figure 3.7 – Résultat de l'estimation bayésienne pour le modèle de probabilité de présence d'oiseaux. Histogrammes des valeurs générées par MCMC pour les paramètres α_0 (a) et α_1 (b). Distribution des courbes de réponse (échantillon de 60 courbes) (c) et réponse moyenne (d).

Étape *iv.* Évaluation

L'étape d'évaluation est moins déterminante ici que dans l'exemple de la section 3.1.5, car un seul modèle logistique a été développé. Il est cependant utile de comparer le modèle logistique à un modèle plus simple, qui ne prend en compte aucune variable explicative. Cette comparaison est réalisée ici en calculant les valeurs des critères AIC et DIC déjà utilisés dans le cas du modèle linéaire (section 3.1.5).

L'AIC est un critère d'évaluation souvent utilisé pour évaluer les modèles linéaires généralisés dont les paramètres ont été estimés par maximum de vraisemblance (Harrel, 2001). Voir Agresti (1990) et Harrel (2001) pour d'autres techniques permettant d'évaluer les modèles de régression logistique dans le cadre de la statistique classique.

L'AIC d'un modèle linéaire généralisé a la même définition que l'AIC d'un modèle linéaire (équation (3.2)). Ici, ce critère a été calculé pour le modèle logistique incluant la variable d'entrée hauteur d'herbe et sa valeur a été comparée à la valeur de l'AIC obtenue pour un modèle plus simple n'incluant pas cette variable d'entrée. Ce deuxième modèle prédit une probabilité de présence constante, indépendante de la hauteur d'herbe.

Les valeurs d'AIC peuvent être obtenues avec l'instruction R `summary` après avoir utilisé `glm`. Dans notre cas, l'AIC du modèle n'incluant pas la hauteur d'herbe est égale à 313.63, alors que l'AIC du modèle incluant cette variable est égale à 310.60. Le modèle incluant la hauteur d'herbe est donc meilleur selon ce critère.

La version bayésienne du modèle logistique peut être évaluée en comparant la valeur du DIC (équation (3.3)) du modèle incluant la hauteur d'herbe à la valeur du DIC du modèle sans cette variable d'entrée. Le DIC peut être calculé à l'aide des instructions `dic.set()` et `dic.stats()` de WinBUGS (voir le fichier script ci-dessus). Dans notre cas, le DIC du modèle qui tient compte de la hauteur d'herbe est égal à 310.57 et celui du modèle sans cette variable d'entrée est égal à 313.62.

Ces résultats confirment que la hauteur d'herbe a une influence sur la probabilité de présence du chevalier gambette au stade incubation. Plus la hauteur d'herbe est élevée, moins il y a de chance que le chevalier gambette soit présent. Cependant, cette variable n'explique qu'une partie de la variabilité de la présence/absence de l'oiseau. La figure 3.7 montre bien qu'il existe une variabilité résiduelle importante qui pourrait probablement être expliquée en ajoutant d'autres variables au modèle telles que le nombre d'animaux d'élevage présents sur la prairie ou la présence d'un étang à proximité.

3.3 Modèle non linéaire

3.3.1 Définition

Il arrive parfois qu'une variable d'intérêt Y ne puisse pas être reliée de manière réaliste à des variables explicatives $X_1, ..., X_p$ à l'aide d'un modèle linéaire ou linéaire généralisé. Dans ce cas, il est possible d'utiliser un modèle non linéaire (Seber et Wild, 2003) défini par :

$$Y = f(\mathbf{X}, \alpha) + \varepsilon \tag{3.6}$$

où $\mathbf{X} = (X_1, ..., X_p)^T$ est le vecteur des p variables d'entrée, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_q)^T$ est un vecteur incluant les q paramètres du modèle (^{T} correspond à l'opérateur transposé), ε est un terme d'erreur aléatoire et f est une fonction mathématique. Les variables d'entrée $X_1, ..., X_p$ peuvent être de diverses natures, quantitative ou qualitative.

Le modèle (3.6) est qualifié de non linéaire lorsque $f(\mathbf{X}, \boldsymbol{\alpha})$ n'est pas une combinaison linéaire des paramètres et ne peut pas se ramener à une combinaison linéaire en utilisant les fonctions de liens définies pour le modèle linéaire généralisé. Un des intérêts des modèles non linéaires est que leurs paramètres ont souvent une interprétation intéressante d'un point vue physique ou biologique.

Par exemple, le modèle de croissance logistique, ou modèle de Verhulst, est défini par $f(X_1, \boldsymbol{\alpha}) = \alpha_0\alpha_1/\big(\alpha_0 + (\alpha_1 - \alpha_0)\exp(-\alpha_2 X_1)\big)$. Il possède $p = 1$ variable explicative X_1 et $q = 3$ paramètres. Entre autres propriétés, les paramètres $\alpha_0, \alpha_1, \alpha_2$ correspondent respectivement à la valeur de f quand $x_1 = 0$, à la limite de f quand x_1 tend vers $+\infty$, et à la dérivée au point d'inflexion. Dans un modèle décrivant la croissance d'une population en fonction du temps (avec densité-dépendance), ils représentent la taille initiale de la population, sa taille maximale et son taux de croissance maximal.

Comme pour le modèle linéaire, le terme d'erreur est défini comme une variable aléatoire centrée : $E(\varepsilon) = 0$. Cela signifie que $f(\mathbf{X}, \boldsymbol{\alpha})$ représente la valeur moyenne de Y pour des valeurs de $X_1, ..., X_p$ données. La distribution de ε est souvent supposée être une loi gaussienne d'espérance nulle et de variance σ^2. Cette distribution est alors notée $\varepsilon \sim N\left(0, \sigma^2\right)$. L'ensemble des paramètres du modèle de l'équation (3.6) est, dans ce cas, composé de $\alpha_1, ..., \alpha_q$, σ^2.

Si $\varepsilon \sim N\left(0, \sigma^2\right)$, si les erreurs sont indépendantes et si N observations $(y_i, x_{1i}, ..., x_{pi})$, $i = 1, \ldots, N$ sont disponibles, la vraisemblance est définie par

$$V(\boldsymbol{\alpha}, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[y_i - f\left(\mathbf{x_i}, \boldsymbol{\alpha}\right)\right]^2\right\}. \tag{3.7}$$

L'équation (3.7) correspond au produit des N densités gaussiennes associées aux N observations. Comme pour les modèles linéaire et linéaire généralisé, l'estimation peut être réalisée avec la méthode du maximum de vraisemblance. Cependant, comme nous le verrons plus loin, la mise en œuvre de cette méthode

d'estimation est plus délicate avec un modèle non linéaire, même si le principe est identique.

Dans le cadre bayésien, comme dans le cas du modèle linéaire et du modèle linéaire généralisé, deux lois de probabilité sont définies pour les paramètres $\boldsymbol{\alpha}$ et σ^2, une loi *a priori* et une loi *a posteriori*. Les calculs sont réalisés à partir de la loi *a priori* définie par les experts du domaine considéré et d'une fonction de vraisemblance calculée à partir des données comme, par exemple, la fonction définie par l'équation (3.7).

Remarque 3.1 *Dans de nombreuses situations, il n'est pas réaliste de définir une valeur unique σ^2 pour toutes les observations, par exemple lorsque certaines observations sont moins précises que d'autres. Les erreurs peuvent également être corrélées et il est alors nécessaire de définir une loi de probabilité plus complexe incluant un nombre plus grand de paramètres pour décrire la distribution de ε. Les algorithmes permettant d'ajuster des modèles non linéaires donnent parfois la possibilité de modéliser également la variance ou les covariances des observations en fonction des variables explicatives X_i ou en fonction de $f(\mathbf{X}, \boldsymbol{\alpha})$ (voir par exemple Huet et al., 2004).*

3.3.2 Exemple : reliquat d'azote dans le sol à la récolte

Les pratiques de gestion des engrais azotés ont un effet sur les risques de pollution de l'eau par les nitrates (Meynard *et al.*, 2002 ; Delgado *et al.*, 2006). Des doses d'engrais trop importantes ou des dates d'apport inadaptées peuvent être à l'origine d'une augmentation de la quantité d'azote minéral présente dans le sol à la récolte et ainsi accroître la teneur en nitrate de l'eau de percolation. L'objectif de ce cas d'étude est de modéliser l'effet de la dose totale d'engrais minéral appliquée sur une culture de blé d'hiver (*Triticum aestivum* L.) sur la quantité d'azote minéral résiduelle, c'est-à-dire l'azote minéral présent dans le sol à la récolte et donc susceptible d'augmenter la teneur en nitrate de l'eau.

Étape *i.* Définition des variables

La variable d'intérêt est la quantité d'azote minéral présent dans le sol (0-90 cm) à la récolte, appelée « reliquat N récolte ». C'est une variable quantitative continue notée Y et exprimée en kg d'azote par hectare. Le modèle ne comporte qu'une variable d'entrée, X, qui correspond à la dose totale d'engrais azoté apportée sous forme minérale durant la culture du blé, appelée « dose d'engrais N ». La variable X est une variable quantitative continue exprimée également en kg d'azote par hectare.

Étape *ii.* Définition des équations

Deux équations proposées par Jauregui et Paris (1985) et Makowski *et al.* (1999) sont considérées ici pour relier le reliquat N récolte (R) à la dose (X) : une équation plateau-plus-linéaire et une équation plateau-plus-quadratique.

Dans les deux équations, on suppose que le reliquat est constant jusqu'à ce que la dose appliquée dépasse un certain seuil. Au-delà du seuil, le reliquat augmente. Les deux équations ne représentent pas de la même façon la phase d'accroissement du reliquat récolte.

L'équation plateau-plus-linéaire est définie par :

$$\begin{cases} R = \mathrm{R_{min}} & \text{si } X < \mathrm{X_{min}}; \\ R = \mathrm{R_{min}} + A\,(X - \mathrm{X_{min}}) & \text{sinon.} \end{cases}$$

L'équation plateau-plus-quadratique est définie par :

$$\begin{cases} R = \mathrm{R_{min}} & \text{si } X < \mathrm{X_{min}}; \\ R = \mathrm{R_{min}} + A\,(X - \mathrm{X_{min}})^2 & \text{sinon.} \end{cases}$$

Ces deux équations incluent trois paramètres à estimer, $\boldsymbol{\alpha} = (\mathrm{R_{min}}, A, \mathrm{X_{min}})^T$. Le paramètre $\mathrm{R_{min}}$ correspond à la valeur minimale de reliquat récolte, A est un paramètre qui détermine l'augmentation du reliquat en fonction de la dose et $\mathrm{X_{min}}$ est la dose seuil au-delà de laquelle le reliquat augmente.

On pose le modèle $Y = R + \varepsilon$, avec $R = f\,(X, \boldsymbol{\alpha})$ défini par l'équation plateau-plus-linéaire ou l'équation plateau-plus-quadratique et les erreurs supposées indépendantes, centrées et de même variance σ^2.

Étape *iii.* Estimation des paramètres

Données Des observations ont été recueillies sur trois parcelles expérimentales en 1994 et cinq en 1995, soit huit sites-années. Les parcelles élémentaires mesuraient 2.3 m sur 10 m, 3 m sur 12 m ou 3 m sur 24 m. Sur chaque site-année, six à sept doses d'engrais comprises entre 0 et 280 kg/ha ont été appliquées. Leur répartition avait été déterminée préalablement selon un plan en blocs complets randomisés, avec quatre blocs complets.

Le reliquat d'azote minéral dans le sol a été mesuré pour chaque dose appliquée et chaque site-année. Le nombre total de mesures disponibles est égal à 52. Chaque mesure de reliquat a été obtenue à partir de quatre ou cinq échantillons de sol par site-année, prélevés sur une profondeur de 0-90 cm. L'azote minéral total ($\mathrm{NO_3^-}$ et $\mathrm{NH_4^+}$) a été mesuré sur chaque échantillon puis les valeurs ont été moyennées. Les mesures de reliquat sont comprises entre 19 et 83.6 kg/ha. Voir Makowski *et al.* (1999) pour plus de détail. Les données sont présentées sur la figure 3.8.

Moindres carrés ordinaires Le principe de la méthode est de calculer la valeur de $\boldsymbol{\alpha} = (\mathrm{R_{min}}, A, \mathrm{X_{min}})^T$ qui minimise la somme des carrés des écarts entre les mesures de Y et les valeurs calculées par le modèle. Cela revient à calculer la valeur de $\boldsymbol{\alpha}$ qui minimise

$$\mathrm{SCE} = \sum_{i=1}^{52} \Big(y_i - f\,(x_i, \boldsymbol{\alpha})\Big)^2,$$

avec y_i la i-ième mesure de reliquat récolte et x_i la i-ième dose d'engrais, pour $i = 1, \ldots, 52$. Cette valeur correspond à l'estimation des moindres carrés ordinaires et elle est notée $\widehat{\boldsymbol{\alpha}} = \left(\widehat{\mathrm{R}}_{\min}, \widehat{A}, \widehat{\mathrm{X}}_{\min}\right)^T$. La variance des erreurs σ^2 peut être estimée par la variance résiduelle $\widehat{\sigma}^2 = \frac{1}{52-3} \sum_{i=1}^{52} \left(y_i - f(x_i, \widehat{\alpha})\right)^2$.

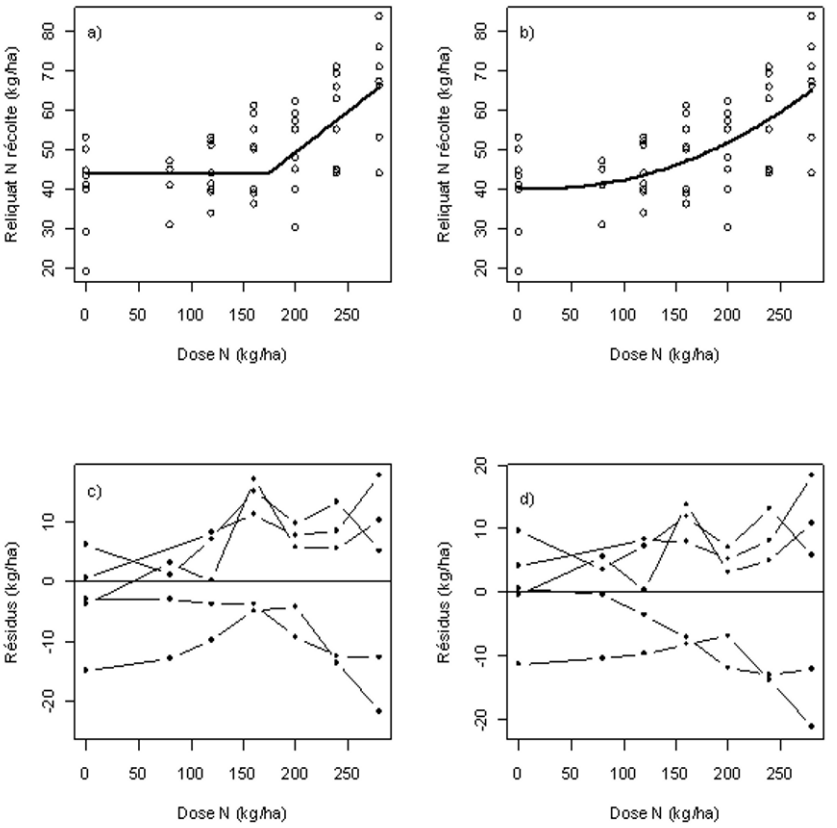

Figure 3.8 – Valeurs observées et prédites de reliquat récolte avec le modèle plateau-plus-linéaire (a, c) et avec le modèle plateau-plus-quadratique (b, d). Les résidus appartenant à un même site-année sont reliés entre-eux (c, d). Seulement cinq des huit sites-années sont présentés. Les paramètres ont été estimés par les moindres carrés ordinaires.

Lorsque la distribution des résidus est définie par $\varepsilon \sim N\left(0, \sigma^2\right)$ et que les erreurs sont indépendantes, la minimisation de SCE est équivalente à la maximisation de la vraisemblance. Le calcul de l'estimation des moindres carrés ordinaires est plus délicat pour un modèle non linéaire que pour un modèle linéaire. Avec un modèle linéaire, cette estimation a une expression analytique

connue et peut être calculée à partir des données disponibles par du calcul matriciel. Ce n'est pas le cas avec un modèle non linéaire. Avec ce type de modèle, il est en effet nécessaire d'utiliser un algorithme itératif (*e.g.*, Gauss-Marquardt) démarrant d'une valeur initiale fournie par l'utilisateur. La valeur optimale des paramètres est alors approchée en plusieurs étapes successives.

Plusieurs algorithmes itératifs ont été développés pour les modèles non linéaires comme, par exemple, l'algorithme de Gauss-Newton ou celui de Gauss-Marquardt (*e.g.* Seber et Wild, 2003). D'un point de vue pratique, les deux différences essentielles par rapport au modèle linéaire sont que : *i)* l'utilisateur doit fournir une valeur initiale pour chaque paramètre qu'il souhaite estimer, *ii)* il doit s'assurer que l'algorithme a bien convergé vers une solution. Si les valeurs initiales sont trop éloignées de l'optimum ou si le modèle est inadapté, l'algorithme aura des difficultés à converger vers la bonne valeur.

La méthode des moindres carrés peut être appliquée à un modèle non linéaire à l'aide de la fonction `nls` de R (Venables et Ripley, 2002) et avec la fonction `nls2` de la librairie du même nom (Huet *et al.*, 2004). Le programme R utilisé pour estimer les paramètres des modèles plateau-plus-linéaire et plateau-plus-quadratique avec `nls` est présenté ci-dessous en trois parties : lecture des données, définition des fonctions f, estimation.

```
# Code R: Lecture des donnees a partir de l'instruction
# read.table.

TAB <- read.table("f:\\Programmes\\ExempleReliquat\\Reliquat.txt",
                  header=T)
```

Le tableau `TAB` comprend deux variables en colonne : la dose d'engrais appliquée (`DOSE`) et la mesure de reliquat récolte (`REL`).

Dans le cas des modèles linéaire et linéaire généralisé, le modèle était défini par une formule fournie comme argument de l'instruction utilisée pour estimer les paramètres (`lm` ou `glm`). Cette approche est également possible avec l'instruction `nls` mais il est souvent plus intéressant de définir les modèles non linéaires dans des fonctions spécifiques qui seront appelées ensuite par `nls`. Deux exemples de fonctions R définissant des modèles non linéaires sont présentés ci-dessous. La première fonction, `Mod.PL`, définit le modèle plateau-plus-linéaire et inclut quatre entrées : le vecteur des doses appliquées et les valeurs des trois paramètres (qui seront générées par `nls` au cours de ses itérations). Un vecteur Y est d'abord initialisé avec des valeurs manquantes (`NA`), puis ses éléments sont remplacés par les valeurs calculées en utilisant la fonction plateau-plus-linéaire. Une technique similaire est utilisée pour définir la fonction `Mod.PQ` qui décrit le modèle plateau-plus-quadratique.

```
#Code R

## Modele plateau-plus-lineaire
```

```
Mod.PL <- function(Dose, Rmin, A, Xmin) {

  Y <- rep(NA, length(Dose))
  Y[Dose<Xmin] <- Rmin
  Y[Dose>=Xmin] <- Rmin + A*(Dose[Dose>=Xmin] - Xmin)

  Y }

## Modele plateau-plus-quadratique

Mod.PQ <- function(Dose, Rmin, A, Xmin) {

  Y <- rep(NA, length(Dose))
  Y[Dose<Xmin] <- Rmin
  Y[Dose>=Xmin] <- Rmin + A*(Dose[Dose>=Xmin] - Xmin)^2

  Y }

#Code R: Estimation des parametres

Mod.pl.res <- nls(REL ~ Mod.PL(DOSE,Rmin, A, Xmin), data=TAB,
                start=list(Rmin=50,A=0.1, Xmin=200), trace=T)
print(summary(Mod.pl.res))

Mod.pq.res <- nls(REL ~ Mod.PQ(DOSE,Rmin, A, Xmin), data=TAB,
                start=list(Rmin=50, A=0.001, Xmin=200), trace=T)
print(summary(Mod.pq.res))
```

L'instruction `nls` fait appel aux modèles (par l'intermédiaire des fonctions `Mod.PL` et `Mod.PQ`), aux données (vecteurs `DOSE` et `REL` du tableau `TAB`) et à trois valeurs initiales des paramètres. Ces valeurs ont été choisies visuellement à partir des données présentées sur la figure 3.8. L'instruction `trace=T` est facultative. Elle permet de suivre les itérations de l'algorithme d'estimation. Ici, pour le modèle plateau-plus-linéaire, le nombre d'itérations est égal à trois. Après les valeurs initiales, les estimations des paramètres $\widehat{\boldsymbol{\alpha}} = \left(\widehat{\mathrm{R}}_{\min}, \widehat{A}, \widehat{\mathrm{X}}_{\min}\right)^T$ sont fixées successivement par `nls` à $(45.18, 0.22, 167.78)^T$, $(43.92, 0.21, 73.12)^T$ et finalement $(43.92, 0.21, 173.5)^T$. Les résultats de `nls` sont stockés dans des objets appelés `Mod.pl.res` et `Mod.pq.res`, puis imprimés à l'écran avec `summary`. Cette instruction permet d'obtenir les valeurs estimées des paramètres, leurs écarts types ainsi que les résultats d'une version approchée du test de Student (tableau 3.5).

Si les valeurs de $\widehat{\mathrm{R}}_{\min}$ (reliquat minimal) obtenues avec les deux modèles sont assez similaires, celles de $\widehat{\mathrm{X}}_{\min}$ (dose seuil au-delà de laquelle le reliquat augmente) sont très différentes. Avec le modèle plateau-plus-linéaire $\widehat{\mathrm{X}}_{\min} =$

Modèle	$\widehat{\mathrm{R}}_{\min}$	$\widehat{A}$	$\widehat{\mathrm{X}}_{\min}$	$\widehat{\sigma}^2$
Plateau-plus-linéaire	43.92 ***	0.21 **	173.50 ***	106.3
	(1.95)	(0.06)	(25)	
Plateau-plus-quadratique	40.33 ***	$3.92\ 10^{-4}$.	28.42	100.6
	(3.32)	$(2.19\ 10^{-4})$	(73.6)	

Tableau 3.5 – Valeurs estimées des paramètres des deux modèles de reliquat N récolte obtenues avec la méthode des moindres carrés ordinaires. Les valeurs entre parenthèses correspondent aux écarts types des estimateurs des paramètres. Les étoiles indiquent si la valeur estimée du paramètre est significativement différente de zéro avec une probabilité d'erreur de type I de 0.001 (***), 0.01 (**), 0.05 (*), 0.1 (.).

173.5 kg/ha alors qu'avec le modèle plateau-plus-quadratique $\widehat{\mathrm{X}}_{\min} = 28.42$ kg/ha. Par ailleurs, les paramètres du modèle plateau-plus-linéaire sont tous estimés avec une bonne précision ; les écarts types sont en effet faibles par rapport aux valeurs estimées et les tests sont très significatifs. Ce n'est pas le cas pour $\widehat{A}$ et $\widehat{\mathrm{X}}_{\min}$ avec le modèle plateau-plus-quadratique. Les valeurs de $\widehat{\sigma}^2$ obtenues avec les deux modèles sont similaires.

Méthode bayésienne Dans le cadre bayésien, $\boldsymbol{\alpha} = (\mathrm{R}_{\min}, A, \mathrm{X}_{\min})^T$ et σ^2 sont aléatoires. La version bayésienne des modèles est présentée graphiquement sur la figure 3.9.

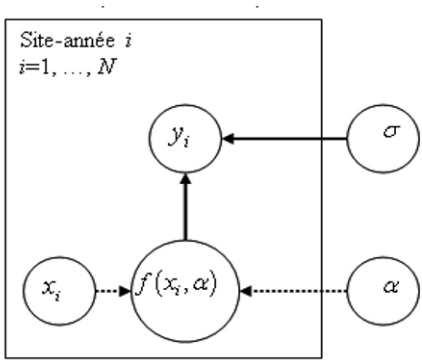

Figure 3.9 – Représentation graphique du modèle non linéaire de reliquat d'azote dans le sol. Le rectangle représente l'unité expérimentale (ici un site-année). Les cercles correspondent à des variables aléatoires. Les flèches continues indiquent une relation stochastique. Les flèches en pointillé correspondent à des relations déterministes.

L'estimation des paramètres est réalisée en trois étapes, comme dans le cas des modèles linéaire et linéaire généralisé :

– définition des distributions *a priori* des paramètres ;

– définition de la vraisemblance ;
– génération d'un échantillon de valeurs de paramètres à l'aide d'une méthode MCMC.

Les distributions *a priori* des paramètres R_{min}, A, X_{min} sont définies comme des lois gaussiennes indépendantes :

$$R_{min} \sim N\left(20, 10^3\right),\ A \sim N\left(0.5{,}10\right),\ X_{min} \sim N\left(140, 10^4\right)$$

pour le modèle plateau-plus-linéaire et

$$R_{min} \sim N\left(20, 10^3\right),\ A \sim N\left(0.005{,}10^{-4}\right),\ X_{min} \sim N\left(140, 10^4\right)$$

pour le modèle plateau-plus-quadratique. L'espérance de R_{min} correspond à la valeur fournie par le Comifer (1996). L'espérance de X_{min} correspond à la valeur médiane des doses d'engrais azoté appliquées sur les parcelles expérimentales (comprises entre 0 et 280 kg/ha). Finalement, l'espérance de A a été fixée à une valeur intermédiaire de 0.5 pour le modèle plateau-plus-linéaire et à une valeur 100 fois plus petite pour le plateau-plus-quadratique. La valeur de la variance de la loi gaussienne a été fixée à une valeur suffisamment grande pour qu'on puisse supposer que la distribution *a priori* n'apporte que peu d'information sur les valeurs des paramètres.

Cependant, des problèmes manifestes de convergence ont été détectés pour le modèle plateau-plus-quadratique : trois séries de valeurs de paramètres ont été générées par MCMC et ces trois séries se sont avérées être très différentes même en augmentant fortement le nombre d'itérations. Il a été nécessaire de réduire la variance de la distribution de A afin d'obtenir des résultats plus acceptables. Concernant la distribution *a priori* de la variance des erreurs, nous supposons, comme pour le modèle linéaire, que $1/\sigma^2 \sim \text{Gamma}\left(10^3, 10^3\right)$.

Les erreurs sont supposées ici indépendantes et distribuées selon une loi gaussienne de variance constante. La vraisemblance est donc définie par

$$V(\boldsymbol{\alpha}, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{52/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{52}\left(y_i - f(x_i, \boldsymbol{\alpha})\right)^2\right).$$

Le logiciel WinBUGS a été utilisé pour générer un échantillon de la loi *a posteriori* des paramètres des modèles à l'aide de l'échantillonneur de Gibbs. Comme dans les exemples précédents, la méthode d'estimation a été appliquée en utilisant trois fichiers WinBUGS, un fichier qui contient les données, un fichier qui inclut les valeurs initiales des paramètres, et un fichier décrivant le modèle. Ces fichiers sont présentés ci-dessous pour le modèle plateau-plus-linéaire.

```
# Code BUGS : Donnees (extrait)

list(
  y=c( 53.2 ,...),
```

```
  d=c( 0 , 120...),
  N=52)

# Code BUGS : Valeurs initiales

list(alpha=c(20,0.5,140), prec=1)
```

Le fichier des données et celui incluant les valeurs initiales des paramètres ont la même structure que ceux utilisés dans les exemples précédents. Trois fichiers incluant des valeurs initiales légèrement différentes ont été définis afin de tester la stabilité des résultats, mais un seul est présenté ici. Il y a trois paramètres dans la fonction plateau-plus-linéaire et une variance des erreurs ; il faut donc initialiser quatre paramètres.

```
# Code BUGS: modele plateau-plus-lineaire

model

# y= Reliquat N recolte
# d = dose d'engrais
# alpha = parametres du modele
# prec = 1/variance

{
for (i in 1:N) {
  y[i]  ~   dnorm(mu[i], prec)
  mu[i] <- alpha[1] + step(d[i]-alpha[3]) * alpha[2] * (d[i]-alpha[3])
  }

## Prior ##
  alpha[1] ~ dnorm(20,1.0E-3)
  alpha[2] ~ dnorm(0.5,1.0E-1)
  alpha[3] ~ dnorm(140,1.0E-4)
  prec ~ dgamma(0.001, 0.001)

## Variance ##

  var <- 1/prec
}
```

Le fichier décrivant le modèle plateau-plus-linéaire présente d'abord la distribution gaussienne utilisée pour calculer la vraisemblance puis les distributions *a priori* des paramètres. La syntaxe de la fonction plateau-plus-linéaire est basée sur l'utilisation de l'instruction `step` qui prend la valeur 1 si l'expression entre parenthèses est supérieure à zéro et qui prend la valeur zéro sinon.

L'estimation a été réalisée avec un fichier script. Trois séries de valeurs de paramètres ont été générées afin de tester la convergence de l'algorithme avec la méthode de Brooks et Gelman (1998) (instruction `gr`). Chaque série est basée sur un fichier de valeurs initiales différent. Une période de chauffe de 30 000 itérations est spécifiée, puis 150 000 valeurs de paramètres sont générées par l'échantillonneur de Gibbs. L'instruction `thin-samples` permet de ne garder qu'une fraction des valeurs générées par l'algorithme, ici une sur 30. Cette procédure est utile lorsque des auto-corrélations sont détectées entre les valeurs générées par l'algorithme. Les tests essai-erreur préliminaires ont révélé l'existence de telles corrélations. Le fait de ne garder qu'une valeur de paramètre sur 30 permet de réduire les auto-corrélations à un niveau quasi nul. Les valeurs générées ont été utilisées pour déterminer la distribution *a posteriori* des paramètres (tableau 3.6).

```
# Code BUGS: fichier Script

display(log)
set.seed(1)

check(f:/Projets/ExempleReliquat/ModelExempleRel.odc)
data(f:/Projets/ExempleReliquat/DataExempleRel.odc)

compile(3)

inits(1, f:/Projets/ExempleReliquat/IniExempleRel.odc)
inits(2, f:/Projets/ExempleReliquat/IniExempleRelb.odc)
inits(3, f:/Projets/ExempleReliquat/IniExempleRelc.odc)

set(alpha)
set(var)

beg(30000)
thin.samples(30)
update(30000)
dic.set()
update(150000)

history(alpha)
gr(alpha)
autoC(alpha)
coda(alpha)

stats(*)
dic.stats()
```

Les résultats sont assez similaires à ceux obtenus par les moindres carrés.

Modèle	R_{min}	A	X_{min}	σ^2
Plateau-plus-linéaire	41.15	0.16	118.5	107.3
	(3.54)	(0.07)	(53.7)	(22.6)
Plateau-plus-quadratique	42.38	$9.58\ 10^{-4}$	89.11	106.2
	(2.85)	$(1.22\ 10^{-3})$	(60.36)	(22.65)

Tableau 3.6 – Moyenne et écart type (entre parenthèses) des distributions *a posteriori* des paramètres des deux modèles non linéaires pour le reliquat N. Calculs réalisés avec 150 000 itérations de l'échantillonneur de Gibbs après une période de chauffe de 30 000 itérations et un filtrage (une valeur sur 30 conservée).

Les moyennes *a posteriori* de R_{min} obtenues avec les deux modèles sont très proches. Par contre, les moyennes *a posteriori* de A et de X_{min} obtenues avec le modèle plateau-plus-linéaire sont très différentes de celles obtenues avec le modèle plateau-plus-quadratique. Par ailleurs, pour le modèle plateau-plus-quadratique, les écarts types *a posteriori* de A et de X_{min} sont très élevés et révèlent des niveaux d'incertitude associés à ces paramètres importants. Ces deux paramètres sont estimés de façon plus précise avec le modèle plateau-plus-linéaire.

Étape *iv.* Évaluation

Comme les modèles linéaire et linéaire généralisé, les modèles non linéaires peuvent être évalués et comparés en calculant les valeurs des critères AIC et DIC. Les AIC des modèles plateau-plus-linéaire et plateau-plus-quadratique sont égaux respectivement à 395.08 et 392.26 après estimation par les moindres carrés ordinaires. Ils sont donc proches. Ce résultat est logique car les deux modèles ont des variances résiduelles semblables et incluent le même nombre de paramètres.

Les valeurs de DIC sont un peu plus contrastées : 392.2 pour le modèle plateau-plus-linéaire et 378.8 pour le modèle plateau-plus-quadratique. Cependant, ces valeurs de DIC sont difficiles à comparer car les deux modèles n'utilisent pas la même distribution *a priori* pour le paramètre A.

Les valeurs des paramètres sont estimées de façon plus précise avec le modèle plateau-plus-linéaire qu'avec le modèle plateau-plus-quadratique. Nous avons en effet noté lors de l'étape d'estimation que les valeurs des paramètres A et X_{min} du modèle plateau-plus-quadratique étaient mal estimées aussi bien avec la méthode fréquentiste qu'avec la méthode bayésienne. Le choix du modèle plateau-plus-linéaire semble donc plus judicieux.

La figure 3.8 présente les résidus obtenus avec les deux modèles pour quatre des huit sites-années, après estimation par les moindres carrés. Les résidus appartenant à un même site-année sont reliés entre eux. Cette présentation des

résidus montre que les résidus d'un même site-année tendent à être tous du même signe, soit négatif soit positif. Ceci indique que les deux modèles ont tendance à sous-estimer systématiquement ou, inversement, à sur-estimer systématiquement les observations d'un même site-année. Il est donc peu réaliste de supposer les erreurs toutes indépendantes. Un constat identique peut être fait avec l'estimation bayésienne.

La fonction de vraisemblance utilisée dans cette étude était basée sur une hypothèse d'indépendance des erreurs. Son utilisation dans un contexte où les erreurs ne sont manifestement pas toutes indépendantes peut conduire à des estimations imprécises et à une vision trop optimiste de la qualité de cette estimation (*e.g.*, Davidian et Giltinan, 1995). Un autre type de modélisation est proposé dans la section suivante pour traiter ce type de problème : les modèles hiérarchiques.

3.4 Modèle hiérarchique

3.4.1 Définition et intérêt

Les modèles hiérarchiques sont utilisés lorsque la variabilité totale des observations peut être décomposée en plusieurs niveaux, par exemple en une variabilité intra-individu et une variabilité inter-individus. Pour les domaines d'application qui nous intéressent, un individu correspondra souvent à un site-année et, dans ce cas, un modèle hiérarchique permettra de décrire à la fois la variabilité des observations au sein d'un site-année donné et la variabilité entre sites-années.

Cette approche est particulièrement utile pour l'analyse des observations dites « répétées » ou « longitudinales », c'est-à-dire des observations qui résultent de mesures réalisées à plusieurs reprises sur une série d'individus. En agronomie, ce cas de figure se rencontre lorsque des mesures (par exemple de biomasse) sont réalisées à plusieurs dates différentes sur les mêmes sites-années ou lorsque plusieurs mesures sont réalisées sur plusieurs traitements expérimentaux (par exemple plusieurs doses d'engrais) appliqués sur plusieurs sous-parcelles toutes localisées sur un même site-année.

Un modèle hiérarchique peut être linéaire, linéaire généralisé ou non linéaire (Davidian et Giltinan, 1995 ; Gilks *et al.*, 1996). Dans tous les cas, il permet d'analyser et de comparer l'importance respective de plusieurs niveaux de variabilité. Il prend également en compte les corrélations intra-individu des erreurs, comme celles mises en évidence dans la section 3.3. Les estimations des paramètres obtenues avec un modèle hiérarchique peuvent ainsi être plus précises et les intervalles de confiance plus réalistes que ceux obtenus avec un modèle standard.

Modèle hiérarchique pour décrire la variabilité inter sites-années On considère l'analyse d'un jeu de données issues d'une expérimentation conduite

dans un réseau de n sites-années. Soit Y_{ij} la j-ième mesure obtenue sur le i-ième site-année, pour $i = 1, \ldots, n$ et $j = 1, \ldots, m_i$, où m_i est le nombre total de mesures dans le site-année i. Le vecteur X_{ij} inclut des variables explicatives caractérisant cette mesure, par exemple une date d'observation ou une dose d'engrais. Ce type de données peut être modélisé avec un modèle hiérarchique, dont la formulation générale se décompose en deux niveaux :

$$\textit{Niveau 1, intra site-année :} \quad Y_{ij} = f(X_{ij}, \boldsymbol{\alpha}_i) + \varepsilon_{ij},\ j = 1, \ldots, m_i, \quad (3.8)$$

$$\textit{Niveau 2, inter sites-années :} \quad \boldsymbol{\alpha}_i = D_i \boldsymbol{\mu} + \boldsymbol{\eta}_i,\ i = 1, \ldots, n. \quad (3.9)$$

Dans l'équation (3.8), la fonction f décrit la réponse intra site-année et inclut un vecteur de paramètres $\boldsymbol{\alpha}_i$ spécifique du site-année i. Le terme $\varepsilon_{ij} = Y_{ij} - f(X_{ij}, \boldsymbol{\alpha}_i)$ correspond à l'erreur intra site-année. Il est aléatoire, par exemple distribué selon une loi gaussienne $\varepsilon_{ij} \sim N(0, \sigma^2)$. L'équation (3.9) définit la façon dont les paramètres varient entre sites-années. Dans cette équation, D_i est une matrice spécifique du i-ième site-année, supposée connue, $\boldsymbol{\mu}$ est un vecteur de paramètres fixes, $\boldsymbol{\eta}_i$ est un vecteur d'effets aléatoires. La matrice D_i permet de prendre en compte des variables caractéristiques des sites-années, qui expliquent au moins partiellement la variabilité inter sites-années de $\boldsymbol{\alpha}_i$. Le vecteur $\boldsymbol{\eta}_i$ permet de modéliser la variabilité inter sites-années des paramètres non expliquée par $D_i\boldsymbol{\mu}$. Ce vecteur est supposé suivre une loi de probabilité, par exemple une loi gaussienne $\boldsymbol{\eta}_i \sim N(0, \Gamma)$, où Γ est une matrice de variance-covariance entre les paramètres.

Dans le modèle ci-dessus, les paramètres décrivant les caractéristiques de la population sont $(\boldsymbol{\mu}, \Gamma, \sigma^2)$. Plusieurs méthodes ont été proposées dans le cadre de la statistique classique pour estimer ces paramètres (Davidian et Giltinan, 1995), notamment la méthode du maximum de vraisemblance et la méthode du maximum de vraisemblance restreint. L'écriture de la vraisemblance est plus complexe pour les modèles hiérarchiques que pour les modèles rencontrés dans les sections précédentes du fait de la présence de corrélations entre observations. La maximisation de la vraisemblance est également plus délicate à mettre en œuvre, notamment lorsque la fonction f est non linéaire. Des algorithmes sont disponibles dans le logiciel R pour les modèles linéaires et non linéaires (Pinheiro et Bates, 2000). Un autre algorithme, plus récent, a été proposé par Kuhn et Lavielle (2005) et peut être mis en œuvre avec le logiciel Monolix (`http://software.monolix.org/index.php`). En plus de l'estimation des paramètres de la population, ces algorithmes permettent également de prédire les valeurs des effets sites-années $\boldsymbol{\alpha}_i$, $i = 1, \ldots, n$.

Les deux niveaux du modèle défini par (3.8) et (3.9) peuvent être complétés par un troisième niveau lorsqu'on souhaite se placer dans un cadre bayésien. Le troisième niveau permet, dans ce cas, de définir des distributions *a priori* pour les paramètres $\boldsymbol{\mu}$, Γ, σ (Gilks *et al.*, 1996 ; Carlin et Louis, 2008). L'estimation des paramètres peut alors être réalisée avec des méthodes MCMC et le logiciel WinBUGS.

Prenons un exemple simple pour illustrer la notion de modèle hiérarchique. Supposons qu'une observation Y (*e.g.*, biomasse d'une culture) soit reliée au temps t (*e.g.*, somme de température) à l'aide du modèle hiérarchique suivant :

$$\begin{aligned} &\textit{Niveau 1, intra site-année :} && Y_{ij} = \alpha_{1i} + \alpha_{2i}\, t_{ij} + \varepsilon_{ij}, j = 1, \ldots, m_i, \\ &\textit{Niveau 2, inter sites-années :} && \alpha_{1i} \sim N\left(\mu_{\alpha 1}, \sigma^2_{\alpha 1}\right) \text{ et } \alpha_{2i} \sim N\left(\mu_{\alpha 2}, \sigma^2_{\alpha 2}\right), \end{aligned}$$

où Y_{ij} est la j-ième mesure obtenue sur le i-ième site-année à la date t_{ij}, où α_{1i} et α_{2i} sont les deux paramètres de la régression pour le site-année i et l'erreur intra site-année ε_{ij} est supposée distribuée selon une loi gaussienne $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$. On suppose que α_{1i} et α_{2i} sont indépendants entre eux et indépendants de ε_{ij}. Dans l'équation de niveau 2, $\mu_{\alpha 1}$ et $\mu_{\alpha 2}$ représentent les espérances de α_{1i} et α_{2i} dans la population de sites-années considérée. Les variances $\sigma^2_{\alpha 1}$ et $\sigma^2_{\alpha 2}$ décrivent la variabilité des valeurs de α_{1i} et α_{2i} entre sites-années.

Selon ce modèle, la relation linéaire est valable pour tous les sites-années, mais les valeurs des deux paramètres α_{1i} et α_{2i} de cette relation varient entre sites-années. Les paramètres de la population de sites-années sont

$$\boldsymbol{\mu} = \left(\mu_{\alpha 1}, \mu_{\alpha 2}\right)^T, \quad \Gamma = \begin{pmatrix} \sigma^2_{\alpha 1} & 0 \\ 0 & \sigma^2_{\alpha 2} \end{pmatrix} \text{ et } \sigma^2.$$

L'expression $\mu_{\alpha 1} + \mu_{\alpha 2}\, t$ représente l'espérance de la réponse Y en fonction du temps t, en moyenne sur la population de sites-années.

Ce modèle permet de décrire des corrélations et des différences de variabilité entre observations d'un même site-année, en utilisant seulement trois paramètres. En effet, la variance de Y_{ij} est égale à $\mathrm{Var}\left(Y_{ij}\right) = \sigma^2_{\alpha 1} + \sigma^2_{\alpha 2}\, t^2_{ij} + \sigma^2$ et augmente donc avec le temps t_{ij}. Considérons une autre observation, $Y_{ij'}$, obtenue sur le même site-année i, mais à une autre date $t_{ij'}$. La covariance de Y_{ij} et $Y_{ij'}$ est égale à

$$\mathrm{Cov}\left(Y_{ij}, Y_{ij'}\right) = \sigma^2_{\alpha 1} + t_{ij} t_{ij'} \sigma^2_{\alpha 2}$$

La covariance de deux observations d'un même site-année Y_{ij} et $Y_{ij'}$ ne dépend donc que des deux paramètres $\sigma^2_{\alpha 1}$ et $\sigma^2_{\alpha 2}$. Les variances et covariances de toutes les observations peuvent être calculées de cette manière en fonction de $\sigma^2_{\alpha 1}$, $\sigma^2_{\alpha 2}$ et σ^2. Le modèle hiérarchique permet donc de modéliser de l'hétéroscédasticité (hétérogénéité des variances) et des corrélations entre observations, de façon parcimonieuse.

Modèles hiérarchiques pour analyser les données spatialisées Les modèles hiérarchiques sont également utiles pour analyser les données spatialisées. Dans certains cas, les mesures sont réalisées dans différentes localisations spatiales plus ou moins éloignées. Du fait de l'effet de facteurs milieu, des mesures

réalisées dans des localisations spatiales proches sont susceptibles de se ressembler et, donc, d'être fortement corrélées. Il est alors utile de modéliser ces corrélations spatiales. C'est notamment le cas lorsque l'objectif de l'étude est de prédire une variable d'intérêt dans des localisations où aucune mesure n'a été réalisée en tenant compte des mesures réalisées autour de ces localisations. Ce type de prédiction par interpolation est appelé *krigeage*.

Les modèles hiérarchiques sont intéressants dans ce contexte car ils permettent de modéliser à la fois :

- les corrélations entre les valeurs prises par une variable dans différentes localisations spatiales, en tenant compte de leurs distances ;
- la variabilité des mesures au sein de chacune de ces localisations spatiales.

Cette utilisation particulière des modèles hiérarchiques fait appel à des techniques spécifiques présentées en détail dans Banerjee *et al.* (2004). Un exemple est décrit dans la section 3.4.3.

3.4.2 Exemple : reliquat d'azote dans le sol à la récolte

Nous reprenons l'exemple traité dans la section 3.3.2 dont l'objectif était de modéliser l'effet de la dose totale d'engrais minéral sur la quantité d'azote minéral résiduel dans le sol. La modélisation est réalisée ici avec un modèle non linéaire hiérarchique. La fonction f du modèle correspond à la fonction plateau-plus-linéaire présentée dans la section 3.3.2.

L'utilisation de ce type de modèle est assez naturelle compte tenu des données disponibles. La figure 3.10 présente les observations de quatre des huit sites-années de la base de données. Cette figure montre que les réponses des différents sites-années ont toutes la même allure mais que les paramètres des courbes de réponse semblent varier d'un site-année à l'autre.

L'exemple est repris à l'étape *ii.*

Étape *ii.* Définition des équations

Le modèle plateau-plus-linéaire peut être défini comme un modèle hiérarchique à deux niveaux. L'idée sous-jacente est que la réponse « reliquat récolte » à la dose suit partout une fonction plateau-plus-linéaire mais que les valeurs des paramètres de cette fonction varient d'un site-année à l'autre. Le premier niveau décrit la fonction plateau-plus-linéaire de chaque site-année, ainsi que la variabilité des mesures autour de cette fonction. Le deuxième niveau du modèle décrit la variabilité inter sites-années de la fonction de réponse.

Soit Y_{ij} la j-ième mesure de reliquat récolte et X_{ij} la j-ième dose d'engrais sur le i-ième site-année, pour $i = 1, \ldots, 8$ et $j = 1, \ldots, m_i$, où m_i est le nombre total de mesures du site-année i. Le modèle s'écrit :
Niveau 1, intra site-année :

$$Y_{ij} = \max\left\{\mathrm{R_{min,i}}\,;\, \mathrm{R_{min,i}} + A_i(X_{ij} - \mathrm{X_{min,i}})\right\} + \varepsilon_{ij},\ j = 1, \ldots, m_i, \quad (3.10)$$

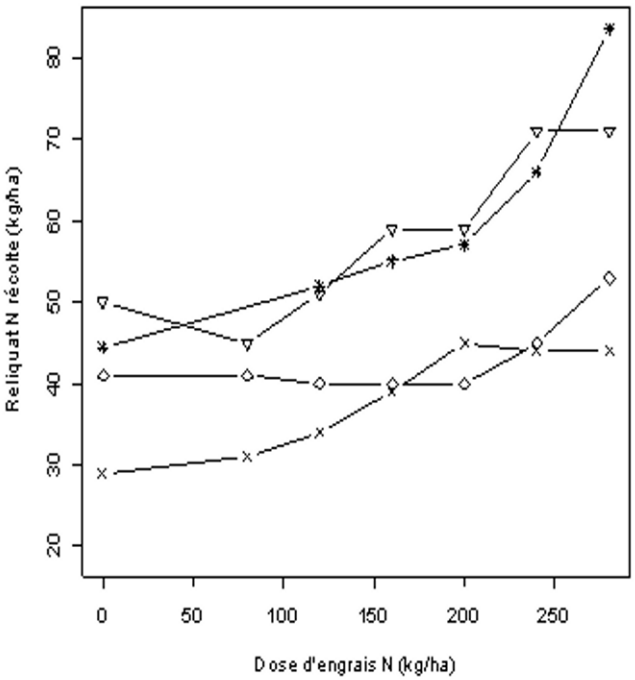

Figure 3.10 – Réponses observées du reliquat N récolte à la dose d'engrais appliquée pour quatre sites-années.

Niveau 2, inter sites-années :

$$\boldsymbol{\alpha}_i = \boldsymbol{\mu} + \boldsymbol{\eta}_i,\ i = 1, \ldots, 8. \tag{3.11}$$

Dans l'équation (3.10), $\boldsymbol{\alpha}_i = (\mathrm{R}_{\mathrm{min,i}}, A_i, \mathrm{X}_{\mathrm{min,i}})^T$ est le vecteur des paramètres du i-ième site-année. L'erreur ε_{ij} est supposée distribuée selon une loi gaussienne $\varepsilon_{ij} \sim N(0, \sigma^2)$. Dans l'équation (3.11), $\boldsymbol{\mu} = (\mu_{\mathrm{Rmin}}, \mu_A, \mu_{\mathrm{Xmin}})^T$ est le vecteur correspondant à la partie fixe des paramètres, c'est-à-dire le vecteur incluant les valeurs moyennes de $\boldsymbol{\alpha}_i$ et $\boldsymbol{\eta}_i$ est un vecteur d'effets aléatoires. La matrice D_i n'apparaît pas dans (3.11) car cette matrice est égale ici à la matrice identité. Le vecteur $\boldsymbol{\eta}_i$ permet de modéliser la variabilité inter sites-années des paramètres. Ce vecteur est supposé suivre une loi gaussienne $\boldsymbol{\eta}_i \sim N(\mathbf{0}, \Gamma)$, où Γ est une matrice de variance-covariance. Nous définissons ici Γ comme une matrice diagonale dont les éléments diagonaux correspondent aux variances des éléments de $\boldsymbol{\alpha}_i$,

$$\Gamma = \begin{pmatrix} \sigma^2_{\mathrm{Rmin}} & 0 & 0 \\ 0 & \sigma^2_A & 0 \\ 0 & 0 & \sigma^2_{\mathrm{Xmin}} \end{pmatrix}.$$

Notez qu'il est possible de considérer une matrice Γ incluant des covariances non nulles, mais une telle matrice inclut un nombre plus élevé de paramètres à estimer ce qui peut poser des problèmes lors de la phase d'estimation (Makowski et Lavielle, 2006).

Selon ce modèle, les paramètres de la population sont

$$\boldsymbol{\mu} = (\mu_{\mathrm{Rmin}}, \mu_A, \mu_{\mathrm{Xmin}})^T, \quad \Gamma = \begin{pmatrix} \sigma^2_{\mathrm{Rmin}} & 0 & 0 \\ 0 & \sigma^2_A & 0 \\ 0 & 0 & \sigma^2_{\mathrm{Xmin}} \end{pmatrix} \text{ et } \sigma^2.$$

Le modèle inclut également des paramètres aléatoires pour chaque site-année, les vecteurs $\boldsymbol{\alpha}_i = (\mathrm{R}_{\mathrm{min,i}}, A_i, \mathrm{X}_{\mathrm{min,i}})^T$, $i = 1, \ldots, 8$, définis à partir de (3.11). Par exemple, $\mathrm{R}_{\mathrm{min,i}} = \mu_{\mathrm{Rmin}} + \eta_{i,1}$, avec $\eta_{i,1}$ le premier élément du vecteur $\boldsymbol{\eta}_i$.

Dans le modèle défini par (3.10) et (3.11), les trois paramètres de la fonction plateau-plus-linéaire sont aléatoires. Il est possible de supposer que seulement certains des trois paramètres de la fonction sont aléatoires. Par exemple, une autre version du modèle consiste à définir $\mathrm{R}_{\mathrm{min}}$ comme un paramètre aléatoire et A et $\mathrm{X}_{\mathrm{min}}$ comme fixes. Un tel modèle inclut moins de paramètres à estimer que la version à trois paramètres aléatoires.

Étape *iii.* Estimation des paramètres

Estimation par maximisation de la vraisemblance Les deux programmes ci-dessous montrent comment les paramètres peuvent être estimés avec la fonction `nlme` de R. Le premier programme décrit la fonction plateau-plus-linéaire. Il est très proche du programme utilisé avec la fonction `nls`. La seule différence est que les arguments associés aux paramètres sont maintenant des

vecteurs et non des scalaires. Les tests booléens `Dose < Xmin` et `Dose >= Xmin` doivent donc être appliqués à ces vecteurs de paramètres.

Le deuxième programme estime les paramètres avec la fonction `nlme` à partir des données disponibles. Les paramètres de deux modèles sont estimés : un modèle incluant des effets aléatoires indépendants sur trois paramètres et un modèle incluant un effet aléatoire sur un seul paramètre R_{min}. Le tableau de données `TAB` est d'abord transformé en un nouveau tableau par l'instruction `groupedData`. L'objectif de cette opération est de définir quelle est la variable de réponse (`REL`), quelle est la variable explicative (`DOSE`) et quelle est la variable définissant les individus (`NOM`). La variable `NOM` correspond à une colonne de `TAB` indiquant les mesures appartenant à un même site-année. L'instruction `nlme` est ensuite appelée pour estimer les paramètres. Les paramètres à effets fixes et à effets aléatoires doivent être identifiés. L'instruction `pdDiag` est utilisée pour définir une matrice Γ diagonale. Des valeurs initiales doivent être fournies pour les effets fixes à l'aide de l'argument `start`.

```
# Code R

Mod.PL <- function(Dose, Rmin, A, Xmin) {

  Y <- Dose
  Y[Dose<Xmin]  <- Rmin[Dose<Xmin]
  Y[Dose>=Xmin] <- Rmin[Dose>=Xmin] + A[Dose>=Xmin] *
                     (Dose[Dose>=Xmin] - Xmin[Dose>=Xmin])
  Y}

TAB.est <- groupedData(REL ~ DOSE | NOM, data=TAB)
Fit1 <- nlme(REL ~  Mod.PL(DOSE, Rmin, A, Xmin), data=TAB.est,
                     fixed = Rmin+A+Xmin ~ 1,
                     random = pdDiag(Rmin+A+Xmin ~ 1),
                     start = c(Rmin=50, A=0.1, Xmin=200))
print(summary(Fit1))

Fit2 <- nlme(REL ~  Mod.PL(DOSE, Rmin, A, Xmin), data=TAB.est,
                     fixed = Rmin+A+Xmin ~ 1,
                     random = pdDiag(Rmin ~ 1),
                     start = c(Rmin=50, A=0.1, Xmin=200))
print(summary(Fit2))
```

Le tableau 3.7 présente les valeurs estimées des paramètres du modèle avec R_{min} aléatoire. Les résultats correspondent à l'objet `Fit2` du programme R ci-dessus. Le tableau 3.7 présente les valeurs estimées de l'espérance de R_{min} (μ_{Rmin}), de sa variance (σ^2_{Rmin}), des valeurs de R_{min} pour les huit sites-années ($R_{min,i}$, pour $i = 1, \ldots, 8$) des paramètres fixes A et X_{min} et finalement de la variance des erreurs σ^2. La fonction `nlme` fournit également les écarts types des

estimateurs des effets fixes, ici les écarts types des estimateurs de μ, $\mathrm{R_{min}}$, A et $\mathrm{X_{min}}$.

Le tableau 3.7 montre que les valeurs estimées de μ_{Rmin}, de A et de $\mathrm{X_{min}}$ sont proches des valeurs des paramètres correspondants reportées dans le tableau 3.5 pour un modèle non hiérarchique. Le tableau 3.7 montre également que la valeur estimée de σ^2_{Rmin} est relativement élevée, ce qui indique que le niveau du plateau varie fortement entre sites-années. La valeur estimée de la variance des erreurs σ^2 est bien plus faible (49.28) avec le modèle hiérarchique qu'avec le modèle sans effet aléatoire (106.3 d'après le tableau 3.7). Ce résultat montre qu'une part importante de la variabilité des observations est expliquée par la variabilité de $\mathrm{R_{min}}$.

Paramètre	Estimation	
μ_{Rmin}	44.10	(2.92)
σ^2_{Rmin}	49.99	
$\mathrm{R_{min,1}}$	33.68	
$\mathrm{R_{min,2}}$	37.93	
$\mathrm{R_{min,3}}$	36.63	
$\mathrm{R_{min,4}}$	47.81	
$\mathrm{R_{min,5}}$	45.11	
$\mathrm{R_{min,6}}$	51.20	
$\mathrm{R_{min,7}}$	48.70	
$\mathrm{R_{min,8}}$	51.70	
A	0.21	(0.05)
$\mathrm{X_{min}}$	174.34	(17.43)
σ^2	49.28	

Tableau 3.7 – Valeurs estimées des paramètres du modèle de reliquat N dans le sol avec $\mathrm{R_{min}}$ aléatoire. Résultats obtenus avec `nlme`. Les nombres entre parenthèses désignent des écarts types.

Méthode bayésienne Les paramètres du modèle (3.10)-(3.11) peuvent également être estimés avec une approche bayésienne. Le principe est de définir des distributions *a priori* pour les paramètres de la population. Nous utilisons ici les mêmes distributions que celles utilisées dans la section 3.3.2. Dans le cas d'un modèle où $\mathrm{R_{min}}$ et $\mathrm{X_{min}}$ sont aléatoires, les distributions *a priori* suivantes sont choisies :

$$\mu_{\mathrm{Rmin}} \sim N(20, 10^3), \qquad A \sim N(0.5, 10),$$
$$\mu_{\mathrm{Xmin}} \sim N(140, 10^4), \qquad 1/\sigma^2 \sim \mathrm{Gamma}(10^3, 10^3),$$
$$1/\sigma^2_{\mathrm{Rmin}} \sim \mathrm{Gamma}(10^3, 10^3), \qquad 1/\sigma^2_{\mathrm{Xmin}} \sim \mathrm{Gamma}(10^3, 10^3).$$

Le programme WinBUGS utilisé pour estimer les distributions *a posteriori* comporte, comme dans les cas précédents, trois fichiers. Le premier inclut les données. Ici, la structure du fichier de données est plus complexe car il est nécessaire de définir la structure hiérarchique des observations et donc de distinguer le niveau intra site-année du niveau inter sites-années. Il existe par ailleurs des données manquantes car certaines doses ne sont pas représentées sur tous les sites-années. Ces données manquantes sont identifiées par `NA`.

```
# Code BUGS

list(
 d=c(0 , 80 , 120 , 160 , 200 , 240 , 280 ),
 N=8, T=7,
 y=structure(
 .Data = c(53.2 , NA , 39.4 , 36.1 , 48 , 69.3 , 66.2 , 19 , NA ,
  41.3 , 50.7 , 30.2 , 44.5 , 67.3 , 43.3 , NA , 53 , 50.2 , 62.3 ,
  55 , 67.3 , 29 , 31 , 34 , 39 , 45 , 44 , 44 , 41 , 41 , 40 , 40 ,
  40 , 45 , 53 , 50 , 45 , 51 , 59 , 59 , 71 , 71 , 40 , 47 , 44 ,
  61 , 55 , 63 , 76 , 44.5, NA , 52 , 55.1 , 57.1 , 66 , 83.6 ),
  .Dim=c(8, 7))
)
```

Le deuxième fichier inclut les valeurs initiales utilisées par l'algorithme MCMC. Ce fichier fournit des valeurs initiales à la fois pour les paramètres de la population et pour ceux des sites-années individuels.

```
# Code BUGS

list(Rmin=c(20,20,20,20,20,20,20,20),
 A=0.5,
 Xmin=c(140,140,140,140,140,140,140,140),
 Rmin.m=20,
 A=0.5,
 Xmin.m=140,
 Rmin.p=1,
 Xmin.p=1,
 prec=1)
```

Le modèle est décrit à l'aide de deux boucles `for`. La première boucle correspond au niveau intra site-année et la deuxième au niveau inter sites-années. Les distributions *a priori* sont ensuite définies. L'indice j correspond à l'indice des doses et l'indice i à celui des sites-années.

```
# Code BUGS

model
```

```
# y= Reliquat N recolte
# d = dose d'engrais

{

for (i in 1:N) {
  for (j in 1:T) {
    y[i,j]  ~   dnorm(mu[i,j],prec)
    mu[i,j] <- Rmin[i] + step(d[j] - Xmin[i]) * A * (d[j]-Xmin[i])
    }

  Rmin[i] ~  dnorm(Rmin.m, Rmin.p)
  Xmin[i] ~  dnorm(Xmin.m, Xmin.p)
  }

## Prior ##

Rmin.m ~ dnorm(20, 1.0E-3)
A ~ dnorm(0.5, 1.0E-1)
Xmin.m ~ dnorm(140, 1.0E-4)
Rmin.p ~ dgamma(0.001, 0.001)
Xmin.p ~ dgamma(0.001, 0.001)
prec ~ dgamma(0.001, 0.001)

## Variance ##

var <- 1/prec
varRmin <- 1/Rmin.p
varXmin <- 1/Xmin.p

}
```

Les trois fichiers ci-dessus sont appelés dans le fichier script ci-dessous. La période de chauffe de l'algorithme MCMC est ici de 30 000 itérations, puis 100 000 itérations sont lancées. Une valeur sur 30 est conservée afin de limiter les problèmes d'auto-corrélation. Les estimations des espérances et variances *a posteriori* sont présentées dans le tableau 3.8.

```
# Code BUGS : fichier Script

display(log)
set.seed(1)

check(f:/Projets/ExempleReliquatMix/ModelExempleRelMixRminXmin.odc)
data(f:/Projets/ExempleReliquatMix/DataExempleRelMix.odc)
```

```
compile(1)

inits(1,f:/Projets/ExempleReliquatMix/IniExempleRelMixRminXmin.odc)
gen.inits()

set(Rmin.m)
set(A)
set(Xmin.m)
set(varRmin)
set(varXmin)
set(var)
set(Rmin)
set(Xmin)

beg(30000)
thin.samples(30)
update(30000)
dic.set()
update(100000)

stats(*)
dic.stats()
```

Étape *iv.* Évaluation des modèles

Les critères AIC et DIC ont été calculés pour déterminer le nombre de paramètres qu'il est nécessaire de considérer comme aléatoires. Les techniques présentées dans la section 3.1.5 ont été utilisées pour ajuster aux données une série de modèles incluant un nombre plus ou moins grand de paramètres aléatoires. Le modèle le plus simple n'inclut aucun paramètre aléatoire. Il correspond au modèle de la section 3.3.2. Le modèle le plus complexe inclut trois paramètres aléatoires.

Les résultats sont présentés dans le tableau 3.9. La fonction `nlme` n'a pas convergé pour deux des modèles testés. Pour ces deux modèles, il n'a donc pas été possible de calculer l'AIC. Dans le cadre fréquentiste, les valeurs d'AIC montrent que le modèle optimal est celui où seulement R_{min} est aléatoire. Dans le cadre bayésien, les valeurs de DIC montrent que le modèle optimal est celui où R_{min} et X_{min} sont aléatoires, mais le DIC du modèle où seulement R_{min} est aléatoire est à peine supérieur.

Utilisation du modèle pour évaluer les risques de pollution

L'utilisation du modèle bayésien avec R_{min} et X_{min} aléatoires est illustrée sur la figure 3.11. La figure 3.11a présente l'ajustement du modèle pour deux des huit sites-années obtenu avec les paramètres spécifiques des sites-années.

Paramètre	Espérance	Variance
μ_{Rmin}	39.39	3.77
σ^2_{Rmin}	76.61	76.21
$\text{R}_{\text{min},1}$	41.29	3.72
$\text{R}_{\text{min},2}$	31.44	5.19
$\text{R}_{\text{min},3}$	43.00	3.85
$\text{R}_{\text{min},4}$	30.61	4.08
$\text{R}_{\text{min},5}$	35.52	3.77
$\text{R}_{\text{min},6}$	46.21	4.55
$\text{R}_{\text{min},7}$	43.22	4.67
$\text{R}_{\text{min},8}$	46.07	4.23
A	0.13	0.02
μ_{Xmin}	90.78	31.86
σ^2_{Xmin}	1524	3386
$\text{X}_{\text{min},1}$	99.00	36.00
$\text{X}_{\text{min},2}$	81.14	47.02
$\text{X}_{\text{min},3}$	87.13	33.76
$\text{X}_{\text{min},4}$	101.60	41.62
$\text{X}_{\text{min},5}$	114.00	46.91
$\text{X}_{\text{min},6}$	81.19	42.18
$\text{X}_{\text{min},7}$	75.28	44.18
$\text{X}_{\text{min},8}$	77.97	36.51
σ^2	49.79	12.04

Tableau 3.8 – Espérance et variance *a posteriori* des paramètres de la population et des paramètres individuels du modèle de reliquat N hiérarchique. Modèle avec R_{min} et X_{min} aléatoires. Calculs réalisés avec WinBUGS.

Effets aléatoires	AIC	DIC
R_{min}, A, X_{min}	380.3	361.7
R_{min}, A	378.3	362.2
R_{min}, X_{min}	378.5	360.9
A, X_{min}	-	366.9
R_{min}	376.5	361.0
A	382.6	371.4
X_{min}	-	365.6
Aucun	395.1	392.2

Tableau 3.9 – AIC et DIC des modèles plateaux-plus-linéaires en fonction du nombre de paramètres aléatoires.

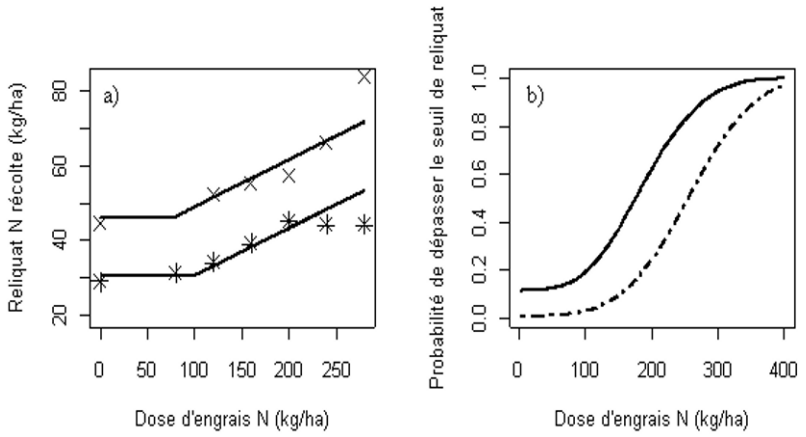

Figure 3.11 – Modèle bayésien du reliquat N récolte avec R_{min} et X_{min} aléatoires. Ajustement obtenu pour deux sites-années avec les paramètres individuels (a). Exemple d'utilisation pratique (b) : probabilités de dépasser des seuils de reliquat récolte de 50 (courbe continue) et 60 kg/ha (courbe en tirets) en fonction de la dose d'engrais appliquée, calculées avec les distributions de probabilité de R_{min} et X_{min}.

La figure 3.11b présente une application pratique basée sur les paramètres de la population. Une série de 1 000 valeurs de R_{min} et X_{min} a été générée à partir de leurs distributions de probabilité. Ces valeurs ont été utilisées pour générer 1 000 courbes de réponse du reliquat d'azote à la dose d'engrais appliquée, puis pour estimer la probabilité d'être supérieur à un seuil de reliquat récolte en fonction de la dose d'engrais appliquée. Deux seuils ont été considérés successivement, 50 et 60 kg/ha. La probabilité de dépasser ces seuils a été estimée par le nombre de valeurs simulées supérieures aux seuils divisé par 1 000. Ce type de calcul permet d'évaluer les risques de pollution nitrique dans les parcelles agricoles. Le code R est donné ci-dessous.

```
# Code R

# Generation de 1000 valeurs de parametres

Rmin.vec <- rnorm(1000, 39.39, sqrt(76.61))
A.vec    <- 0.1266
Xmin.vec <- rnorm(1000, 90.78, sqrt(1524.0))

# Initialisation du vecteur de reliquats simules

Rel.vec <- 1:1000

# Initialisation du vecteur des probabilites (une
#   probabilite par dose testee)

Proba.vec <- 1:400

#Boucles sur doses (i) et sur valeurs de parametres (j)

for (i in 1:400) {
  for (j in 1:1000) {
    Rel.vec[j] <- Mod.PL(i, Rmin.vec[j], A.vec, Xmin.vec[j])
    }

  #Calcul de la probabilite de depasser 50 kg/ha

  Proba.vec[i] <- sum(Rel.vec>=50)/length(Rel.vec)
  }
```

3.4.3 Exemple : variabilité intra-parcellaire des densités de mauvaises herbes

Considérons une parcelle divisée en N quadrats. Supposons que le nombre de mauvaises herbes ait été mesuré sur m des N quadrats une année donnée. Le problème pratique considéré ici est d'estimer la densité de mauvaises herbes sur l'ensemble des N quadrats pour cette même année en utilisant les m mesures déjà disponibles.

Nous proposons ci-dessous un modèle hiérarchique qui permet à la fois de gérer des données de comptage (ici des nombres de mauvaises herbes) et d'intégrer des corrélations spatiales entre les mesures réalisées dans différents quadrats. Le modèle fait appel à la loi de Poisson, qui est souvent utilisée pour modéliser des données de comptage comme indiqué dans la section 3.2.

Étapes *i* et *ii* : définition des variables et des équations

Le modèle bayésien hiérarchique suivant est utilisé pour simuler les densités et nombres de mauvaises herbes dans les N quadrats. Il comporte trois niveaux :

Niveau 1, intra-quadrat : $Y_i \,|\, \mu_i \sim \mathrm{Po}\left(\mu_i s_i\right), i = 1, \ldots, N,$

Niveau 2, inter-quadrats :

$$\begin{aligned} \log\left(\mu_i\right) &= \alpha + \varphi_i + e_i, \\ \varphi_i \,|\, \tau, \bar{\varphi}_i &\sim N\left(\bar{\varphi}_i, \tau/n_i\right), \\ e_i \,|\, \sigma^2 &\sim N\left(0, \sigma^2\right), \quad i = 1, \ldots, N, \end{aligned}$$

Niveau 3, distributions a priori :

$$\begin{aligned} \alpha &\sim \mathrm{dflat}(), \\ 1/\tau &\sim \mathrm{Gamma}(0.1, 0.1), \\ 1/\sigma^2 &\sim \mathrm{Gamma}(0.1, 0.1). \end{aligned}$$

Au niveau 1, Y_i est le nombre de mauvaises herbes observé sur une surface s_i du i-ième quadrat, μ_i est la vraie densité de plantes dans le i-ième quadrat et $\mathrm{Po}(\mu_i s_i)$ désigne la loi de Poisson d'intensité $\mu_i s_i$.

Au niveau 2, on note $\bar{\varphi}_i = \frac{1}{n_i} \sum \varphi_j$, où la somme porte sur les n_i quadrats j situés dans le voisinage du quadrat i. Le fait de conditionner chaque φ_j par la moyenne de ses voisins rend le modèle théoriquement impropre, c'est-à-dire mal défini d'un point de vue probabiliste. En pratique, cela n'empêche pas d'utiliser ce type de modèle, considéré comme une approximation d'un modèle proprement défini. Les erreurs e_i, $i = 1, \ldots, N$, sont supposées indépendantes. Le voisinage d'un quadrat est défini ici par l'ensemble des quadrats adjacents mais d'autres choix sont possibles (Banerjee *et al.*, 2004). Ce deuxième niveau

modélise la variabilité de la densité de mauvaises herbes μ_i entre quadrats. Il comporte un niveau moyen (α), un effet aléatoire décrivant la dépendance spatiale entre les quadrats (φ) et un effet aléatoire correspondant à une erreur résiduelle (e). Une transformation log est utilisée pour contraindre les densités de mauvaises herbes à des valeurs positives. La dépendance spatiale des φ_i, $i = 1, \ldots, N$, est décrite par des lois gaussiennes ; la loi gaussienne décrivant la distribution de chaque terme φ_i est centrée sur la valeur moyenne des φ_j situés dans le voisinage de φ_i, ici dans les quadrats adjacents.

Pour le niveau 3, des distributions *a priori* standard et peu informatives sont utilisées dans cet exemple.

Étape *iii* : estimation des paramètres

Notons $\boldsymbol{\theta}$ le vecteur des paramètres du modèle, $\boldsymbol{\theta} = (\alpha, \tau, \sigma)^T$. Lorsque m mesures $y_1, \ldots, y_m$, sont collectées sur m des N quadrats, ces mesures peuvent être utilisées pour estimer la distribution *a posteriori* de $\boldsymbol{\theta}$, $p(\boldsymbol{\theta} \mid y_1, ..., y_m)$ et pour estimer la distribution *a posteriori* de la densité de mauvaises herbes $p(\mu_i \mid y_1, ..., y_m)$ pour l'ensemble des quadrats, $i = 1, \ldots, N$, y compris ceux qui n'ont pas été observés. Ces distributions peuvent être estimées avec un algorithme MCMC (Banerjee *et al.*, 2004) et l'espérance ou la médiane de $p(\mu_i \mid y_1, ..., y_M)$ peut être utilisée pour estimer la densité de mauvaises herbes de chaque quadrat i, pour $i = 1, \ldots, N$.

On considère une parcelle de blé en 1996 constituée de $N = 92$ quadrats de même surface (5 m $\times$5 m) répartis sur 23 lignes et 4 colonnes. Des mesures de nombre de plantes de vulpin (*Alopecurus myosuroides* Huds) ont été réalisées dans chacun des 92 quadrats. Chaque mesure correspond à un nombre de plantes compté sur une placette de 0.04 m^2 localisée au centre du quadrat (Chauvel *et al.*, 2001, 2009). Les données sont présentées sur la figure 3.12a.

On suppose que le nombre de plantes de vulpin n'a été mesuré que dans $m = 31$ quadrats et on cherche à prédire les densités de vulpin sur les 61 quadrats non observés. Les localisations des 31 quadrats sont présentées sur la figure 3.12b. Le modèle hiérarchique décrit ci-dessus a été ajusté aux 31 données à l'aide de WinBUGS avec 15 000 itérations MCMC (dont une période de chauffe de 5 000 itérations).

```
# Code BUGS: modele Poisson log

model{

for (i in 1:N) {

  #Niveau 1

  O[i] ~ dpois(mu[i])
```

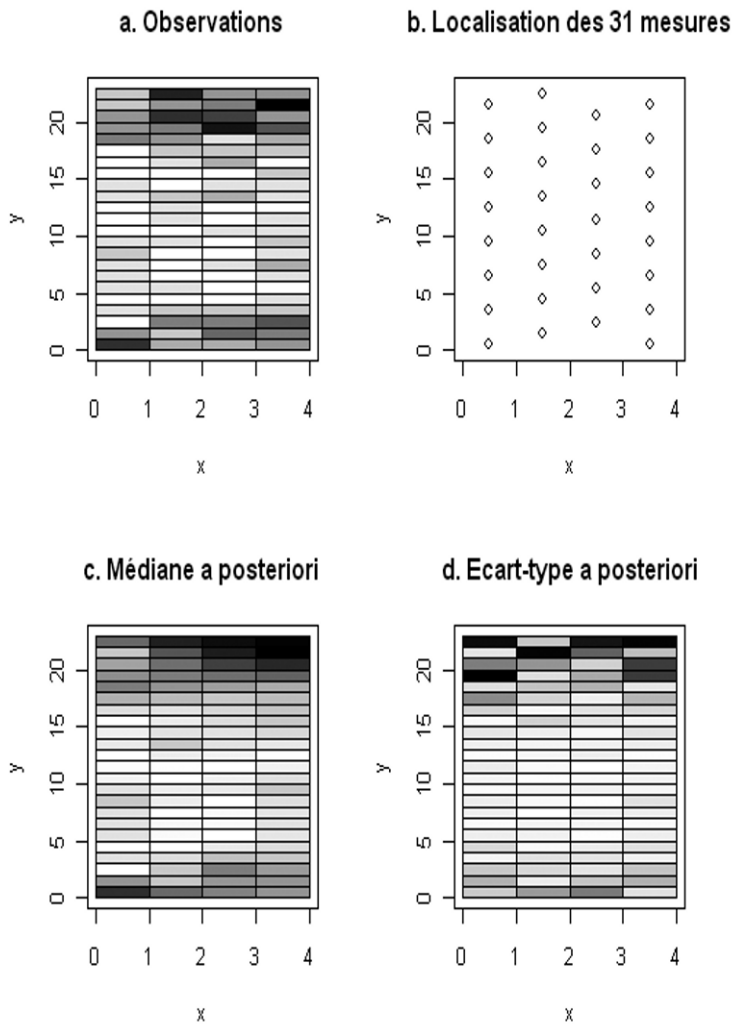

Figure 3.12 – Mesures de nombres de plantes de vulpin réalisées sur les $N = 92$ quadrats de la parcelle (a) ; plan d'expérience des $M = 31$ mesures utilisées pour estimer les paramètres du modèle (b) ; valeurs des médianes *a posteriori* estimées à l'aide du modèle et des 31 mesures (c) ; écarts types *a posteriori* des densités de mauvaises herbes estimées à l'aide du modèle hiérarchique et des 31 mesures (d). La valeur est d'autant plus forte que la couleur est foncée. Pour les figures a et c, la couleur blanche correspond à une densité de plantes nulle et la couleur noire correspond à une densité supérieure à 300 plantes/m^2.

```
  #Niveau 2

  log(mu[i]) <- alpha0+phi[i]+e[i]
  e[i] ~ dnorm(0, tau2)
  }

phi[1:N] ~ car.normal(adj[], weights[], num[], tau1)
for (j in 1:sumNumNeigh) {
  weights[j]<-1 }

#Niveau 3

alpha0 ~ dflat()
tau1 ~ dgamma(0.1, 0.1)
tau2 ~ dgamma(0.1, 0.1)
```

Étape *iv* : évaluation

Les médianes *a posteriori* et écarts types *a posteriori* déterminés à partir des simulations MCMC sont présentées sur les figures 3.12c et 3.12d respectivement pour l'ensemble des 92 quadrats. Les médianes *a posteriori* obtenues sur les 61 quadrats non observés peuvent être utilisées pour prédire les densités de plantes dans ces quadrats. Ces prédictions correspondent aux valeurs ayant 50% de chance d'être dépassées. Elles sont présentées et comparées aux observations correspondantes sur la figure 3.13. Cette figure montre que les prédictions du modèle ont tendance à être plus fortes pour les quadrats où la densité observée est élevée et, inversement, plus faibles pour les quadrats où la densité observée est faible. Les prédictions ne sont cependant pas parfaites ; l'éloignement des points à la bissectrice indique que les erreurs de prédiction peuvent être importantes dans certains quadrats.

Les erreurs de prédiction du modèle peuvent être quantifiées en calculant la racine carrée de l'erreur quadratique moyenne de prédiction (en anglais, RMSE pour *Root Mean Squared Error*) définie par :

$$\mathrm{RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(y_j - \widehat{y}_j)^2}$$

où J est le nombre de valeurs prédites, $\widehat{y}_j$ la j-ième prédiction et y_j l'observation correspondante, $j = 1, \ldots, J$. Lorsque les données utilisées pour calculer le RMSE sont indépendantes de celles utilisées pour développer le modèle, le RMSE est un critère d'évaluation qui indique l'erreur de prédiction moyenne du modèle. Avec nos données, ce critère d'évaluation est égal à RMSE = 47.22 plantes/m^2. Cette valeur confirme que les erreurs de prédiction du modèle peuvent être importantes. Elles pourraient être réduites en augmentant le nombre de quadrats mesurés dans la parcelle.

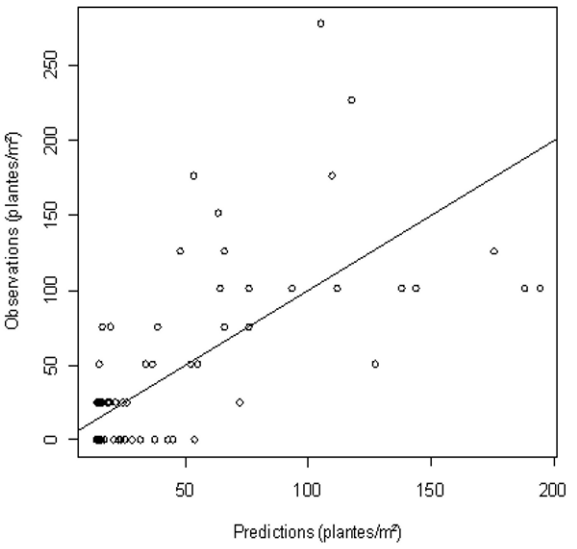

Figure 3.13 – Densités de vulpin observées et prédites par le modèle hiérarchique sur les 61 quadrats non observés. Chaque point correspond à un quadrat.

3.5 Estimation de valeurs extrêmes par régression quantile

3.5.1 Définition

Les modèles présentés dans les sections précédentes permettent de modéliser l'espérance d'une variable réponse Y en fonction d'une ou plusieurs variables d'entrée X. Dans certains cas, ce n'est pas la réponse moyenne qui intéresse le modélisateur mais des réponses plus extrêmes correspondant à des quantiles particuliers. La régression quantile est une technique qui permet de modéliser les quantiles d'une variable d'intérêt en fonction d'une ou plusieurs variables d'entrée (Koenker et Basset, 1978 ; Koenker et Park, 1996 ; Koenker et Machado, 1999). Elle est souvent utilisée en économie (Koenker et Basset, 1978), en écologie (Cade *et al.*, 1999) et, de façon plus récente, en agronomie (Makowski *et al.*, 2007). Les méthodes bayésiennes sont relativement peu développées dans le contexte de la régression quantile et nous nous limiterons donc à des méthodes fréquentistes.

Le principe est de définir $f(X, \boldsymbol{\alpha})$ comme une fonction telle que

$$P[Y < f(X, \boldsymbol{\alpha}) \mid X] = \tau.$$

Avec cette définition, $f(X, \boldsymbol{\alpha})$ représente le quantile de niveau τ de la variable réponse Y. Autrement dit, pour une valeur donnée X des variables d'entrée,

une observation y est inférieure à $f(X, \boldsymbol{\alpha})$ avec une probabilité τ et supérieure ou égale à $f(X, \boldsymbol{\alpha})$ avec une probabilité $1-\tau$. Par exemple, si $\tau = 0.5$, $f(X, \boldsymbol{\alpha})$ correspond à la réponse médiane de Y à X. Si $\tau = 0.9$, $f(X, \boldsymbol{\alpha})$ correspond au neuvième décile.

La fonction f peut être linéaire ou non linéaire. Dans les deux cas, le vecteur des paramètres $\boldsymbol{\alpha}$ peut être estimé à l'aide d'une méthode non paramétrique qui consiste à minimiser la somme pondérée des différences absolues entre observations et prédictions. Pour un quantile donné, un estimateur de $\boldsymbol{\alpha}$ est le vecteur qui minimise (Koenker et Basset, 1978)

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \rho_\tau\big(y_i - f(x_i, \boldsymbol{\alpha})\big), \tag{3.12}$$

où y_i et x_i sont les i-ièmes observations de Y et X, et $\rho_\tau[.]$ est une fonction définie par

$$\begin{cases} \rho_\tau(z) = \tau\, z & \text{si } z \geq 0, \\ \rho_\tau(z) = (1-\tau)\, z & \text{si } z < 0. \end{cases}$$

Plusieurs algorithmes ont été développés pour minimiser (3.12) selon que la fonction f est linéaire ou non linéaire. Ces algorithmes peuvent être appliqués avec la librairie `quantreg` de R.

La démarche habituelle consiste à définir une série de valeurs de quantiles τ, par exemple 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, et à estimer la valeur de $\boldsymbol{\alpha}$ en minimisant (3.12) pour chaque quantile. Le résultat est une série de valeurs estimées des paramètres, $\widehat{\alpha}_{0.5}$, $\widehat{\alpha}_{0.6}$, $\widehat{\alpha}_{0.7}$, ..., $\widehat{\alpha}_{0.95}$. Ces paramètres peuvent être présentés graphiquement ou dans un tableau. Des représentations graphiques des fonctions de réponse correspondantes, $f(X, \widehat{\alpha}_{0.5})$, $f(X, \widehat{\alpha}_{0.6})$, ..., $f(X, \widehat{\alpha}_{0.95})$, peuvent également être réalisées.

Il est recommandé de déterminer les écarts types et intervalles de confiance des estimateurs des paramètres (*e.g.*, Koenker et Basset, 1978). Quand la fonction f est non linéaire, les écarts types et intervalles de confiance peuvent être calculés par bootstrap comme illustré par Makowski *et al.* (2007), c'est-à-dire en ré-échantillonnant un grand nombre de fois avec remise dans les données et en appliquant la procédure d'estimation à chaque échantillon généré de cette manière. Les écarts types et intervalles de confiance permettent d'identifier les quantiles les plus extrêmes qui peuvent être estimés avec une précision restant raisonnable. L'intérêt de l'approche non paramétrique décrite ci-dessus est qu'elle ne fait aucune hypothèse sur la distribution de probabilité des observations.

3.5.2 Exemple : risque de sclérotinia du colza

Sclerotinia sclerotiorum Lib. de Bary est un champignon qui attaque le colza. Au printemps, les sclérotes forment des apothécies qui libèrent des spores

dans l'atmosphère. Ces spores disséminées vont polluer et contaminer les pétales des fleurs de colza qui, en tombant, se fixent sur les feuilles et permettent au mycélium de coloniser le limbe de la feuille, puis le pétiole et la tige.

Le sclérotinia peut être à l'origine de pertes de rendement importantes certaines années mais l'incidence de sclérotinia est très variable. Dans de nombreuses situations, l'utilisation de fongicide est inutile et accroît les risques de pollution de l'environnement. Il est donc important de pouvoir évaluer les risques de sclérotinia dès la floraison du colza afin de pouvoir décider de l'opportunité d'un traitement fongicide.

Dans cette application, deux modèles sont développés. Le premier modèle permet de prédire le pourcentage de plantes qui seront malades à la récolte (appelé incidence de la maladie) à partir d'un pourcentage de fleurs malades mesuré à la floraison. Ce modèle peut être utilisé pour décider de l'application d'un traitement fongicide à la floraison sur la base d'une estimation de l'incidence finale de la maladie. Le deuxième modèle permet d'estimer les pertes de rendement induites par un niveau donné d'incidence de sclérotinia dans une culture de colza. Ce deuxième modèle ne peut pas être utilisé pour raisonner un traitement fongicide mais permet d'estimer le seuil de nuisibilité de la maladie, c'est-à-dire le niveau d'incidence au-delà duquel les pertes de rendement sont importantes.

Étape *i.* Définition des variables

La variable d'entrée du premier modèle, M_1, est la proportion de fleurs malades mesurées à floraison et sa variable de sortie est la proportion de plantes malades mesurées à la récolte. Pour le second modèle, M_2, la variable d'entrée est la proportion de plantes malades mesurées à la récolte et la variable de sortie est le rendement du colza.

Étape *ii.* Définition des équations

Chaque modèle est basé sur une équation linéaire $f(X, \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 X$, où $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$ est le vecteur des paramètres à estimer et X est la variable d'entrée, égale selon le modèle, soit à la proportion de fleurs malades à floraison, soit à la proportion de plantes malades à la récolte.

Étape *iii.* Estimation des paramètres

Données Nous utilisons les données issues de 759 sites-années de culture de colza non traitée contre le sclérotinia. Ces données ont été collectées par le Cetiom (`www.cetiom.fr`) au cours des années 2002-2006 sur de nombreux sites répartis dans les principales zones de culture de colza en France. La fraction de fleurs malades a été mesurée sur chaque site-année à partir de 40 à 80 fleurs collectées en début de floraison. La fraction de plantes malades a également été mesurée sur chaque site-année à partir de 200 plantes collectées quelques jours avant la récolte. Le rendement du colza a été mesuré sur 155 des 759

sites-années. Voir Makowski *et al.* (2005) et Makowski *et al.* (2008) pour une description détaillée des données.

Régression quantile Les paramètres des modèles M_1 et M_2 ont été estimés pour les quantiles $\tau = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95$. Cela revient à définir la contrainte

$$P\big(Y < \alpha_0 + \alpha_1 X \mid X\big) = \tau$$

pour chaque modèle et chaque quantile. L'estimation a été réalisée avec la fonction `rq` de la librairie `quantreg` en utilisant le programme R décrit ci-dessous en trois parties.

```
#Code R : lecture des donnees

library(quantreg)

TAB <- read.table("f:\\Programmes\\ReqQuantSclero\\Sclero.txt",
                  header=T, sep="\t")

KIT <- TAB$KIT
TX  <- TAB$TXATTNT
RDT <- TAB$RDTNT

TAB.1 <- data.frame(KIT,TX)
TAB.1 <- TAB.1[is.na(TX)==F & is.na(KIT)==F,]

TAB.2 <- data.frame(TX,RDT)
TAB.2 <- TAB.2[is.na(RDT)==F & is.na(TX)==F,]
```

La première ligne du programme permet de charger la librairie `quantreg` qui doit être préalablement installée depuis internet. Les données sont ensuite lues à partir d'un fichier `.txt`. Les données manquantes sont alors éliminées et deux tableaux de données sont créés, un pour chaque modèle.

```
#Code R: Estimation avec rq

Tau.vec <- c(0.5, 0.6, 0.7, 0.8, 0.9, 0.95)

ParaMat.1 <- matrix(ncol=2, nrow=6)
ParaMat.2 <- matrix(ncol=2, nrow=6)

for (i in 1:6) {

  Fit <- rq(TX ~  KIT, data=TAB.1, tau=Tau.vec[i])
  print(summary(Fit))
  ParaMat.1[i,] <- coef(Fit)
```

```
  Fit <- rq(RDT ~  TX, data=TAB.2, tau=Tau.vec[i])
  print(summary(Fit))
  ParaMat.2[i,] <- coef(Fit)

  }
```

Les paramètres sont estimés pour chaque quantile et chaque modèle en utilisant une boucle `for`. Les valeurs estimées sont affichées avec `summary` et sont archivées dans les matrices avec l'instruction `coef`.

```
# Code R: realisation des graphiques

par(mfrow=c(2,2))

x.vec <- 1:100
x.vec <- x.vec/100

plot(TAB.1$KIT, TAB.1$TX, xlab="Proportion de fleurs malades",
                 ylab="Proportion de plantes malades recolte")

for (i in 1:6) {
  lines(x.vec, ParaMat.1[i,1]+ParaMat.1[i,2]*x.vec,lwd=2)
  }

plot(TAB.2$TX, TAB.2$RDT,
      xlab="Proportion de plantes malades recolte",
      ylab="Rendement (q/ha)")

for (i in 1:6) {
  lines(x.vec, ParaMat.2[i,1]+ParaMat.2[i,2]*x.vec,lwd=2)
  }

plot(Tau.vec, ParaMat.1[,2], xlab="Quantile", ylab="Pente 1",
                                                    type="b")

plot(Tau.vec, ParaMat.2[,2], xlab="Quantile", ylab="Pente 2",
                                                    type="b")
```

Les données expérimentales sont affichées avec l'instruction `plot`, puis les régressions quantiles sont superposées aux données avec l'instruction `lines` en utilisant une boucle `for`. Finalement, les valeurs estimées des paramètres sont affichées en fonction des quantiles, en utilisant deux autres instructions `plot`. Les graphiques sont présentés sur la figure 3.14 pour M_1 et M_2.

Concernant le modèle reliant la proportion de plantes malades récolte à la proportion de fleurs malades (figure 3.14ac), le graphique montre que la

proportion de plantes malades augmente en fonction de la proportion de fleurs malades. La pente est d'autant plus grande que le quantile est élevé, ce qui est lié au fait que la variabilité de la réponse augmente avec le nombre de fleurs malades. Lorsque le pourcentage de fleurs malades est inférieur à 20%, l'incidence reste relativement faible dans toutes les parcelles, toujours inférieur à 40% et le plus souvent inférieur à 20%. Par contre, lorsque le pourcentage de fleurs malades est supérieur à 80%, le niveau d'incidence varie entre 0 et 100%. Concernant le modèle reliant le rendement à la proportion de plantes malades (figure 3.14bd), le graphique montre que le rendement est d'autant plus faible que la fraction de plantes malades est forte. Par contre, la pente n'évolue que faiblement en fonction du quantile.

Les réponses présentées sur la figure 3.14a peuvent être utilisées pour déterminer la probabilité d'atteindre ou dépasser une proportion donnée de plantes malades à la récolte, connaissant la proportion de fleurs malades. Par exemple, lorsque le pourcentage de fleurs malades est de 80%, l'incidence de la maladie à 5% de chance de dépasser 60%. La proportion de fleurs malades étant mesurable en début de floraison, la figure 3.14a peut être utilisée pour anticiper le risque de sclérotinia et, si nécessaire, décider d'un traitement fongicide.

Les réponses de la figure 3.14b ont un autre intérêt. Elles sont utiles pour analyser les pertes de rendement pouvant être induite pas le sclérotinia, sans se limiter à une valeur moyenne mais en considérant des risques à différents niveaux de probabilité. Ces réponses peuvent ainsi permettre la détermination de seuils de nuisibilité en fonction de la probabilité de subir une perte de rendement donnée.

Étape *iv.* Évaluation

Les résultats de la régression quantile peuvent être évalués en déterminant les intervalles de confiance des estimateurs des paramètres et en réalisant des tests statistiques (Koenker et Basset, 1978 ; Cade *et al.*, 1999 ; Makowski *et al.*, 2007). Lorsque plusieurs équations candidates sont disponibles, il est également utile de les comparer à l'aide des critères proposés par Koenker et Machado (1999). Dans notre exemple, les pentes des régressions sont toutes significativement différentes de zéro, sauf celle de la réponse du rendement correspondant au quantile 0.95. Ce résultat indique que cette pente n'est pas estimée précisément, probablement à cause d'un nombre de données insuffisant.

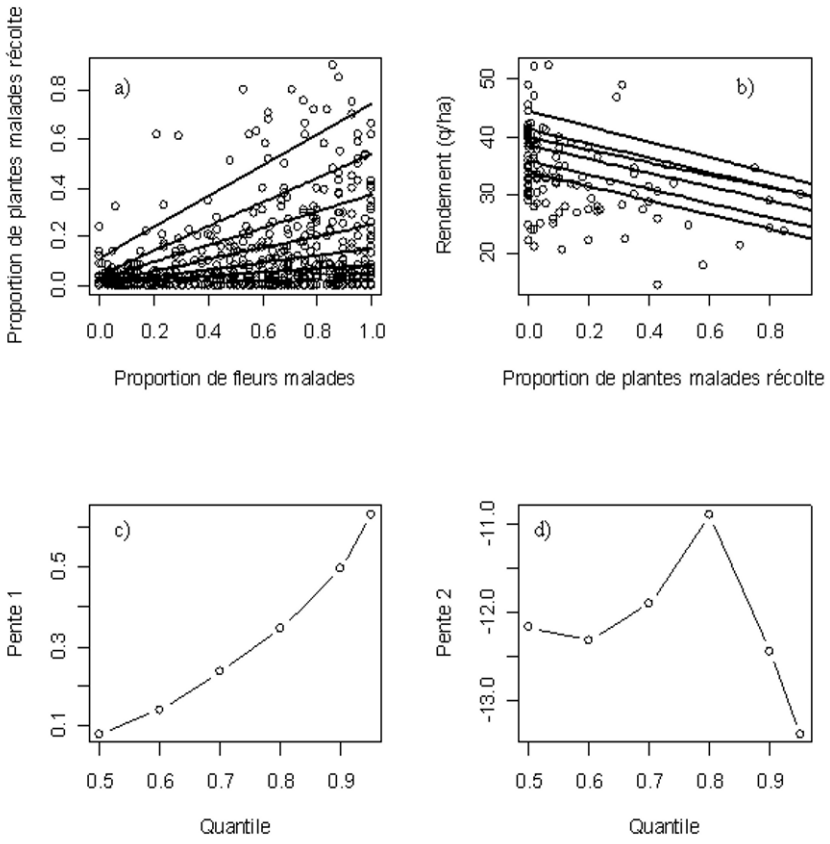

Figure 3.14 – Résultats de la régression quantile pour les modèles M_1 (a, c) et M_2 (b, d). a) et b) présentent les réponses pour les quantiles (de bas en haut) 0.5, 0.6, 0.7, 0.8, 0.9, 0.95. c) et d) présentent les valeurs estimées des pentes.

3.6 Exercices

Exercice 3.1 La méligèthe est un petit insecte noir qui attaque le colza mais aussi certaines plantes ornementales des jardins sur lesquelles elle peut causer de graves dégâts. Un bureau d'étude vous demande de modéliser la présence de méligèthes dans un jardin en fonction de la distance du jardin au champ de colza le plus proche. Vous disposez pour cela d'une base de données incluant les résultats d'expérimentations conduites dans 142 jardins en France. Trois variables ont été mesurées dans chaque jardin : la présence/absence de méligèthes, le nombre moyen de méligèthes par bouton de rose, la distance au champ de colza le plus proche.

- Quel(s) modèle(s) allez-vous utiliser ?
- Écrire leurs équations.
- Comment estimer leurs paramètres ?
- Comment évaluer le(s) modèle(s) ?

Exercice 3.2 Définir un modèle linéaire reliant le nombre de grains par m^2 d'une culture de blé à deux facteurs limitants potentiels (indice de nutrition azotée au stade épi 1 cm et densité de mauvaises herbes par m^2 au même stade) et au type variétal (précoce, tardif).

- Déterminer la nature des variables d'entrée et de sortie.
- Écrire le modèle dans sa version fréquentiste et bayésienne.
- Écrire le code R permettant d'estimer les paramètres par les moindres carrés en supposant que des mesures des variables d'entrée et de sorties ont été collectées sur N sites-années pour estimer les paramètres et que 50% de ces sites-années correspondent à chacun des deux types variétaux.
- Écrire le code WinBUGS permettant d'estimer les paramètres par MCMC.

Exercice 3.3 Vous souhaitez prédire la présence/absence d'une espèce végétale d'intérêt écologique dans une prairie. Vous disposez de N mesures de présence/absence de cette espèce obtenue dans N prairies différentes. La date d'implantation de la prairie et le nombre moyen de coupes annuel sont connus pour chaque prairie. Construire un modèle pour prédire la présence/absence de l'espèce dans une nouvelle prairie en fonction de sa date d'implantation et de son nombre moyen de coupes.

- Quel type de modèle pouvez-vous utiliser ?
- Combien de modèles pouvez-vous définir ? Écrire leurs équations.
- Comment estimer les paramètres ?
- Comment évaluer les modèles et choisir le(s) meilleur(s) ?

Exercice 3.4 La norme internationale ISPM15 pour le traitement thermique du bois destiné à l'exportation consiste à chauffer le bois à 56 °C pendant 30 min. Vous souhaitez évaluer l'efficacité de ce traitement thermique pour éliminer un insecte nuisible, *Agrilus planipennis*, en utilisant les données expérimentales de Myers, Fraser et Mastro (2009). Dans cette expérimentation, des

pièces de bois de frêne ont été exposées à cinq températures différentes pendant 30 min. Le nombre d'insectes vivants mesuré sur les pièces de bois après le traitement thermique est égal à 182, 112, 43, 0, 0 pour les températures 45, 50, 55, 60, 65 °C respectivement.

- Définir un modèle Poisson reliant le nombre d'insectes vivants à la température.
- Estimer les paramètres du modèle avec les données.
- Utiliser le modèle pour estimer le nombre moyen d'insectes vivants après application du traitement thermique défini par la norme ISPM15 (56 °C pendant 30 min).
- Utiliser le modèle pour calculer la probabilité d'avoir au moins un insecte vivant après application de ce traitement.

Exercice 3.5 Des longueurs de limbes ont été mesurées dans un champ de maïs à différentes dates sur différentes plantes mais toujours sur le même phytomère. La somme de températures depuis le semis a été déterminée à chaque date de mesure. Une à trois mesures de longueur de limbe ont été obtenues à chaque date. Les mesures sont présentées sur la figure 3.15 en fonction de la somme de températures.

- Quel type de modèle(s) peut-on utiliser pour modéliser la longueur du limbe en fonction de la somme de température ?
- Définissez un ou plusieurs modèles.
- Quelle(s) méthode(s) utiliser pour estimer les paramètres ?

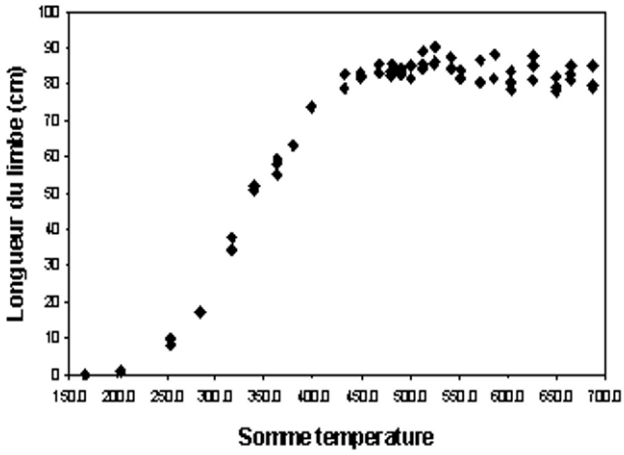

Figure 3.15 – Longueur du limbe en fonction de la somme de températures sur des plantes de maïs.

Exercice 3.6 Une expérimentation a été réalisée sur une parcelle de blé homogène pour étudier la réponse du rendement à la dose d'engrais azoté. La parcelle a été divisée en six sous-parcelles de même surface. Chaque sous-parcelle a reçu une dose totale d'engrais azoté égale à 0, 50, 100, 200, 250 ou 300 kg/ha. Les rendements obtenus pour ces doses sont respectivement 5.1, 7.5, 9.2, 9.8, 9.6, 9.7 t/ha.

- Créer un tableau de données sous R incluant deux colonnes, une pour les doses d'engrais, une pour les rendements.
- Présenter graphiquement les données.
- Définir une ou plusieurs équations pour modéliser le rendement à la dose d'engrais.
- Estimer les paramètres de ces modèles avec R en utilisant une méthode de votre choix.
- Choisir le modèle qui s'ajuste le mieux aux données.

Exercice 3.7 Supposons que l'expérience décrite dans l'exercice 3.6 ait été répétée dans 141 parcelles de blé supplémentaires.

- Quel(s) type(s) de modèles pouvez-vous utiliser pour modéliser la réponse du rendement à la dose sur l'ensemble des parcelles ?
- Écrire leurs équations.
- Comment estimer leurs paramètres avec R ou WinBUGS ?

Chapitre 4

Optimisation des décisions et gestion des risques

4.1 Les quatre étapes de l'optimisation

4.1.1 Présentation

Les méthodes d'optimisation de la décision sont utiles pour la deuxième étape de l'analyse des risques : *la gestion des risques*. Elles permettent d'identifier des solutions techniques conduisant à une diminution de la probabilité d'occurrence d'un événement potentiellement dommageable (*e.g.*, perte de rendement, perte de revenu, pollution) ou à la diminution de son impact. Le terme « optimisation » ne doit cependant pas faire croire à l'opérateur que le résultat fournit la vraie décision optimale. Il faut par conséquent non seulement estimer quelle est la décision optimale, mais aussi évaluer la qualité de cette estimation et rendre aussi explicites que possible les hypothèses sur lesquelles elle repose.

Dans ce chapitre, la décision est représentée par une variable dite décisionnelle sur laquelle un opérateur peut agir. Dans notre contexte, une variable décisionnelle correspond, par exemple, à un nombre de traitements fongicides, une dose d'engrais ou une surface affectée à un type de culture. L'optimisation d'une variable décisionnelle consiste à déterminer la valeur de cette variable qui permet d'optimiser un ou plusieurs objectifs, par exemple de maximiser le revenu des agriculteurs, d'atteindre un niveau de qualité satisfaisant pour une production, d'éviter une perte de rendement ou de respecter des contraintes environnementales.

La procédure d'optimisation d'une variable décisionnelle se décompose en quatre étapes :

1. définition de la variable décisionnelle et si nécessaire de la règle de décision (*e.g.*, dose d'engrais, décision d'appliquer un fongicide) ;
2. définition précise des objectifs de l'optimisation (*e.g.*, maximiser un

rendement ou une marge économique) ;

3. à partir d'informations existantes ou à acquérir, calcul des valeurs optimales de la variable décisionnelle ou des paramètres de la règle de décision ;
4. évaluation du résultat.

La première étape permet de formaliser la prise de décision. Dans les cas les plus simples, la décision est le choix d'une unique valeur $d^\star$ de la variable décisionnelle D (dose d'engrais à appliquer, par exemple), cette valeur étant considérée comme optimale pour toutes les situations envisagées. Mais dans de nombreux cas, la décision est prise en fonction des valeurs d'une ou plusieurs variables I (mesure de la fertilité du sol, par exemple), appelées indicateurs, qui fournissent une information sur l'état du milieu. Le but est alors de définir une fonction $d^\star(I)$ donnant la décision à appliquer en fonction de la valeur du ou des indicateurs I. Une telle fonction correspond à une *règle de décision* qui permet d'adapter les valeurs des variables décisionnelles aux caractéristiques du milieu (*e.g.*, adapter la dose d'engrais à la fertilité du sol). Une règle de décision peut inclure un ou plusieurs paramètres dont les valeurs doivent être estimées.

La deuxième étape consiste à définir la ou les variables par lesquelles sont mesurés les impacts des décisions (*e.g.*, revenu de l'agriculteur), puis à formuler les critères à optimiser sur ces variables et les contraintes à respecter (*e.g.*, maximiser ce revenu sans dépasser une dose maximale d'engrais). Le lien avec la décision est fait en élaborant la *fonction objectif*, qui exprime le critère à optimiser en fonction des variables décisionnelles et des indicateurs éventuels.

La troisième étape requiert un *algorithme d'optimisation*. Le choix de l'algorithme dépendra du nombre et de la nature des variables décisionnelles (binaires, continues), et de la nature de la ou des fonctions objectifs. Lorsque l'objectif est d'aboutir à une règle de décision, l'optimisation consiste à déterminer les valeurs optimales des paramètres de la règle. Cette étape nécessite généralement un travail de modélisation et d'estimation statistique pour préciser les relations entre les variables d'impact, les variables décisionnelles, les indicateurs et parfois d'autres variables reflétant l'état du milieu, mais différentes des indicateurs car supposées inconnues de l'opérateur.

La dernière étape consiste à évaluer la qualité du résultat de l'optimisation, c'est-à-dire à déterminer la précision des valeurs calculées des variables décisionnelles et de leurs impacts. Comme dans le cas de la modélisation, cette étape d'évaluation pourra aboutir à une remise en cause de la procédure utilisée.

L'optimisation est souvent réalisée avec des méthodes issues d'une autre discipline que la statistique : *la recherche opérationnelle* (*e.g.*, Winston, 1994). Les techniques issues de cette discipline, comme la programmation linéaire, sortent du champ couvert par cet ouvrage. Les méthodes disponibles sont nombreuses et nous encourageons le lecteur intéressé par cette question à consulter des ouvrages spécialisés en recherche opérationnelle.

La statistique joue également un rôle important dans la définition de règles de décision, notamment dans le cadre de certains problèmes d'optimisation de règles de décision binaires et à travers la théorie de la décision (Lindsay, 2004). Nous présentons ici des techniques d'optimisation et des applications mobilisant des concepts issus de la statistique, dont certaines méthodes présentées dans les chapitres précédents. Après une illustration des différentes étapes sur un exemple simple, nous détaillons l'optimisation d'une règle de décision binaire par analyse ROC (section 4.2), puis l'optimisation d'une variable décisionnelle par simulations intensives (section 4.3).

4.1.2 Exemple : détermination d'une température optimale pour le traitement thermique du bois destiné à l'exportation

Le bois destiné à l'exportation doit souvent être soumis à un traitement thermique avant d'être exporté. Un tel traitement permet de limiter les risques d'invasion du pays importateur par des organismes nuisibles (nématodes, insectes, champignons). L'efficacité du traitement dépend de sa durée et de la température utilisée et il est important d'optimiser ces deux paramètres.

On s'intéresse au cas particulier d'un insecte nuisible appelé *Agrilus planipennis* s'attaquant au bois de frêne. Pour mener le travail d'optimisation du traitement, on dispose des données issues de l'expérience conduite par Myers *et al.* (2009). Des pièces de bois de frêne ont été exposées à cinq températures différentes pendant 30 min. Le nombre d'insectes vivants mesuré sur les pièces de bois après le traitement thermique était égale à 51, 35, 12, 0, 0 par m^2 de bois pour les températures 45, 50, 55, 60, 65 °C respectivement.

Etape *i.* Définition de la règle de décision

La durée du traitement est fixée à 30 min. La seule variable décisionnelle est la température T de traitement thermique pour contrôler l'insecte *Agrilus planipennis* dans le bois de frêne. Il n'y a pas d'indicateur dans cet exemple : on cherche une unique valeur t^* de T supposée optimale sur l'ensemble des situations à considérer.

Étape *ii.* Définition de l'objectif

L'impact de la décision se mesure par la variable aléatoire « nombre d'insectes vivants par m^2 après le traitement », notée Y. Il se mesure aussi par la température, que l'on veut minimiser. L'objectif précis qui est retenu est un compromis : on choisit de déterminer la plus petite température qui permet de retrouver au plus un insecte vivant par m^2 avec une probabilité au moins égale à 0.8. Ce niveau de probabilité a été choisi pour illustrer la démarche. D'autres niveaux peuvent être considérés.

Formellement, on recherche la température t^* qui minimise t sous la contrainte $P(Y \leq 1 \mid T = t) \geq 0.8$. Un examen rapide des données de Myers *et al.* (2009) suggère que t^* se situe vraisemblablement entre 50 °C et 75 °C .

Étape *iii.* Optimisation

Pour relier le nombre d'insectes vivants Y après le traitement (variable d'impact) à la température t appliquée pendant ce traitement (variable décisionnelle), on utilise un modèle linéaire généralisé Poisson log-linéaire, déjà évoqué dans les sections 3.2 et 3.4. Le modèle est défini par :

$$\begin{cases} Y \sim \text{Poisson}(\lambda_t) \\ \log(\lambda_t) = \alpha + \beta t \end{cases}$$

avec λ_t l'espérance du nombre d'insectes survivants, α et β deux paramètres.

Les paramètres α et β sont estimés par maximum de vraisemblance avec la fonction `glm` de R, en utilisant le code ci-dessous appliqué aux données de Myers *et al.* (2009). L'algorithme utilisé dans cet exemple pour optimiser t est très simple :

- calculer $P(Y \leq 1 \mid T = t)$ pour une série de valeurs t de T entre 50 °C et 75 °C ;
- retenir pour t^* la plus petite valeur t satisfaisant la contrainte « $P(Y \leq 1 \mid T = t) \geq 0.8$ ».

```
X <- c(45, 50, 55, 60, 65)
Y <- c(51, 35, 12, 0, 0)

Model.0 <- glm(Y~1, family=poisson)
summary(Model.0)

Model <- glm(Y~X, family=poisson)
summary(Model)

TempTest <- 50:75
ProbaTempTest <- ppois(1,lambda=exp(coef(Model)[1]+
                 coef(Model)[2]*TempTest))
print(ProbaTempTest)

Temp.opt <- TempTest[ProbaTempTest>=0.8]
print(Temp.opt)

plot(TempTest, ProbaTempTest)
abline(h=0.8)
abline(v=68)
```

Les valeurs estimées de α et β sont égales à 12.479 et -0.186 et les écarts types des estimateurs sont égaux à 1.07 et 0.022 respectivement. Selon ces

estimations, la plus petite température entière respectant la contrainte « $P(Y \leq 1 \mid T = t) \geq 0.8$ » est 69 °C et il y a 84.9% de chance que le nombre d'insectes survivants après un traitement thermique à 69 °C pendant 30 min soit inférieur à 1 par m^2. Si on se limite à des valeurs entières, la décision optimale est donc $t^* = 69\,°\mathrm{C}$.

Les paramètres du modèle peuvent également être estimés avec une méthode bayésienne en utilisant le code WinBUGS ci-dessous.

```
model
  {
    for( i in 1 : Temp) {
      y[i] ~ dpois(mu[i])
      log(mu[i]) <- alpha + beta * x[i]
      }
    alpha ~ dnorm(0.0,1.0E-6)
    beta ~ dnorm(0.0,1.0E-6)
  }

list(alpha=0, beta=0)
list(Temp=5, y= c(51, 35, 12, 0, 0), x= c(45, 50, 55, 60, 65))
```

Les résultats obtenus avec WinBUGS sont très proches de ceux obtenus avec la fonction `glm` de R ; les moyennes *a posteriori* de α et β obtenues après 100 000 itérations sont égales à 12.63 et -0.1896 et les écarts types *a posteriori* sont égaux à 1.116 et 0.023 pour α et β respectivement.

Étape *iv.* Évaluation

Le modèle décrit ci-dessous a été comparé à un modèle plus simple n'incluant pas la température comme variable explicative. La valeur d'AIC du modèle simplifié (ajusté par maximum de vraisemblance) est 143.86 alors que celle obtenue pour le modèle tenant compte de la température est 36.62. L'effet de la température est significatif avec une probabilité d'erreur de première espèce de 0.001. Parmi les deux modèles testés, le modèle tenant compte de la température est donc le plus adapté.

Une évaluation plus complète pourrait tenir compte du fait que les paramètres α et β ne sont pas connus de façon exacte mais estimés à partir d'un nombre limité de données. Nous n'entrerons pas dans ces détails sur cet exemple.

4.2 Optimisation d'une règle de décision binaire par analyse ROC

4.2.1 Introduction

Les règles de décision binaires sont souvent utilisées en protection des cultures (Yuen *et al.*, 1996 ; Hughes *et al.*, 1999 ; Makowski *et al.*, 2005 ; Primot *et al.*, 2006). La variable de décision D correspond alors à une décision de traitement fongicide, herbicide ou mécanique ($D = 1$ si on décide d'appliquer le traitement, $D = 0$ sinon). La décision se fonde souvent sur un *indicateur* I, par exemple une mesure précoce d'incidence d'une maladie, une mesure de densité de mauvaises herbes ou la prédiction d'un modèle. Lorsque la variable I dépasse un seuil de décision I_s, un traitement est recommandé. Dans le cas inverse, aucun traitement n'est prescrit. L'intérêt de ce type de règle de décision est qu'elle permet d'ajuster la décision aux caractéristiques du milieu et, ainsi, d'éviter des traitements systématiques.

La méthode d'analyse appelée *Receiver Operating Characteristic* (ROC) est une méthode statistique permettant à la fois d'évaluer la précision d'une règle de décision binaire, de comparer plusieurs indicateurs et d'optimiser leurs seuils de décision (Swets, 1988 ; Murtaugh, 1996). Cette méthode est très utilisée dans le domaine médical (Pepe, 1998 ; Pepe, 2003). Le principe de l'analyse ROC est décrit ci-dessous. Il suppose que l'on dispose d'un jeu de données comprenant pour chaque situation une mesure de l'indicateur et la connaissance de la bonne décision.

4.2.2 Règle de décision binaire et ses deux types d'erreur

De façon générale, une règle de décision binaire basée sur un indicateur qui prend des valeurs numériques est défini de la façon suivante :

$$\begin{cases} D = 0 & \text{si } I < I_s, \\ D = 1 & \text{si } I \geq I_s, \end{cases}$$

avec D la variable décisionnelle, I l'indicateur dont la valeur est déterminée pour la situation concernée (*e.g.*, une parcelle agricole) et I_s le seuil de décision. La mise au point de la règle décrite ci-dessus suppose de déterminer deux éléments : l'indicateur I le plus approprié et la valeur optimale du seuil I_s.

Une règle de décision binaire peut conduire à deux types d'erreur, c'est-à-dire deux types de mauvaise décision, appelés faux positif et faux négatif. Pour les décrire, on définit une variable R de référence correspondant à la véritable décision optimale. Par exemple, dans le cadre de la protection des cultures, $R = 1$ si la situation nécessite réellement un traitement vis-à-vis d'un objectif donné et $R = 0$ sinon. Dans ce cadre, *un faux positif* correspond au cas où la règle de décision conduit à $D = 1$ ($I \geq I_s$) alors que la meilleure décision est $R = 0$. Un *faux négatif* correspond au cas où la règle de décision conduit à $D = 0$ ($I < I_s$) alors que la meilleure décision est $R = 1$.

Les fréquences de faux positif et de faux négatif dépendent de la variable I utilisée par la règle de décision et du seuil I_s. Il est important de choisir un indicateur I et un seuil I_s conduisant à des fréquences d'erreur aussi faibles que possible. Deux critères sont fréquemment utilisés pour quantifier les deux types d'erreurs et exprimer les objectifs de l'optimisation de la règle de décision. Il s'agit de la *sensibilité* et la *spécificité* définies par :

$$\begin{aligned}
\text{Sensibilité} &= P(D = 1 \mid R = 1) = 1 - P(D = 0 \mid R = 1) \\
&= P(I \geq I_s \mid R = 1), \\
\text{Spécificité} &= P(D = 0 \mid R = 0) = 1 - P(D = 1 \mid R = 0) \\
&= P(I < I_s \mid R = 0).
\end{aligned}$$

La sensibilité est égale à « 1 moins la probabilité de faux négatif » (*i.e.*, à la probabilité de vrai positif) et la spécifité est égale à « 1 moins la probabilité de faux positif » (*i.e.*, à la probabilité de vrai négatif). Dans le cadre de la protection des cultures, la sensibilité correspond à la probabilité que la règle de décision recommande un traitement dans les situations où un tel traitement est réellement nécessaire. La spécificité correspond à la probabilité de ne pas recommander de traitement dans les situations où un tel traitement est inutile. La valeur de « 1− Spécificité » représente la proportion de traitements inutiles et potentiellement dommageables pour l'environnement.

4.2.3 Estimation et évaluation par la méthode ROC

Supposons que N situations expérimentales (sites-années) soient disponibles et que les valeurs d'une variable binaire de référence R et les valeurs de I aient été déterminées sur chacune. Les N situations expérimentales sont ainsi divisées en deux sous-groupes selon que $R = 1$ ou $R = 0$. Dans l'exemple de la protection des cultures, la valeur de R est déterminée *a posteriori* à partir d'une mesure de perte de rendement ou d'incidence de la maladie à un stade clé sur des parcelles non traitées (Hughes *et al.*, 1999).

Plusieurs techniques statistiques, fréquentistes et bayésiennes, ont été proposées pour estimer les valeurs de sensibilité et de spécificité (Pepe, 1998 ; Farragi et Reiser, 2002 ; Pepe, 2003). Certaines sont non paramétriques et ne nécessitent aucune hypothèse sur la distribution des valeurs de l'indicateur I dans la population de sites-années. D'autres sont au contraire des méthodes paramétriques basées sur une modélisation de la distribution des valeurs de I dans le cas où $R = 1$ et dans le cas où $R = 0$. L'intérêt de l'approche paramétrique est qu'elle permet d'étudier l'effet d'une ou plusieurs covariables sur le niveau de précision de la règle de décision. Sa mise en œuvre est cependant plus délicate et nous ne présentons ici que l'approche non paramétrique. Le lecteur intéressé par les approches paramétriques pourra consulter, par exemple, l'article de Pepe (1998) pour une approche fréquentiste et les articles de Choi *et al.* (2006) ou Makowski *et al.* (2008) pour l'approche bayésienne.

L'approche non paramétrique consiste à estimer de façon empirique la sensibilité et la spécificité. Pour un seuil I_s donné, on compare pour chacune des N situations expérimentales disponibles la valeur de l'indicateur I au seuil I_s. On en déduit la répartition des situations selon les quatre catégories résumées dans le tableau 4.1 : vrais négatifs (VN), faux négatifs (FN), faux positifs (FP), vrais positifs (VP). Les résultats sont utilisés pour calculer la proportion PVP de vrais positifs parmi les cas $R = 1$ (PVP = VP/(FN + VP)) et la proportion PVN de vrais négatifs parmi les cas $R = 0$ (PVN = VN/(VN + FP)). La sensibilité est alors estimée par PVP et la spécificité est estimée par PVN.

Décision donnée par la règle	**Décision optimale**	
	$R = 0$	$R = 1$
$D = 0$ $(I < I_s)$	Vrais négatifs (VN)	Faux négatifs (FN)
$D = 1$ $(I \geq I_s)$	Faux positifs (FP)	Vrais positifs (VP)

Tableau 4.1 – Répartition des situations selon la décision optimale et la décision prise.

Cette technique peut être répétée pour tous les seuils I_s possibles. Pour des valeurs croissantes du seuil I_s, la spécificité augmente mais la sensibilité diminue. L'ensemble des valeurs obtenues peut être représenté sous la forme d'une courbe, appelée courbe ROC, qui relie la sensibilité à la spécificité lorsque I_s varie (voir la figure 4.1c). Une courbe qui passe près du point (1, 1) montre que l'indicateur donne un résultat satisfaisant en terme de sensibilité et de spécificité. Une telle courbe indique qu'il est possible d'atteindre à la fois un bon niveau de sensibilité et de spécificité en choisissant un seuil de décision adapté. Par contre, une courbe qui passe près de la droite reliant les points (0, 1) et (1, 0) montre que l'indicateur n'est pas plus utile qu'une règle de décision aléatoire. Souvent, la sensibilité est présentée en fonction de « 1 – Spécificité » et le point optimal a alors comme coordonnées (0, 1).

Lorsque plusieurs indicateurs candidats sont disponibles, il est utile de résumer les performances de chacun en calculant l'aire sous la courbe ROC. Celle-ci est comprise entre zéro et un. Elle peut s'interpréter comme la probabilité que la valeur de l'indicateur pour une situation $R = 1$ soit plus grande que sa valeur pour une situation $R = 0$ (Hanley et McNeil, 1982). L'indicateur le plus précis sera celui avec l'aire sous la courbe ROC la plus grande.

Deux exemples sont présentés ci-dessous.

4.2.4 Exemple : gestion du risque d'invasion par les mauvaises herbes

L'utilisation systématique d'herbicide pour lutter contre les mauvaises herbes en parcelles agricoles peut conduire à un risque accru de pollution du sol et de l'eau et à des coûts de traitement importants pour les agriculteurs (van Der Werf, 1996). Une alternative au traitement systématique consiste à ne traiter

que les situations qui présentent des risques élevés d'invasion par les mauvaises herbes. La mise en œuvre de cette approche nécessite la définition de règles de décision permettant de décider, précocement dans le cycle de la culture, de l'opportunité d'un traitement. Dans l'exemple ci-dessous, une règle de décision basée sur une mesure de densité de mauvaises herbes est mise au point pour le colza. Cette règle est inspirée des travaux réalisés par Primot *et al.* (2006).

Étape *i.* Définition de la règle de décision

Une règle de décision binaire basée sur l'équation (4.1) est définie en utilisant comme indicateur I la densité de mauvaise herbe mesurée sur un site-année de colza, quatre à six semaines après le semis (à l'automne). Si cette densité est supérieure à un seuil de décision I_s, le risque d'invasion est considéré comme étant élevé et un traitement est recommandé. Cette règle néglige un certain nombre de facteurs pouvant avoir un effet sur la population de mauvaises herbes mais elle présente l'avantage d'être simple et d'être basée sur une mesure précoce compatible avec un traitement herbicide post-levée.

Étape *ii.* Définition de l'objectif

Notre objectif est de déterminer une valeur du seuil de décision I_s conduisant à des valeurs de sensibilité et spécificité égales à au moins 0.7. Notez qu'il ne s'agit que d'un exemple utilisé ici pour illustrer la démarche et que d'autres objectifs pourraient être considérés en fonction des besoins de l'utilisateur. Les valeurs de sensibilité et spécificité sont définies par rapport à une variable de référence R correspondant à un seuil de nuisibilité de biomasse d'adventice. Ici, $R = 1$ si la biomasse de mauvaise herbe en hiver dépasse le seuil de nuisibilité de 0.15 t ha^{-1} et $R = 0$ sinon (Primot *et al.*, 2006). La sensibilité correspond donc à la probabilité que la densité de mauvaise herbe à l'automne ait été supérieure à I_s sachant que la biomasse de mauvaise herbe en hiver est supérieure à 0.15 t ha^{-1}. La spécificité correspond à la probabilité que la densité de mauvaise herbe à l'automne ait été inférieure à I_s sachant que la biomasse de mauvaise herbe en hiver est inférieure à 0.15 t ha^{-1}.

Étape *iii.* Calcul de la valeur optimale du paramètre de la règle de décision

Données Un nombre $N = 268$ de situations expérimentales (sites-années) est utilisé pour estimer les valeurs de sensibilité et de spécificité. Les situations expérimentales correspondent à des parcelles de colza cultivées dans plusieurs régions françaises en 1998, 1999 et 2002. La densité de mauvaises herbes à l'automne et la biomasse de mauvaises herbes en hiver ont été mesurées dans chacune des 268 situations (Primot *et al.*, 2006).

Le programme R ci-dessous permet de présenter la distribution des valeurs de densité sous forme d'histogramme pour les situations à faibles et fortes biomasses de mauvaises herbes.

```
# Code R: Presentation des donnees

TAB <- read.table("f:\\Exemple2\\Exemple2.txt",header=T,sep="\t")

Biom <- TAB$MSMHeh
Biom.t <- 0.15
Biom[Biom < Biom.t] <- 0
Biom[Biom >= Biom.t] <- 1

Ind <- TAB$DMHAUT

par(mfrow=c(2,2))

hist(Ind[Biom==0], main=" ",
                   xlab="Densite automne (plantes/m^2)$",
                   ylab="Frequence")
text(200,60, "moy.=38.3 plantes/m^2$", cex=0.8)

hist(Ind[Biom==1], main=" ",
                   xlab="Densite automne (plantes/m^2)$",
                   ylab="Frequence")
text(350,25, "moy.=98.2 plantes/m^2$", cex=0.8)
```

Une variable `Biom` est créée à partir du tableau de données `TAB`. Cette variable continue est ensuite transformée en variable binaire en utilisant le seuil de nuisibilité de 0.15 t ha^{-1}. Finalement, une variable `Ind`, contenant les mesures de densité, est définie à partir du tableau de données. Cette variable est utilisée pour présenter la distribution des densités sous forme d'histogramme pour les situations à faibles et fortes biomasses en hiver. Les résultats montrent que la densité de mauvaises herbes est, en moyenne, plus élevée dans les situations à fortes biomasses que dans les situations à faibles biomasses (figure 4.1ab).

Analyse ROC Les données sont utilisées pour estimer la sensibilité et la spécificité pour toutes les valeurs de I_s possibles. Les valeurs du seuil de décision permettant d'atteindre des valeurs de sensibilité et de spécificité au moins égales à 0.7 sont ensuite identifiées.

L'analyse ROC est réalisée ici avec la librairie `ROCR` (Sing *et al.*, 2005). Deux fonctions R sont utilisées : prediction et performance. La fonction prediction transforme les données (les valeurs de l'indicateur et celles de la variable de référence) en un format standard qui est ensuite utilisé par performance pour évaluer la précision de l'indicateur. La fonction `performance` permet de calculer de nombreux critères, mais ici seul le calcul de la sensibilité et de la spécificité est réalisé. Les valeurs de sensibilité et de spécificité sont déterminées automatiquement pour tous les seuils de décision possibles et récupérées dans `perf`. L'instruction `perf@` permet d'extraire les résultats. Ceux-ci sont stockés dans

les vecteurs nommés ici `spec.1` et `sens.1`. La première instruction `plot` du programme ci-dessous trace la courbe ROC et la seconde présente les valeurs de sensibilité et de spécificité en fonction des seuils de décision.

```
# Code R: Utilisation de ROCR

library(ROCR)

pred <- prediction(Ind, Biom)
perf <- performance(pred, "sens", "spec")

spec.1 <- perf@"x.values"[[1]]
sens.1 <- perf@"y.values"[[1]]

plot(spec.1, sens.1, xlab="Specificite",
                  ylab="Sensibilite", type="l", lty=1, lwd=3)
abline(1,-1)

plot(perf@"alpha.values"[[1]], spec.1,
                   xlab="Densite automne (plantes/m$^{2})$",
                   ylab="Sensibilite/Specificite",
                   lty=1, lwd=2, type="l")
lines(perf@"alpha.values"[[1]], sens.1,
                   xlab="Densite automne (plantes/m^2)$",
                   lty=4, lwd=2)
```

La figure 4.1cd présente les résultats de l'analyse ROC. La courbe ROC (figure 4.1c) passe largement au-dessus de la bissectrice ce qui indique que l'indicateur « densité de mauvaises herbes » discrimine les situations à forte biomasse des situations à faible biomasse. Une courbe ROC proche de ou sur la bissectrice aurait révélé que l'indicateur ne permet pas de classer les situations de façon plus précise qu'un classement aléatoire. Notez cependant que, bien que notre courbe ROC soit au-dessus de la bissectrice, cette courbe reste relativement éloignée du point idéal (1, 1).

La figure 4.1d présente les valeurs de sensibilité et de spécificité en fonction du seuil de décision. Plus le seuil est élevé, moins la règle de décision est sensible mais plus elle est spécifique. Lorsqu'un seuil de décision élevé est utilisé, l'application de la règle de décision conduit à ne classer dans la catégorie « forte biomasse » (*i.e.*, biomasse de mauvaises herbes supérieure à 0.15 t ha^{-1}) qu'un petit nombre de situations expérimentales ce qui réduit la sensibilité mais augmente la spécificité.

La figure 4.1d permet d'identifier les seuils de décision qui conduisent à des valeurs élevées de sensibilité et de spécificité. Ici, un seuil de 43 plantes/m^2 permet d'obtenir une sensibilité et une spécificité supérieures à 0.7 (0.73 pour la spécificité et 0.74 pour la sensibilité).

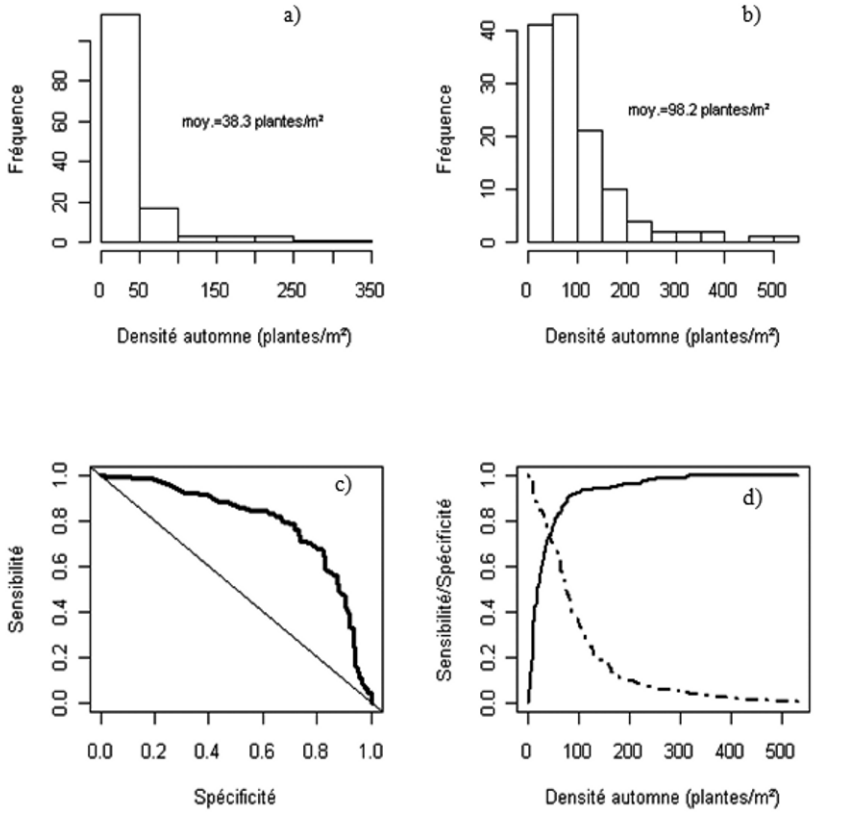

Figure 4.1 – Distributions des mesures de densité de mauvaises herbes dans 268 sites-années de colza lorsque la biomasse des mauvaises herbes en hiver est faible (a) ou forte (b). Courbe ROC pour l'indicateur « densité » (c) et valeurs de sensibilité (tirés) et de spécificité (continue) en fonction du seuil de décision (d).

Étape *iv.* Évaluation

Plusieurs critères peuvent être utilisés pour évaluer la qualité de la règle de décision. Une première possibilité est de calculer l'aire sous la courbe ROC présentée sur la figure 4.1c. La librairie ROCR permet d'estimer cette aire (ASC) avec une méthode non paramétrique. Il suffit pour cela d'utiliser l'instruction `performance` en précisant que l'on souhaite calculer l'élément « auc » comme indiqué ci-dessous. Il est également possible d'estimer cette aire sous la courbe sans la librairie ROCR en calculant la statistique du test de Wilcoxon directement avec des instructions R de base (*e.g.*, Pepe, 2003).

```
# Code R : Calcul de l'aire sous la courbe ROC

perf <- performance(pred,"auc")
auc.1 <- perf@"y.values"

print("ASC de l'indicateur")
print(auc.1)
```

La valeur estimée de l'aire sous la courbe ROC de la figure 4.1c est ASC = 0.79. Cette valeur est proche de la meilleure valeur obtenue par Primot *et al.* (2006) (ASC = 0.81) pour des règles de décision plus complexes basées sur des combinaisons d'indicateurs. On peut donc considérer que la performance de l'indicateur « densité de mauvaises herbes à l'automne » est satisfaisante.

L'ASC fournit une mesure globale de la performance de l'indicateur considéré indépendamment du seuil de décision utilisé. Il peut cependant être intéressant d'évaluer la performance de la règle de décision elle-même, en considérant la valeur retenue pour le seuil de décision. Yuen et Hughes (2002) ont proposé de calculer un critère de type rapport de vraisemblance défini par « Sensibilité / (1 − Spécificité) » qui représente la probabilité que $I \geq I_s$ dans une situation à forte biomasse hiver divisée par la probabilité que $I \geq I_s$ dans une situation à faible biomasse hiver. Ce critère dépend du seuil de décision. Les résultats de l'étape *iii* nous ont permis de fixer ce seuil à $I_s = 43$ plantes/m^2. Pour ce seuil, Sensibilité = 0.74 et Spécificité = 0.73, ce qui conduit à un rapport « Sensibilité / (1 − Spécificité) » égal à 2.74. Il y a donc 2.74 fois plus de chance que $I \geq I_s$ (et donc qu'un traitement soit recommandé) dans une situation à forte biomasse hiver que dans une situation à faible biomasse hiver.

Une autre façon d'utiliser la règle de décision définie ci-dessus est de se placer dans un contexte bayésien et de calculer la probabilité que la biomasse soit forte en hiver (supérieure à 0.15 t ha^{-1}) conditionnellement à $I \geq I_s$. Cela revient à calculer la probabilité *a posteriori* que $R = 1$, c'est-à-dire $P(R = 1 \,|I \geq I_s)$. Cette probabilité peut être reliée à la sensibilité et à la spécificité en utilisant le théorème de Bayes (Yuen et Hughes, 2002) :

$$P(R = 1 \,|I \geq I_s) = \frac{\text{Sensibilité} \times P(R = 1)}{\text{Sensibilité} \times P(R = 1) + (1 - \text{Spécificité})[1 - P(R = 1)]} \tag{4.1}$$

Cette probabilité dépend de la probabilité *a priori* que la biomasse soit élevée en hiver $P(R=1)$ qui peut être estimée par la proportion des situations où la biomasse mesurée en hiver est supérieure à 0.15 t ha^{-1}, c'est-à-dire ici $P(R=1) = 0.47$. Pour le seuil de décision retenu, on obtient alors $P(R=1 \,|\, I \geq I_s) = 0.71$.

4.2.5 Exemple : gestion du risque de sclérotinia du colza

Dans cet exemple, nous considérons de nouveau le champignon *Sclerotinia sclerotiorum* Lib. de Bary qui attaque le colza. L'incidence (pourcentage de plantes affectées) de la maladie induite par le champignon étant très variable, des règles de décision ont été proposées par le Cetiom (Centre technique des oléagineux) pour identifier les situations à risque, décider de l'opportunité d'un traitement fongicide et éviter des applications de fongicides inutiles préjudiciables pour l'environnement (Maisonneuve *et al.*, 1997 ; Taverne, 2003). Contrairement à l'exemple précédent, plusieurs indicateurs I différents sont disponibles ici pour décider de l'opportunité d'un traitement et nous montrons ci-dessous comment évaluer l'intérêt de combiner les indicateurs disponibles.

Étape *i.* Définition des règles de décision

Trois règles de décision binaires basées sur l'équation (4.1) sont définies. La première est basée sur l'indicateur I_1 « fraction de fleurs malades ». Cette variable est mesurée début floraison. Si sa valeur dépasse un seuil de décision, un traitement est recommandé. La deuxième règle est basée sur l'indicateur I_2 qui correspond à un score reflétant un niveau de risque. Un traitement est recommandé si le score dépasse un seuil de décision. Le score est calculé pour chaque situation à partir d'une série de facteurs définis dans une grille de risque (tableau 4.2). Ces facteurs sont liés au climat et aux techniques culturales appliquées par l'agriculteur (Maisonneuve *et al.*, 1997). Finalement, la troisième règle est basée sur une combinaison de la mesure de fraction de fleurs malades et du score. Ces deux indicateurs sont combinés à l'aide d'une régression logistique (voir la section 3.2 et l'étape iii) et déterminent un troisième indicateur défini par

$$I_3 = \frac{\exp(\alpha_0 + \alpha_1 I_1 + \alpha_2 I_2)}{1 + \exp(\alpha_0 + \alpha_1 I_1 + \alpha_2 I_2)}, \tag{4.2}$$

où I_1 est la fraction de fleurs malades, I_2 est le score de la grille de risque et α_0, α_1, α_2 sont trois paramètres à estimer. Avec la troisième règle de décision, un traitement est recommandé si I_3 est supérieur à un seuil de décision.

Étape *ii.* Définition des objectifs

Notre objectif est d'abord d'identifier l'indicateur qui, parmi les trois définis ci-dessus, permet d'obtenir les valeurs de sensibilité et de spécificité les plus grandes. Ces indicateurs utilisent des informations différentes. Le premier

Facteur de risque	Niveau	Points
Nombre de cultures en colza	> 5	30
au cours des 10 années	3-5	20
précédentes	2-3	10
	1	0
Autre cultures hôtes au cours	oui	15
des 5 dernières années	non	0
Niveau d'infection de la	élevé	15
culture précédente	moyenne	5
	faible	0
Type de parcelle	humide	10
	sèche	0
Densité de plantes	élevé	10
	normale	5
	faible	0
Pluie au cours du mois	élevée	10
précédant la floraison	normale	5
(normale = 50-60 mm)	faible	0
Nombre de jours pluvieux	10-14	30
(1 à 10 mm) au cours des	9-5	20
deux semaines avant floraison	< 5	10
	0	0
Température moyenne au cours	$> 15\,^{\circ}\mathrm{C}$	15
des 5 jours avant floraison	10-15	10
	< 10	0
Prévision météo	pluie	30
	variable	15
	sec	0
Sol	très humide	15
	humide	10
	sec	0

Tableau 4.2 – Facteurs de risque du sclérotinia du colza et nombres de points associés dans le calcul de l'indicateur de risque I_2.

est basé sur une mesure d'une caractéristique du couvert végétal, le deuxième indicateur tient compte du climat et des techniques culturales, le troisième combine l'ensemble des informations. Le deuxième objectif est de déterminer une valeur du seuil de décision I_s conduisant à des valeurs de sensibilité et spécificité égales à au moins 0.8.

Etape *iii.* Calcul de la valeur optimale du paramètre de chaque règle

Données Quatre-vingt-cinq sites-années de colza non traités contre le sclérotinia sont utilisés dans cette application. Ils correspondent à deux années d'expérimentations réalisées par le Cetiom en 2002 et 2003 dans le nord-est, le centre et l'ouest de la France. Quatre-vingt fleurs ont été prélevées dans chaque site-année et ont été incubées pendant 4 jours dans des boîtes de Petri à 22-23 °C . La présence ou l'absence de sclérotinia a ensuite été notée pour chaque fleur. L'histoire culturale et le climat ont été déterminés pour chacun des 85 sites-années.

L'incidence de la maladie (% plantes malades) a été mesurée sur chaque site-année environ 3 semaines avant la récolte à partir d'échantillons de 200 plantes. Voir Makowski *et al.* (2005) pour une description plus détaillée des expérimentations.

Dans cette étude, la variable de référence R est définie à partir de la mesure d'incidence de sclérotinia : $R = 1$ si l'incidence est supérieure à 10%, $R = 0$ sinon. Dans notre base de données, $R = 1$ pour 20% des sites-années.

Estimation des paramètres du modèle logistique pour l'indicateur I_3 Les paramètres α_0, α_1, α_2 de l'équation (4.2) définissant I_3 ont été estimés à partir des 85 sites-années décrits ci-dessus en posant le modèle logistique :

$$R \sim \text{Bern}(p)$$
$$\text{logit}(p) = \alpha_0 + \alpha_1 I_1 + \alpha_2 I_2.$$

L'estimation est faite par la méthode du maximum de vraisemblance en utilisant la fonction `glm` de R. En inversant la fonction logit avec les paramètres estimés, on obtient :

$$I_3 = \widehat{P}(R = 1 \mid I_1, I_2) = \frac{\exp(\widehat{\alpha}_0 + \widehat{\alpha}_1 I_1 + \widehat{\alpha}_2 I_2)}{1 + \exp(\widehat{\alpha}_0 + \widehat{\alpha}_1 I_1 + \widehat{\alpha}_2 I_2)}.$$

Le code R pour l'estimation des paramètres est le suivant :

```
# Code R: Estimation des parametres du modele logistique

TAB <- read.table("f:\\Exemples\\Sclero0203.txt",
                                          header=T, sep="\t")
TAB <- TAB[is.na(TAB[,1]) == F,]
```

```
Ind.1 <- TAB$KIT
Ind.2 <-
   TAB[,3]+TAB[,4]+TAB[,5]+TAB[,6]+TAB[,7]+TAB[,8]+TAB[,9]
   +TAB[,10]+TAB[,11]+TAB[,12]

Incidence <- TAB$TxAttNT
Incidence.t <- 0.10
Incidence[Incidence<Incidence.t]  <- 0
Incidence[Incidence>=Incidence.t] <- 1

Fit <- glm( Incidence ~ Ind.1 + Ind.2, family=binomial)
```

Les variables `Ind.1` et `Ind.2` sont définies à partir du tableau de données `TAB`. La variable `Ind.2` est est égale à la somme des points calculée comme indiqué dans le tableau 4.2. Une variable binaire `Incidence` est ensuite définie. Elle est égale à 1 si l'incidence en fin de saison est supérieure à 10% et à 0 sinon. Finalement les paramètres sont estimés en reliant `Incidence` à `Ind.1` et `Ind.2` avec `glm`.

Les résultats de l'estimation sont $\widehat{\alpha}_0 = -2.53$, $\widehat{\alpha}_1 = 2.65$, $\widehat{\alpha}_2 = 0.02$. Les écarts types des estimateurs sont égaux à 0.83, 0.88 et 0.01 respectivement. $\widehat{\alpha}_0$ et $\widehat{\alpha}_1$ sont significativement différents de zéro au seuil 0.001 (probabilité d'erreur de type I) mais $\widehat{\alpha}_2$ n'est pas significativement différent de zéro au seuil 0.1.

Analyse ROC L'analyse ROC est réalisée comme dans l'exemple précédent, avec la librairie ROCR. Le programme utilisé pour appliquer l'analyse ROC au premier indicateur est présenté ci-dessous. Les valeurs de sensibilité et de spécificité sont calculées avec les fonctions `performance` et `prediction`.

```
# Code R: Estimation des valeurs de sensibilite et de specificite
#         pour Ind.1.

library(ROCR)

pred <- prediction(Ind.1, Incidence)
perf <- performance(pred, "sens", "spec")

spec.1 <- perf@"x.values"[[1]]
sens.1 <- perf@"y.values"[[1]]

plot(spec.1, sens.1, ylab="Sensibilite",
                     xlab="Specificite",
                     type="l", lty=1, lwd=3)
abline(1,-1)
```

Un programme similaire est utilisé pour le deuxième indicateur basé sur la grille de risque. Par contre, le calcul des valeurs de sensibilité et de spécificité de l'indicateur I_3 basé sur le modèle logistique (4.2) nécessite un programme différent. En effet, la base de données disponible pour l'analyse ROC a déjà été utilisée une première fois pour estimer les trois paramètres du modèle logistique. L'utilisation des mêmes données pour calculer les valeurs de sensibilité et de spécificité peut conduire à une surestimation de ces valeurs et, par conséquent, à une vision trop optimiste des performances des règles de décision basées sur le troisième indicateur. Pour éviter ce problème, une procédure appelée *validation croisée* est mise en œuvre avec le modèle logistique. Le principe est d'utiliser $N-1$ données pour estimer les trois paramètres du modèle et d'utiliser la N-ieme pour prédire la probabilité que $R=1$. Cette technique est répétée N fois en changeant de donnée pour la prédiction. La validation croisée est mise en œuvre dans une boucle `for` comme indiquée ci-dessous.

```
# Code R: Validation croisee pour l'indicateur Ind.3

Pred.cv <- NA

for (i in (1:length(TAB[,1]))) {

  TAB.est.i <- data.frame(Ind.1[-i], Ind.2[-i], Incidence[-i])
  TAB.pred.i <- c(Ind.1[i], Ind.2[i])

  Fit.cv <- glm(
    TAB.est.i$Incidence~ TAB.est.i[,1] + TAB.est.i[,2],
    family=binomial, data=TAB.est.i)
  Para <- as.vector(Fit.cv$coefficients)

  Pred.i <-
   exp(Para[1] + Para[2]*TAB.pred.i[1] + Para[3]*TAB.pred.i[2])/
   (1 +
    exp(Para[1] + Para[2]*TAB.pred.i[1] + Para[3]*TAB.pred.i[2]))
  Pred.cv <- c(Pred.cv, Pred.i)

  }

pred <- prediction(Pred.cv[-1], Incidence)
perf <- performance(pred, "sens", "spec")
```

Les prédictions calculées à chaque itération sont archivées dans un objet appelé `Pred.cv`. A chaque itération, un nouveau fichier de données incluant 84 éléments est créé. Ce fichier est ensuite utilisé pour estimer les paramètres du modèle logistique. Les valeurs estimées des paramètres sont extraites et sont ensuite utilisées pour prédire la probabilité de forte incidence qui est stockée dans

`Pred.cv`. La boucle s'arrête lorsque chacune des 85 données disponibles a été extraite du fichier initial. Finalement, l'analyse ROC est réalisée sur `Pred.cv`.

Les résultats des analyses ROC sont présentés sur les figures 4.2 et 4.3. La figure 4.2 présente les courbes ROC pour les indicateurs `Ind.1` et `Ind.2`, c'est-à-dire les indicateurs I_1 « Fraction de fleurs malades » et I_2 « Grille de risque ». La courbe obtenue pour `Ind.1` est proche du point idéal (1, 1). L'indicateur « fraction de fleurs malades » peut donc conduire à des valeurs élevées de sensibilité et de spécificité. La courbe ROC obtenue pour `Ind.2` est par contre proche de la bissectrice ce qui révèle que l'indicateur « Grille de risque » est à peine meilleur qu'une décision aléatoire.

La figure 4.3 présente la courbe ROC de l'indicateur I_3 basé sur le modèle logistique obtenue avec et sans validation croisée. La courbe obtenue avec validation croisée est légèrement en dessous de celle obtenue sans validation croisée. Les valeurs de sensibilité et de spécificité obtenues avec validation croisée sont donc un peu plus faibles. Celles obtenues sans validation croisée sont probablement légèrement surestimées.

Visuellement, les figures 4.2 et 4.3 ne semblent pas montrer que l'indicateur I_3 soit plus performant. Pour le confirmer, il est cependant nécessaire d'utiliser des critères quantitatifs comme, par exemple, l'aire sous la courbe. Celle-ci est calculée dans la section suivante.

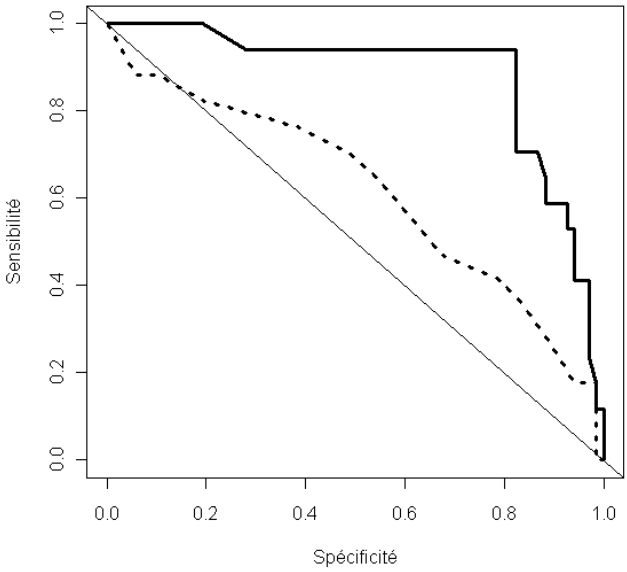

Figure 4.2 – Courbes ROC pour les indicateurs « Fraction de fleurs malades » (continue) et « Grille de risque » (pointillés).

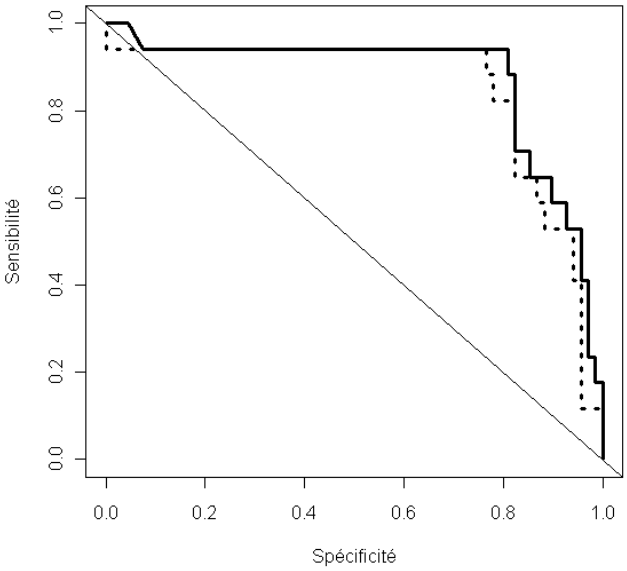

Figure 4.3 – Courbes ROC pour le modèle logistique avec (pointillés) et sans (continue) validation croisée.

Étape 4. Évaluation

Les indicateurs sont comparés en calculant les valeurs d'aire sous la courbe ROC (ASC). Celles-ci sont déterminées avec une technique non paramétrique en utilisant la librairie ROCR. Voir l'exemple précédent pour plus de détail sur les programmes. Les résultats sont présentés dans le tableau 4.3. La valeur la plus élevée est obtenue pour l'indicateur I_1, c'est-à-dire la fraction de fleurs malades. La valeur d'ASC obtenue pour la grille de risque (indicateur I_2) est nettement plus faible : 0.62. L'utilisation du modèle logistique (indicateur I_3) ne conduit pas à une valeur d'ASC supérieure à celle obtenue avec I_1. Il n'y a donc pas d'intérêt à combiner les indicateurs « Fraction de fleurs malades » et « Grille de risque ». Il est préférable de n'utiliser que la fraction de fleurs malades.

La procédure déjà utilisée dans le premier exemple a été appliquée ici pour identifier un seuil de décision conduisant à des valeurs de sensibilité et de spécificité supérieures à 0.80. Les résultats montrent que le seuil de décision de 0.36 permet d'obtenir une sensibilité et une spécificité de 0.82. Avec ce seuil, un traitement est recommandé lorsque le pourcentage de fleurs malades est supérieur à 36%. La probabilité *a posteriori* $P(R = 1 \,|I \geq 0.36)$ définie par

Indicateurs	ASC
Fraction de fleurs malades	0.88
Grille de risque	0.62
Logistique (sans validation croisée)	0.87
Logistique (avec validation croisée)	0.85

Tableau 4.3 – Valeurs d'aires sous la courbe ROC (ASC) pour les indicateurs « Fraction de fleurs malades », « Grille de risque » et pour la régression logistique.

l'expression (4.1) est égale à 0.53 et est donc nettement supérieure à la proportion des situations expérimentales ayant une forte incidence à la récolte, égale à 0.20.

4.3 Optimisation d'une variable décisionnelle par simulation

4.3.1 Méthode

Nous considérons maintenant une variable décisionnelle quantitative D (par exemple, une dose d'engrais ou une décision de traiter contre une maladie) et nous cherchons à déterminer la valeur d de D qui maximise une fonction objectif notée $J(d, \omega)$, où ω est un indicateur ou un vecteur d'indicateurs décrivant les caractéristiques du site-année (caractéristiques du sol, par exemple) en fonction desquelles la variable décisionnelle doit être optimisée. La fonction J pourra représenter, par exemple, la marge économique d'un agriculteur.

Lorsque des mesures de $J(d, \omega)$ sont disponibles pour toutes les valeurs d possibles, il est facile d'identifier celle qui maximise la fonction objectif. Bien souvent, aucune valeur mesurée de $J(d, \omega)$ n'est cependant disponible sur le site-année d'intérêt, ou bien seulement pour quelques valeurs d particulières. Une solution est alors d'utiliser un modèle pour prédire les valeurs de $J(d, \omega)$ en fonction de d et ω. L'optimisation est ensuite réalisée à partir des valeurs prédites par le modèle $\widehat{J}(d, \omega)$. D'un point de vue formel, ce problème revient à déterminer la valeur définie par :

$$d^* = \arg\max_d \widehat{J}(d, \omega).$$

La valeur d^* est déterminée en fonction de ω. En répétant l'opération pour différentes valeurs de ω, on obtient une fonction qui relie d^* à ω. Cette fonction correspond à une règle de décision permettant d'adapter la variable décisionnelle aux caractéristiques du milieu.

Si le modèle n'est pas trop imprécis, la valeur d^* maximisant $\widehat{J}(d, \omega)$ correspondra à une bonne approximation de la vraie valeur optimale. Par contre, si les erreurs de prédiction du modèle sont trop importantes, l'approximation

pourra être mauvaise. La qualité de la valeur optimale calculée peut être évaluée en estimant la *perte* qui résulte de l'application de d^* plutôt que de la vraie décision optimale. L'utilisation de d^* plutôt que de la vraie valeur optimale conduit à une perte égale à $\max_d [J(d,\omega)] - J(d^*,\omega)$.

Différents algorithmes peuvent être utilisés pour calculer d^*. Dans certains cas, il est possible de déterminer l'expression analytique de la fonction reliant d^* à ω. Dans d'autres cas, ce n'est pas possible. Des algorithmes numériques doivent alors être utilisés. Une approche simple consiste à procéder par *simulations*, c'est-à-dire à calculer les valeurs de $\widehat{J}(d,\omega)$ pour un nombre suffisamment grand de valeurs d différentes, puis à déterminer la valeur qui maximise $\widehat{J}(d,\omega)$. Cette approche est illustrée dans l'exemple ci-dessous, où le modèle $J(d,\omega)$ n'est pas déterministe mais contient des paramètres aléatoires pour refléter la variabilité des conditions de milieu à prendre en compte.

4.3.2 Exemple : calcul de doses optimales d'engrais

On cherche dans cet exemple à optimiser la dose d'engrais azoté minéral appliquée sur blé d'hiver. Nous utilisons ici un modèle non linéaire mixte simulant le rendement de la culture en fonction de la dose d'engrais appliquée et d'une caractéristique du sol.

Étape *i.* Définition de la règle de décision

Une partie des besoins en azote d'une culture de blé est satisfaite par l'azote minéral disponible dans le sol. Il est important de tenir compte de cette quantité d'azote afin d'appliquer suffisamment d'engrais pour satisfaire les besoins de la culture mais d'éviter un apport excessif dommageable pour l'environnement et coûteux sur le plan économique. Nous cherchons ici à définir une règle de décision qui permet de calculer une dose d'engrais en fonction d'une mesure d'azote minérale du sol ω réalisée avant la date d'apport. L'utilisation de cette règle permet d'ajuster les apports d'engrais azoté à la quantité d'azote déjà présente dans le sol sous forme minéral et, ainsi, de limiter les risques de pollution de l'eau par les nitrates.

Étape *ii.* Définition de l'objectif

La fonction objectif considérée ici est la marge brute économique de l'agriculteur calculée pour une parcelle de blé et définie par

$$J(d,\omega) = r\, E\,Y(d,\omega) - cd \tag{4.3}$$

avec $E\,Y(d,\omega)$ l'espérance du rendement connaissant la mesure d'azote minéral ω et la dose d'engrais d, r le prix d'une unité de rendement (euros par tonne de grains) et c le coût d'une unité d'engrais (euros par kg d'azote). La dose d'engrais optimale d^* est la dose qui maximise (4.3). Notez que la variable rendement Y est considérée comme aléatoire. La marge brute est donc également aléatoire et on choisit de s'intéresser à son espérance.

Etape *iii.* Optimisation

Notons $\widehat{Y}(d,\omega)$ la valeur prédite de $EY(d,\omega)$. La valeur correspondante de la fonction objectif est définie par $\widehat{J}(d,\omega) = r\,\widehat{Y}(d,\omega) - cd$. En calculant $\widehat{J}(d,\omega)$ pour différentes doses, il est possible de calculer la dose d'engrais optimale associée à ω.

Le modèle utilisé pour générer les valeurs de $\widehat{Y}(d,\omega)$ est un modèle linéaire-plus-plateau à effets mixtes (voir section 3.4) reliant le rendement à la dose d'engrais azotée et à une mesure d'azote minéral du sol réalisée sortie hiver (avant la date d'apport d'engrais) (Makowski et Lavielle, 2006). La fonction de réponse de ce modèle, notée $f(d,\boldsymbol{\alpha})$, est définie par

$$\begin{cases} Y_{\mathrm{MAX}} + B(d - X_{\mathrm{MAX}}) & \text{si } d < X_{\mathrm{MAX}}, \\ Y_{\mathrm{MAX}} & \text{si } d \geq X_{\mathrm{MAX}}. \end{cases}$$

Cette fonction inclut un vecteur de trois paramètres, $\boldsymbol{\alpha} = (Y_{\mathrm{MAX}}, X_{\mathrm{MAX}}, B)^T$. Le paramètre Y_{MAX} est le rendement maximal dans le site-année, X_{MAX} est la dose d'engrais qui maximise le rendement et B est l'augmentation de rendement par unité de dose d'azote supplémentaire.

Le vecteur $\boldsymbol{\alpha}$ est défini comme un vecteur aléatoire :

$$\boldsymbol{\alpha} = A_\omega \boldsymbol{\mu} + \boldsymbol{\eta}$$

avec $\boldsymbol{\eta} \sim N(0,\Gamma)$, A_ω est une matrice incluant les caractéristiques du site-année, $\boldsymbol{\mu}$ est un vecteur incluant des paramètres fixes, $\boldsymbol{\eta}$ est un vecteur (3×1) d'effets aléatoires, Γ est une matrice 3×3 de variance-covariance supposée diagonale, avec les éléments diagonaux (les variances des effets aléatoires) notés $\sigma^2_{Y_{\mathrm{MAX}}}$, $\sigma^2_{X_{\mathrm{MAX}}}$ et σ^2_B. La distribution de probabilité de $\boldsymbol{\alpha}$ décrit la variabilité inter sites-années des paramètres de la fonction linéaire-plus-plateau.

La matrice A_ω est définie par

$$A_\omega = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \omega & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

avec ω la mesure d'azote minéral. Le vecteur $\boldsymbol{\mu}$ inclut quatre éléments, $\boldsymbol{\mu} = (\mu_{Y_{\mathrm{MAX}}}, \mu_{X_{\mathrm{MAX}}}, \beta, \mu_B)^T$. Avec ces définitions, l'espérance conditionnelle de X_{MAX} dépend linéairement de ω :

$$E(X_{\mathrm{MAX}} \mid \omega) = \mu_{X_{\mathrm{MAX}}} + \beta\,\omega.$$

Pour utiliser ce modèle, il est nécessaire de connaître les valeurs de $\boldsymbol{\mu} = (\mu_{Y_{\mathrm{MAX}}}, \mu_{X_{\mathrm{MAX}}}, \beta, \mu_B)^T$ et de $\sigma^2_{Y_{\mathrm{MAX}}}$, $\sigma^2_{X_{\mathrm{MAX}}}$ et σ^2_B. Dans la section 3.4, plusieurs méthodes ont été proposées pour estimer les paramètres de ce type de modèle : maximum de vraisemblance ou méthode bayésienne. Le tableau 4.4 présente les valeurs estimées par maximum de vraisemblance en utilisant des

Paramètre	Unité de mesure	Valeur estimée	Écart type
$\mu_{Y_{\mathrm{MAX}}}$	t ha^{-1}	9.18	(0.19)
$\sigma_{Y_{\mathrm{MAX}}}$	t ha^{-1}	1.15	(0.29)
μ_B	t kg^{-1}	0.026	(0.0012)
σ_B	t kg^{-1}	0.0056	(0.0017)
$\mu_{X_{\mathrm{MAX}}}$	kg ha^{-1}	217.11	(16.03)
$\sigma_{X_{\mathrm{MAX}}}$	kg ha^{-1}	31.31	(8.74)
β	t kg^{-1}	-1.11	(0.18)

Tableau 4.4 – Valeurs estimées des paramètres du modèle linéaire-plus-plateau et écarts types des estimateurs (entre parenthèses).

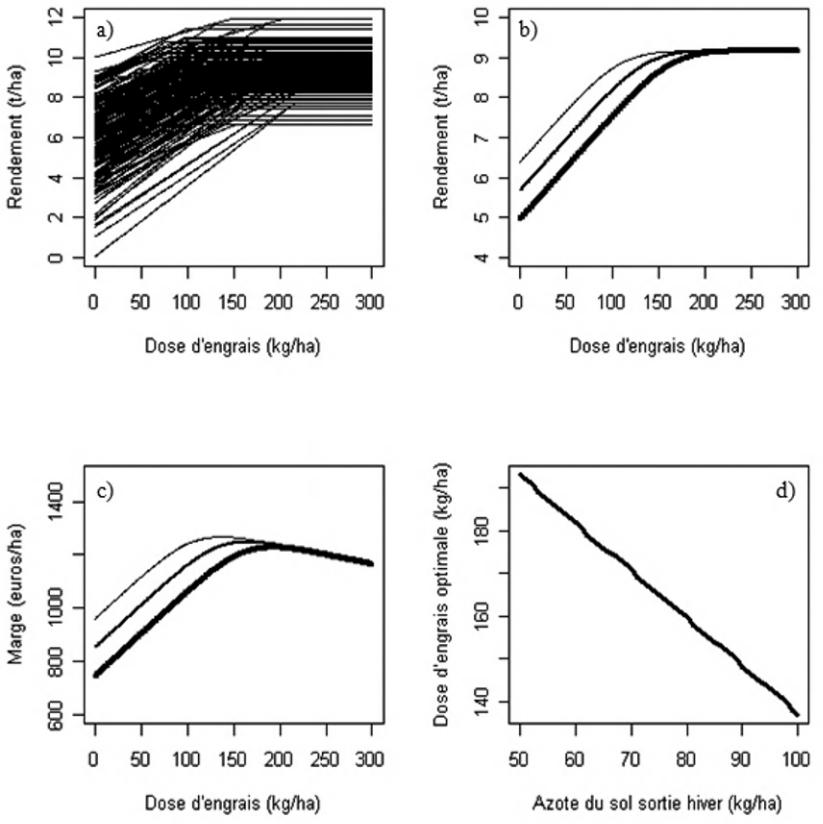

Figure 4.4 – Courbes de réponses du rendement à la dose d'engrais générées pour $\omega = 50$ kg ha^{-1} (a), courbes de réponses moyennes obtenues pour $\omega = 50$ (trait gras), 75 (trait moyen) et 100 kg ha^{-1} (trait fin) (b), réponses correspondantes de la marge économique (c) et doses optimales d'engrais en fonction de ω (d).

données expérimentales issues d'essais de fertilisation. Voir Makowski et Lavielle (2006) pour plus de détails sur cette méthode et sa mise en œuvre avec le modèle ci-dessus.

La valeur simulée de rendement $\widehat{Y}(d,\omega)$ est déterminée en générant un grand nombre de valeurs de $\boldsymbol{\eta}$ à partir de la distribution $\boldsymbol{\eta} \sim N(0,\Gamma)$, en calculant les valeurs correspondantes de $\boldsymbol{\alpha} = (Y_{\mathrm{MAX}}, X_{\mathrm{MAX}}, B)^T$, puis en moyennant les réponses $f(d,\boldsymbol{\alpha})$. Plus précisément, $\widehat{Y}(d,\omega)$ est définie par

$$\widehat{Y}(d,\omega) = \frac{1}{K}\sum_{k=1}^{K} f\left(d,\boldsymbol{\alpha}_k\right)$$

avec $\boldsymbol{\alpha}_k = A_\omega \boldsymbol{\mu} + \boldsymbol{\eta}_k$, où $\boldsymbol{\eta}_k$ est tirée aléatoirement dans $\boldsymbol{\eta} \sim N(0,\Gamma)$, pour $k = 1,\ldots,K$. Le nombre K représente le nombre total de vecteurs de paramètres générés. Il est fixé ici à 1 000. La valeur de la fonction objectif (4.3) est estimée par $\widehat{J}(d,\omega) = r\hat{Y}(d,\omega) - cd$ pour des doses d'engrais d comprises entre 0 et 300 kg ha^{-1}. Un pas de 1 kg ha^{-1} a été considéré pour générer les doses d'engrais. Finalement, la dose maximisant $\widehat{J}(d,\omega)$ est identifiée. Le prix du blé a été fixé à 150 euros t^{-1} et celui de l'engrais à 0.7 euro kg^{-1}. Les calculs ont été réalisés pour plusieurs valeurs de ω et les résultats sont présentés sur la figure 4.4.

Le programme R utilisé pour générer la figure 4.4a est présenté ci-dessous. La partie la plus intéressante de ce programme concerne la génération aléatoire des valeurs des paramètres. Celle-ci est réalisée avec l'instruction `rnorm` utilisée ici pour générer `Num = 1000` séries de valeurs de paramètres. Les 1 000 courbes de réponse correspondantes sont déterminées à l'aide d'une boucle `for` puis présentées une à une sur le graphique avec l'instruction `lines`.

```
# Code R

par(mfrow=c(2,2))
set.seed(1)

Num <- 1000
Dose <- 0:300
Rsh <- 50

#Matrice des Num*301 valeurs simulees de rendement

Rdt <- matrix(nrow=Num, ncol=length(Dose))
Ymax <- rnorm(Num, 9.18,1.15)
B    <- rnorm(Num, 0.026, 0.0056)
Xmax <- rnorm(Num, 217.11, 31.31)
Alpha <- -1.11

#Creation d'un graphique
```

```
plot(Dose, Dose, xlab="Dose d'engrais (kg/ha)",
                 ylab="Rendement (t/ha)",
                 ylim=c(0,12), pch=" ")

# Calcul des Num reponses du rendement vs. la dose d'engrais et
# impression dans le graphique

for (i in 1:Num) {
  Rdt[i,][Dose < (Xmax[i] + Alpha*Rsh)]  <- Ymax[i] +
        B[i]*(Dose[Dose < (Xmax[i] + Alpha*Rsh)] -
        Xmax[i] - Alpha*Rsh)
  Rdt[i,][Dose >= (Xmax[i] + Alpha*Rsh)] <- Ymax[i]

  lines(Dose, Rdt[i,])
  }
```

Étape *iv.* Évaluation

Deux types d'évaluation peuvent être envisagés ici : une évaluation du modèle et une évaluation de la qualité des doses optimales. L'évaluation du modèle peut se faire en utilisant les méthodes présentées dans le Chapitre 3, c'est-à-dire en analysant la précision des estimateurs des paramètres, la distribution des résidus, la valeur de l'AIC, etc. L'évaluation de la qualité des doses optimales peut être réalisée en estimant la *perte économique* qui résulte de l'application des doses optimales calculées ci-dessus plutôt que des vraies doses optimales. Cette perte est estimée ici en calculant les marges économiques maximales associées aux 1000 courbes de réponse générées, puis en soustrayant à ces valeurs la marge obtenue en appliquant la dose optimale calculée. La valeur moyenne des différences des marges constitue une estimation de la perte économique moyenne.

La perte moyenne estimée à partir de nos simulations est de 33 euros ha^{-1} pour toutes les valeurs de ω considérées. Cela signifie qu'un agriculteur perdra en moyenne 33 euros ha^{-1} en appliquant les doses optimales calculées plutôt que les vraies doses optimales. Cela signifie également que la connaissance des vraies doses optimales ne conduirait pas à un gain de plus de 33 euros ha^{-1} par rapport à l'application des doses recommandées par le modèle.

Le programme R utilisé pour calculer la perte est présenté ci-dessous. Il utilise la matrice `Rdt` et le vecteur `Dose` générés à l'aide du programme précédent. Voir Makowski et Wallach (2001) pour une présentation plus détaillée de ce type de calcul.

```
# Code R

Rsh <- 50:100
```

```
DoseOpt <- Rsh
MB.max  <- Rsh
MB.opt  <- Rsh

#Boucle sur les valeurs de reliquat sortie hiver

for (i in 1:length(Rsh)) {

  Rdt <- matrix(nrow=Num, ncol=length(Dose))
  MB.vec <- 1:Num

  #Boucle sur les courbes de reponses

  for (j in 1:Num) {

    Rdt[j,][Dose < (Xmax[j] + Alpha*Rsh[i])] <- Ymax[j] +
        B[j] * (Dose[Dose < (Xmax[j] + Alpha*Rsh[i])] -
        Xmax[j] - Alpha*Rsh[i])
    Rdt[j,][Dose >= (Xmax[j] + Alpha*Rsh[i])] <- Ymax[j]
    MB.vec[j] <- max(prixBle * Rdt[j,] - prixN*Dose)
    }

  #Calcul de la dose optimale

  RdtMoy <- apply(Rdt, 2, mean)
  MargeMoy.i <- prixBle * RdtMoy - prixN * Dose
  DoseOpt[i] <- Dose[MargeMoy.i == max(MargeMoy.i)]
  MB.opt[i] <- max(MargeMoy.i)
  MB.max[i] <- mean(MB.vec)
  }

#Calcul de la perte moyenne

print(MB.max - MB.opt)
```

4.4 Exercices

Exercice 4.1 Un écologiste souhaite prédire la présence/absence d'une espèce d'oiseau vivant dans les prairies à l'aide d'un ou plusieurs indicateurs. Il hésite entre trois indicateurs : la hauteur d'herbe de la prairie, la présence/absence d'un point d'eau dans la prairie, la fréquence de coupe. L'écologiste vous demande de l'aider à choisir le meilleur indicateur ou la meilleure combinaison d'indicateurs.

– Définir la sensibilité et la spécificité de chaque indicateur.

- Comment estimer la sensibilité et la spécificité ? Quelles données expérimentales sont nécessaires ? Comment procéder une fois les données disponibles ?
- Proposer un critère pour choisir l'indicateur le plus précis.
- Définir des modèles permettant de prédire la présence/absence en combinant deux ou trois indicateurs.
- Décrire une méthode pour évaluer ces modèles et choisir la plus précise.

Exercice 4.2 Un agriculteur doit décider d'appliquer un traitement fongicide dans sa parcelle blé contre le piétin verse, un champignon pathogène. Il a prélevé des pieds de blé sur une placette de 2 m^2 et a noté que le pourcentage de tiges nécrosées était de 25%.

Supposons que la recommandation habituelle dans la région de l'agriculteur soit de ne traiter que si le pourcentage de tiges nécrosées est supérieur à 30%. Supposons également que la sensibilité de cette règle de décision par rapport à une perte de rendement soit égale à 0.9 et que sa spécificité soit de 0.75.

Exprimer la probabilité d'avoir une perte de rendement conditionnellement au résultat du test de l'agriculteur en fonction de :

- la sensibilité de la règle de décision ;
- la spécificité de la règle de décision ;
- la probabilité d'avoir une perte de rendement due à cette maladie dans la région considérée. On supposera que cette perte est égale à 0.2 pour l'application numérique.

Exercice 4.3 Reprendre l'exemple présenté dans la section 4.3.2 et modifier le programme R pour étudier la sensibilité des doses d'engrais optimales aux prix du blé et de l'engrais. On supposera que ces prix peuvent varier de ±30%.

Exercice 4.4 Reprendre l'exemple présenté dans la section 4.3.2 et modifier le programme R pour déterminer le nombre minimal de valeurs de paramètres qu'il est nécessaire de générer pour calculer des doses d'engrais optimales fiables. On supposera que la dose optimale doit être calculée avec une précision relative de 10%.

Exercice 4.5 Modifier l'expression de la fonction objectif (4.3) pour tenir compte d'un prix du blé proportionnel à la teneur en protéines des grains de blé. Quelles variables d'entrée et de sortie doit-on simuler pour optimiser cette fonction objectif ?

Chapitre 5

Analyse et communication de l'incertitude

5.1 Les différents types d'incertitude et leurs conséquences

Les méthodes présentées dans le chapitre 3 permettent d'estimer la probabilité d'occurrence d'un événement et son impact. Celles décrites dans le chapitre 4 permettent d'optimiser des variables décisionnelles et de gérer des risques. Dans ce chapitre associé à la communication du risque, nous allons insister sur la quantification, l'analyse et la représentation de l'incertitude. Le terme « incertitude » a de très nombreuses acceptions, mais précisons que nous limitons notre propos à l'incertitude que l'on peut raisonnablement quantifier par des distributions de probabilité en s'appuyant sur la modélisation et la statistique.

Il est utile en pratique de distinguer deux types d'incertitude : l'incertitude due à la variabilité du milieu et l'incertitude due à un manque de connaissance. L'incertitude due à un manque de connaissance peut être réduite en accumulant de nouvelles connaissances, par exemple en réalisant de nouvelles expérimentations. L'incertitude due à la variabilité du milieu est par contre intrinsèque au système étudié et ne peut pas disparaître même avec une connaissance parfaite du fonctionnement du système. En général, ces deux types d'incertitude coexistent quelle que soit la question étudiée. Leur importance relative est, cependant, plus ou moins grande.

En agronomie, les modèles hiérarchiques appliqués à des réseaux d'essais (chapitre 3) peuvent s'interpréter sous l'angle de ces deux types d'incertitude. Considérons par exemple l'impact de l'apport d'une dose d'engrais azoté particulière (disons 200 kg ha^{-1}) sur le reliquat d'azote minéral présent dans le sol d'une parcelle à la récolte. Dans le modèle hiérarchique, on considère qu'un

même modèle M reliant le reliquat à la dose d'azote s'applique à l'ensemble des sites-années du réseau et aux parcelles pour lesquelles il est considéré représentatif. Les équations et les paramètres fixes de ce modèle sont imparfaits et représentent donc des sources d'incertitude sur les reliquats prédits par le modèle. Cependant, on peut considérer que ces incertitudes résultent principalement d'un manque de connaissances et qu'elles pourraient être réduites en intensifiant et en améliorant les mesures et les méthodes d'analyse. Par ailleurs, on considère que plusieurs paramètres du modèle M varient d'un site-année à l'autre de façon aléatoire. En pratique, la variance de ces paramètres entre sites-années est due en partie à un défaut de connaissances, mais aussi en grande partie à de la variabilité aléatoire que l'on peut considérer comme intrinsèque au phénomène modélisé.

Du fait de l'existence d'incertitudes, il est souvent peu satisfaisant de répondre à une question en estimant les variables d'intérêt par des valeurs ponctuelles. Il est préférable d'associer à ces estimations une information sur le niveau d'incertitude qui leur est associé. Par exemple, si l'on souhaite évaluer les conséquences d'une réduction de la dose d'engrais azoté appliquée (par exemple, 180 plutôt que 200 kg ha^{-1}) sur le reliquat d'azote minéral à la récolte, il sera important de fournir une information sur le niveau d'incertitude associé à l'estimation réalisée.

Les sections suivantes présentent quelques méthodes pour représenter et analyser l'incertitude associée à des observations, à des modèles et à des règles de décision. La section 5.2 décrit les calculs de propagation d'incertitude, dont le principe est de générer des informations sur les distributions de probabilité des sorties d'un modèle à partir des incertitudes sur ses variables d'entrée et ses paramètres. La section 5.3 présente des méthodes d'analyse de sensibilité globale, dont l'objectif est de décomposer l'incertitude en sortie en fonction des facteurs incertains d'un modèle.

5.2 Décrire l'incertitude par des distributions de probabilité

5.2.1 Objectif

Différentes techniques peuvent être utilisées pour calculer des probabilités, des espérances, modes et variances de variables aléatoires. Dans certains cas, ces valeurs peuvent être calculées analytiquement. C'est le cas lorsque la variable aléatoire suit une loi de probabilité usuelle, par exemple une loi de Poisson, une loi uniforme, une loi gaussienne ou une loi Beta. Avec ces lois usuelles, les valeurs des probabilités, espérances, variances et modes ont une expression mathématique connue. Cependant, dans de nombreuses situations, aucune expression mathématique n'est disponible et les calculs doivent être réalisés à partir de tirages aléatoires et de simulations de Monte-Carlo. Trois exemples sont décrits ci-dessous.

5.2.2 Exemple basé sur des calculs analytiques : risque d'invasion par une espèce nuisible

Cet exemple illustre des calculs d'incertitude sur une variable aléatoire discrète.

Présentation du modèle

La carie de Karnal est une maladie du blé due au champignon *Tilletia indica*. Il s'agit d'un organisme de quarantaine qui est présent dans plusieurs pays, notamment au Moyen-Orient, mais absent dans diverses régions du monde comme l'Australie ou la France. Les pays non atteints par la maladie cherchent à empêcher son introduction sur leurs territoires et il est important pour ces pays de modéliser les risques d'entrée de *Tilletia indica* afin de définir les mesures de contrôle les plus adaptées.

Nous nous intéressons ici à l'estimation du risque d'entrée du champignon en Australie par la filière « voyageurs » à l'aide du modèle de Stansbury *et al.* (2002). Selon ce modèle, l'entrée du pathogène dépend du nombre de voyageurs arrivant en Australie, de la probabilité qu'un voyageur transporte le pathogène et de la probabilité que ce pathogène soit détecté. Le nombre d'entrées N_E du pathogène sur le territoire après un temps T est modélisé à l'aide d'une loi de Poisson :

$$N_E \sim \text{Poisson}\,(n\,p_c\,p_q\,T)$$

avec N_E le nombre d'entrées du pathogène au bout de T années, n le nombre de voyageurs venant de régions contaminées et entrant en Australie chaque année, p_c la probabilité qu'un voyageur soit porteur de la maladie, p_q la probabilité que la maladie ne soit pas détectée par l'inspection de quarantaine. D'après Stansbury *et al.* (2002), $n = 1\,000$, $p_c = 0.001$ et $p_q = 0.01$.

La loi de Poisson est une loi discrète souvent utilisée pour modéliser le nombre d'événements qui se produit dans un certain intervalle de temps. Avec cette distribution, le nombre moyen d'entrée par année est égal à $n \times p_c \times p_q$ et le nombre moyen d'entrées au bout de T années est égal à $\mu_E = n \times p_c \times p_q \times T$. La valeur de μ_E représente une estimation du nombre attendu d'entrées. Avec les valeurs des paramètres présentées ci-dessus, $\mu_E = 0.05$ avec $T = 5$ ans et $\mu_E = 0.5$ avec $T = 50$ ans.

Incertitude générée par la variabilité des entrées

Le nombre d'entrées de l'organisme nuisible est susceptible de varier selon l'échantillon de voyageurs arrivant dans le territoire. Il est ainsi possible que le nombre réel d'entrées diffère sensiblement de la moyenne μ_E. Il est donc utile de compléter les valeurs de μ_E en calculant les probabilités associées à différents nombres d'entrées possibles, par exemple en calculant les probabilités qu'il y ait zéro entrée, une entrée, deux entrées, trois entrées etc. au bout de T années.

Avec la loi de Poisson, ces probabilités sont définies par :

$$P(N_E = x) = \frac{\exp(-n\,p_c\,p_q\,T)(n\,p_c\,p_q\,T)^x}{x!}$$

où x est le nombre d'entrées du pathogène dans la zone géographique considérée.

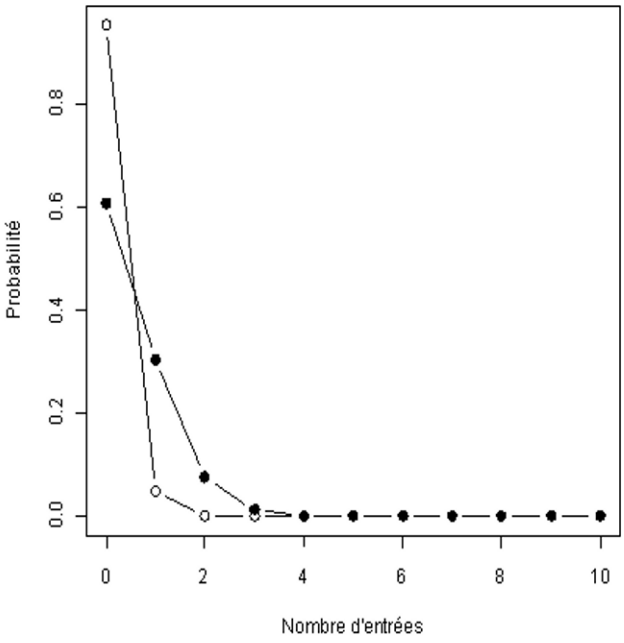

Figure 5.1 – Probabilités d'avoir $0, 1, 2, \ldots, 10$ entrées du pathogène de la carie de Karnal en Australie au bout de 5 ans (points blancs) et au bout de 50 ans (points noirs) par l'intermédiaire de la filière « voyageurs ».

Bien que cette formule soit assez simple, le programme R ci-dessous est utile pour les calculs. Dans ce programme, la probabilité d'avoir $x = 0, 1, 2, \ldots,$ 10 entrées est déterminée pour deux valeurs de T, soit 5 ans et 50 ans (`T.1` et `T.2`) à l'aide de la fonction `dpois` qui calcule la densité de la loi de Poisson pour des valeurs données des paramètres n, p_c et p_q. L'expression `0:10` crée un vecteur incluant les valeurs de x que l'on souhaite analyser. Les probabilités sont ensuite calculées puis archivées dans les objets `probas.1` et `probas.2`. Ces objets sont ensuite utilisés dans les instructions `plot` et `lines` de façon à présenter les probabilités sur un graphique (figure 5.1). Les résultats montrent que la probabilité d'avoir zéro entrée du pathogène au bout de 5 ans est égale à 0.95. Il est donc probable qu'aucune entrée ne se produise au bout de 5 ans. Ce n'est plus le cas au bout de 50 ans où la probabilité de n'avoir aucune entrée n'est égale qu'à 0.61, celle d'avoir une entrée est égale à 0.30 et celle d'avoir

deux entrées est égale à 0.08. Il n'est donc pas exclu qu'une ou deux entrées se produisent sur le territoire australien au bout de 50 ans par l'intermédiaire des voyageurs.

```
# Code R

T.1 <- 5
T.2 <- 50

n = 1000
pc = 0.001
pq = 0.01

Lambda.1 <- n * pc * pq * T.1
Lambda.2 <- n * pc * pq * T.2

probas.1 <- dpois(0:10, Lambda.1)
probas.2 <- dpois(0:10, Lambda.2)

plot(0:10, probas.1, type="b", xlab="Nombre d'entrees",
                               ylab="Probabilite",
                               cex=1.3)
lines(0:10, probas.2, type="b", pch=19, cex=1.3).
```

5.2.3 Exemple basé sur des simulations de Monte-Carlo : reliquat d'azote dans le sol

Rappels sur le modèle

Nous reprenons ici l'exemple de la section 3.3.2 qui traite du risque de pollution par l'azote minéral résiduel dans le sol. Des doses d'engrais trop importantes peuvent être à l'origine d'une augmentation de la quantité d'azote minéral présente dans le sol à la récolte et ainsi accroître la teneur en nitrate de l'eau de percolation. L'objectif de ce cas d'étude est de modéliser l'effet de la dose totale d'engrais minéral appliquée sur le blé d'hiver (*Triticum aestivum* L.) sur la quantité d'azote minéral résiduel, c'est-à-dire l'azote minéral présent dans le sol à la récolte et donc susceptible d'augmenter la teneur en nitrate de l'eau.

La variable d'intérêt est la quantité d'azote minéral présent dans le sol (0-90 cm) à la récolte, appelée « Reliquat N récolte » et notée R. Cette variable est reliée à la dose totale d'engrais azoté appliqué par une équation plateau-plus-linéaire définie par :

$$\begin{cases} R = \mathrm{R_{min}} & \text{si } x < \mathrm{X_{min}} \\ R = \mathrm{R_{min}} + A(x - \mathrm{X_{min}}) & \text{sinon.} \end{cases}$$

Cette équation inclut un vecteur de trois paramètres $\boldsymbol{\alpha} = (\mathrm{R_{min}}, A, \mathrm{X_{min}})^T$. Le paramètre $\mathrm{R_{min}}$ représente la valeur minimale de reliquat récolte, A est un taux d'augmentation du reliquat en fonction de la dose et $\mathrm{X_{min}}$ est la dose seuil au-delà de laquelle le reliquat augmente.

Les paramètres de ce modèle ont été estimés à partir de données expérimentales dans la section 3.4.2. Nous avons montré qu'il existait une forte incertitude dans les valeurs de $\mathrm{R_{min}}$ et $\mathrm{X_{min}}$ due à une variabilité significative des valeurs de ces paramètres entre sites-années. Leurs valeurs ne peuvent pas être raisonnablement considérées comme constantes. Nous avons montré dans la section 3.4.2 que la variabilité inter sites-années de $\mathrm{R_{min}}$ et $\mathrm{X_{min}}$ pouvait être décrite à l'aide de deux lois gaussiennes indépendantes définies par

$$\mathrm{R_{min}} \sim N\,(39.39, 76.61)\,,$$

$$\mathrm{X_{min}} \sim N\,(90.78, 1524.0)\,.$$

Ces deux distributions décrivent la variabilité entre sites-années des valeurs de $\mathrm{R_{min}}$ et $\mathrm{X_{min}}$. La variabilité entre sites-années de A n'étant pas significative, ce paramètre a été estimé par une valeur fixe égale à 0.13 kg/kg (voir la section 3.3.2).

Propagation de l'incertitude

Les paramètres $\mathrm{R_{min}}$ et $\mathrm{X_{min}}$ étant variables, il n'est pas satisfaisant de calculer une valeur unique de reliquat récolte pour une dose d'engrais x donnée. Il est préférable de déterminer la distribution des valeurs de reliquat récolte associée à une dose x donnée. Cette distribution de reliquat récolte peut être déduite des distributions de probabilités $\mathrm{R_{min}}$ et $\mathrm{X_{min}}$ et des équations du modèle par *propagation de l'incertitude*.

Deux approches sont envisageables. La première consiste à déterminer l'expression analytique de la distribution de reliquat récolte, c'est-à-dire sa formule mathématique. La seconde consiste à générer une série de valeurs de $\mathrm{R_{min}}$ et $\mathrm{X_{min}}$ tirées de façon aléatoire dans leurs distributions gaussiennes, puis à calculer à partir du modèle les valeurs correspondantes R de reliquat N à la récolte. La première approche n'est pas facile à réaliser ici car, du fait du seuil $\mathrm{X_{min}}$, le modèle n'est pas linéaire. La distribution de reliquat récolte n'a donc pas une expression analytique simple. Nous avons donc adopté la deuxième approche. Elle est basée sur le tirage aléatoire de valeurs de paramètres et est souvent appelée méthode de Monte-Carlo en référence aux jeux de hasard pratiqués dans le casino de cette ville.

Lorsqu'on réalise des tirages aléatoires, une question importante est : combien ? Un nombre de tirages trop faible conduira à des résultats imprécis. Le programme ci-dessous a été utilisé pour réaliser quatre séries de tirages de valeurs du paramètre $\mathrm{R_{min}}$: un tirage de 5 valeurs, un tirage de 50 valeurs, un de 500 valeurs et un de 5 000 valeurs. Les tirages sont réalisés à l'aide de l'instruction `rnorm` qui permet de générer un échantillon de valeurs tirées aléatoirement

dans une loi gaussienne. Cette fonction utilise en entrée le nombre de valeurs à générer $(5, 50, 500, 5000)$, l'espérance de la loi de probabilité (39.39), l'écart type de cette loi (la racine carrée de la variance). Les valeurs générées sont ensuite présentées graphiquement sous la forme d'un histogramme en utilisant l'instruction `hist`.

La figure 5.2 montre que les histogrammes des échantillons de taille 5 et 50 ne permettent pas de retrouver l'aspect de la loi gaussienne dont ils sont issus ; les histogrammes sont très dissymétriques. Les échantillons de taille 500 et 5 000 ont une allure plus conforme à celle que l'on attend avec une loi gaussienne. Pour vérifier qu'une taille de 5 000 est suffisante, nous avons répété quatre fois le tirage des 5 000 valeurs et comparé les caractéristiques de chaque échantillon avec l'instruction `summary`. Les caractéristiques étant très proches (par exemple la médiane estimée à partir des quatre échantillons est comprise entre 39.41 et 39.49), on peut considérer qu'une taille de 5 000 est suffisante. Il est toujours utile de vérifier ainsi la stabilité des sorties du modèle avec le nombre de simulations choisi.

```
# Code R: Generation de valeurs de parametres

par(mfrow=c(2,2))

Rmin.vec <- rnorm(5, 39.39, sqrt(76.61))
hist(Rmin.vec, xlab="Valeur de Rmin (kg/ha)", xlim=c(0,100),
                                              main="N=5")

Rmin.vec <- rnorm(50, 39.39, sqrt(76.61))
hist(Rmin.vec, xlab="Valeur de Rmin (kg/ha)", xlim=c(0, 100),
                                              main="N=50")

Rmin.vec <- rnorm(500, 39.39, sqrt(76.61))
hist(Rmin.vec, xlab="Valeur de Rmin (kg/ha)", xlim=c(0, 100),
                                              main="N=500")

Rmin.vec <- rnorm(5000, 39.39, sqrt(76.61))
hist(Rmin.vec, xlab="Valeur de Rmin (kg/ha)", xlim=c(0, 100),
                                              main="N=5000")
```

La génération des valeurs des paramètres correspond à la première étape de l'analyse d'incertitude. La deuxième étape consiste à utiliser le modèle plateau-plus-linéaire pour calculer les valeurs de reliquat récolte associées aux valeurs de paramètres générées. Si 5 000 valeurs de paramètres sont générées à la première étape, 5 000 valeurs de reliquat doivent être obtenues lors de la seconde étape pour chaque dose d'engrais testée. Finalement, la dernière étape consiste à décrire la distribution des valeurs de reliquat récolte.

```
# Code R: Calcul de reliquat recolte
```

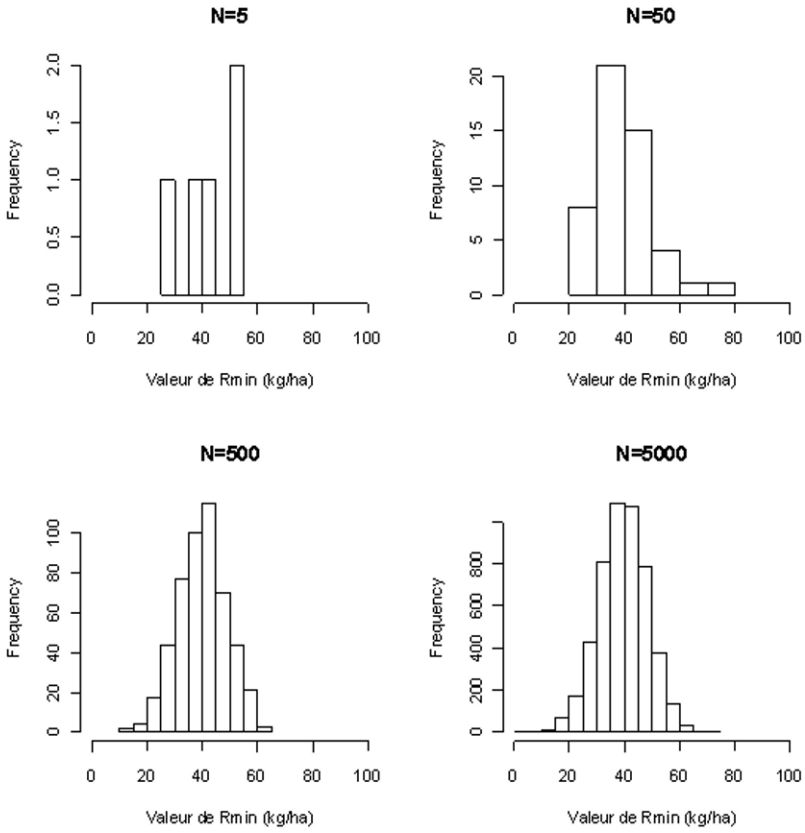

Figure 5.2 – Histogrammes des valeurs du paramètre R_{min} tirées aléatoirement dans une loi gaussienne. Quatre échantillons de taille 5, 50, 500 et 5 000 ont été générés.

```
par(mfrow=c(2,3))

Rmin.vec <- rnorm(5000, 39.39, sqrt(76.61))
A.vec    <- 0.1266
Xmin.vec <- rnorm(5000, 90.78, sqrt(1524.0))
Rel.vec <- 1:5000

for (j in 1:5000) {
  Rel.vec[j] <- Mod(50, Rmin.vec[j], A.vec, Xmin.vec[j])
  }

hist(Rel.vec, xlim=c(0, 100), xlab="Reliquat N (kg/ha)",
                              main="Dose=50 kg/ha")

boxplot(Rel.vec, range=0, ylim=c(0, 100))
summary(Rel.vec)

Proba.vec <- 1:100

for (i in 1:100) {
  Proba.vec[i] <- length(Rel.vec[Rel.vec<i]) / length(Rel.vec)
  }

plot(1:100, Proba.vec, xlab="Reliquat N (kg/ha)",
                       ylab="Probabilite Reliquat<x",
                       ylim=c(0,1), type="l", lwd=2)

#Fonction plateau-plus-lineaire

Mod <- function(Dose, Rmin, A, Xmin) {

  if (Dose < Xmin) Y <- Rmin
  else Y <- Rmin + A * (Dose - Xmin)

  return(Y)
}
```

Le programme R ci-dessus permet d'enchaîner les trois étapes. Cinq mille valeurs de R_{min} et X_{min} sont générées à l'aide de `rnorm`. Le paramètre constant A est fixé à sa valeur estimée 0.1266. Une boucle `for` est ensuite utilisée pour calculer les 5 000 valeurs de reliquat récolte. Celles-ci sont obtenues à l'aide d'une fonction appelée `Mod` qui prend en entrée la valeur de la dose d'engrais (fixée ici à 50 kg/ha) et les valeurs des trois paramètres. La fonction `Mod` est basée sur une équation plateau-plus-linéaire écrite en langage R (voir ci-dessus).

Les 5 000 valeurs de reliquat sont stockées dans un objet appelé `Rel.vec`. Elles sont ensuite présentées de trois façons : sous forme d'histogramme (à l'aide de l'instruction `hist`), sous forme de *boîte à moustaches* (avec `boxplot`) et en utilisant l'instruction `summary`. Cette dernière retourne les résultats suivants :

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.632	33.760	39.630	39.720	45.610	71.520

Ces six valeurs correspondent respectivement aux valeurs minimale, 1er quartile (valeur dépassée 3 fois sur 4), médiane, moyenne, 3e quartile (valeur dépassée 1 fois sur 4) et maximale. Les dernières lignes du programme permettent de calculer les probabilités cumulées, c'est-à-dire les probabilités que le reliquat soit inférieur à des valeurs données. Ici, des valeurs de reliquat comprises entre 1 et 100 kg/ha ont été considérées successivement. Les 100 probabilités associées ont été calculées à l'aide d'une boucle. La présentation graphique des probabilités cumulées permet de visualiser la fonction de répartition de la distribution de reliquat récolte.

Ce programme peut être mis en œuvre avec différentes doses d'engrais afin de comparer les distributions de reliquat associées à des apports d'engrais plus ou moins importants. Les résultats obtenus avec les doses 50 et 200 sont présentés sur la figure 5.3. Ils montrent que les reliquats associés à une dose de 200 kg/ha sont plus élevés que ceux associés à une dose de 50 kg/ha mais qu'il y a un recouvrement partiel des deux distributions. La valeur minimale de reliquat obtenue avec une dose de 200 kg/ha est en effet égale à 15.6 kg/ha et cette valeur est inférieure à la valeur de reliquat maximale obtenue avec une dose de 50 kg/ha (71.52 kg/ha). Les probabilités cumulées sont cependant très différentes. Ainsi, lorsqu'une dose de 50 kg/ha est appliquée, la probabilité que le reliquat récolte soit inférieur à 40 kg/ha est de 0.51, soit environ une chance sur deux. Par contre, lorsqu'une dose de 200 kg/ha est appliquée, cette même probabilité tombe à 0.09, soit moins d'une chance sur 10.

5.2.4 Exemple combinant un modèle dynamique et des mesures en cours de saison : estimation du carbone du sol

Présentation du modèle

Les activités agricoles sont à l'origine d'émission de gaz à effet de serre ; CO_2, N_2O, CH_4. L'étude de la contribution de l'agriculture aux émissions et au stockage de CO_2 passe par la détermination des quantités de carbone stockées dans le sol. Ces quantités peuvent être déterminées par des mesures expérimentales, par simulation à l'aide de modèles mathématiques ou, comme proposé par Jones *et al.* (2004) et par Jones et Graham (2006), en combinant les simulations d'un modèle avec des mesures acquises en cours de saison.

Nous montrons ici comment l'incertitude associée aux prédictions d'un modèle dynamique simulant le carbone du sol peut être décrite à l'aide d'une loi

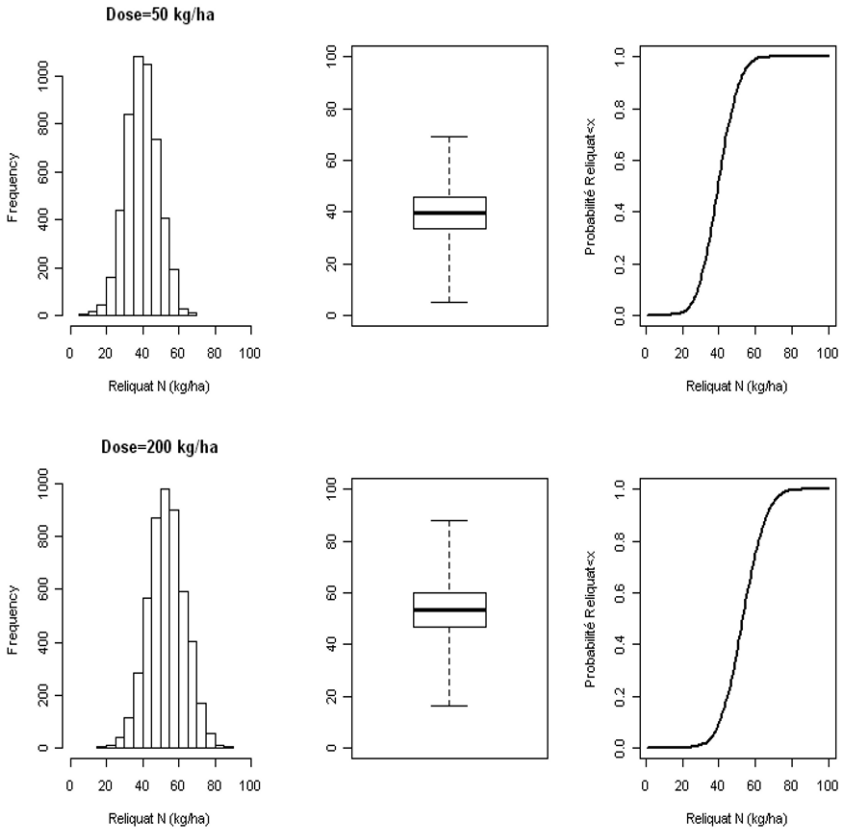

Figure 5.3 – Histogramme, boxplot et probabilités cumulées des 5 000 valeurs simulées de reliquat récolte pour une dose d'engrais de 50 kg/ha et une dose de 200 kg/ha. Les éléments du boxplot correspondent (de haut en bas) aux valeurs maximale, 3^{e} quartile, médiane, 1er quartile, minimale.

de probabilité. Nous montrons également comment cette distribution peut être mise à jour en tenant compte des mesures réalisées en cours de saison.

Pour illustrer cette approche, nous reprenons ici le modèle dynamique stochastique présenté dans Jones et Graham (2006) et défini par :

$$Z_{t+1} = (1 - R)\, Z_t + B\, U_t + \varepsilon_t \tag{5.1}$$

où Z_t est la quantité de carbone dans le sol l'année t (kg ha^{-1}), R est le taux de décomposition annuel du carbone (supposé ici être égal à 0.01), U_t est la quantité de carbone en provenance de la culture qui est ajoutée au sol l'année t (fixé ici à 2 000 kg ha^{-1}an^{-1}), B est la fraction de cette quantité qui subsiste dans le sol au bout d'un an (fixée à 0.2), ε_t est un terme aléatoire représentant l'erreur du modèle supposé distribué selon une loi gaussienne et $\varepsilon_t \sim N\,(0, 20\,000)$.

Assimilation de données dans l'analyse d'incertitude

Ce modèle simule la quantité de carbone dans le sol à un pas de temps annuel. Si une mesure de carbone du sol M_t est réalisée l'année t, il est possible d'utiliser cette mesure pour améliorer la prédiction du modèle en calculant la loi de probabilité de Z_t conditionnelle à M_t. Dans le cas du modèle de Jones et Graham (2006), cette approche conduit au même résultat que le filtre de Kalman. Nous le montrons ci-dessous pour $t = 1$.

La prédiction du modèle à $t = 1$ est $(1-R) \times Z_0 + B \times U_O$. Cette prédiction comporte deux sources d'erreur :

- $\varepsilon_0 \sim N\,(0, 20\,000)$, erreur liée à la structure du modèle ;
- $Z_0 \sim N\,(16\,000, 20\,000)$, erreur liée à la connaissance imparfaite de l'état initial Z_0.

La loi de Z_1 peut être déduite des lois de ε_0 et de Z_0 et de l'équation (5.1) du modèle. Un calcul analytique conduit à

$$Z_1 \sim N\left(16\,000 - 0.01 \times 16\,000 + 0.2 \times 2\,000, (1 - 0.01)^2 \times 20\,000 + 20\,000\right)$$
$$\text{soit } Z_1 \sim N\,(16\,240, 39\,800)\,.$$

Cette loi de probabilité correspond à la loi *a priori* de Z_1, c'est-à-dire la distribution de Z_1 avant la réalisation de la mesure.

Si une mesure M_1 est disponible, il est possible de modifier cette distribution et de calculer une loi *a posteriori* tenant compte de la mesure. Nous supposons ici que $M_t = Z_t + \eta_t$, avec $\eta_t \sim N\,(0, 500\,000)$. La mesure M_t conditionnelle à Z_t est donc distribuée selon une loi gaussienne

$$M_t \mid Z_t \sim N\,(Z_t, 500\,000)\,.$$

On en déduit que

$$M_1 \sim N\,(16\,240, 39\,800 + 500\,000)\,;$$
$$\mathrm{Cov}(Z_1, M_1) = \mathrm{Var}(Z_1).$$

La loi *a posteriori* correspond à la loi de Z_1 conditionnelle à M_1. Elle peut être calculée analytiquement car le modèle est linéaire et la loi conjointe de Z_1 et M_1 est gaussienne. Le calcul de la loi conditionnelle à partir de la loi conjointe conduit à :

$$Z_1 \mid (M_1 = m_1) \sim N\left[E(Z_1) + K_1\left(m_1 - E(M_1)\right), (1 - K_1)\ \mathrm{Var}(Z_1)\right]$$

où $E(Z_1)$ est l'espérance *a priori* (*i.e.*, 16 240), $\mathrm{Var}(Z_1)$ est la variance *a priori* (*i.e.*, 39 800), K_1 est le coefficient de Kalman défini par

$$K_1 = \frac{\mathrm{Var}(Z_1)}{\mathrm{Var}(Z_1) + \mathrm{Var}(M_1 \mid Z_1)}.$$

Comme $\mathrm{Var}(M_1 \mid Z_1) = 500\,000$, la distribution *a posteriori* est définie par

$$Z_1 \mid (M_1 = m_1) \sim N\left[16\,240 + 0.44(m_1 - 16240), 22\,160.36\right].$$

Cette expression montre que la variance de la distribution *a posteriori* est plus faible que la variance de la distribution *a priori* (22 160.36 contre 39 800). L'utilisation d'une mesure permet donc de réduire l'incertitude du modèle, même si cette mesure est elle-même imparfaite. L'espérance de la distribution *a posteriori* est égale à $16\,240 + 0.44 \times (m_1 - 16\,240)$. Elle représente un compromis entre l'espérance *a priori* (16 240) et la mesure. Par exemple, si la mesure est égale à 20 000, l'espérance *a posteriori* est égale à 17 894.4.

La figure 5.4 présente les distributions *a priori* et *a posteriori* obtenues pour $m_1 = 20\,000$. Cette figure a été générée à l'aide du programme R suivant :

```
# Code R: Presentation graphique des distributions a priori et
#         a posteriori

Carb<-15000:20000

Proba1<-dnorm(Carb,16240,sqrt(39800))
Proba2<-dnorm(Carb,17894.4,sqrt(22160.36))

par(mfrow=c(1,2))
plot(Carb, Proba1, xlab="Carbone du sol (kg/ha)",
                   ylab="Densite de probabilite", type="l")
plot(Carb, Proba2, xlab="Carbone du sol (kg/ha)",
                   ylab="Densite de probabilite", type="l")
```

Cette approche peut être appliquée à chaque pas de temps pour corriger séquentiellement le modèle. Ici, les calculs ont pu être réalisés analytiquement car le modèle considéré était linéaire. Pour les modèles plus complexes, le calcul analytique n'est généralement pas possible et d'autres techniques doivent être utilisées, par exemple des techniques basées sur des simulations de Monte-Carlo. Voir par exemple Doucet *et al.* (2000) pour une description de ces techniques, ainsi que Makowski *et al.* (2006) et Naud *et al.* (2007) pour des applications dans un contexte agronomique.

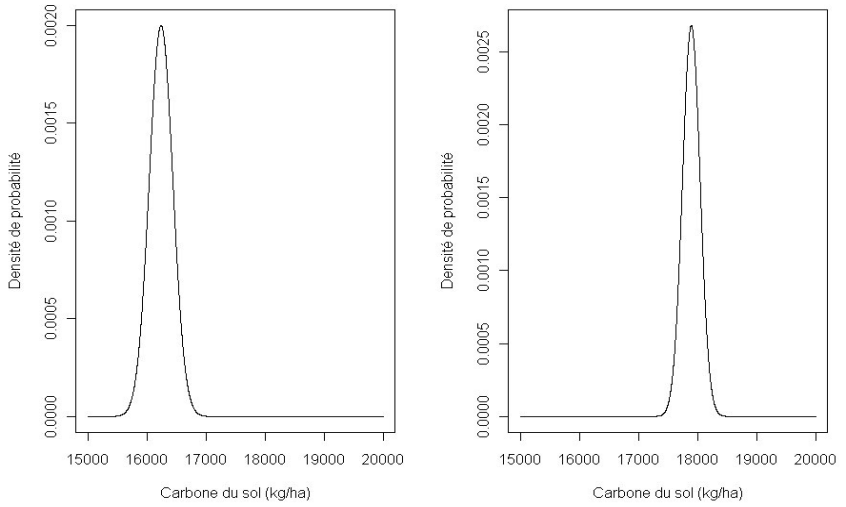

Figure 5.4 – Distributions *a priori* (a) et *a posteriori* (b) du carbone du sol lorsque la mesure utilisée pour corriger le modèle est égale à 20 000 kg/ha.

5.3 Calculer des indices de sensibilité

5.3.1 Objectifs et définitions

L'analyse de sensibilité a pour objectif d'évaluer l'influence de différents composants d'un modèle (appelés facteurs d'entrée de l'analyse de sensibilité) sur les sorties de ce modèle.

L'analyse de sensibilité est souvent effectuée à la suite des calculs d'incertitude présentés dans les sections 5.1 et 5.2. Elle permet alors d'identifier les paramètres et les variables d'entrée qui sont à l'origine des incertitudes sur les prédictions fournies par le modèle. Dans ce cas, les facteurs d'entrée de l'analyse de sensibilité sont les paramètres et les variables d'entrée incertains du modèle considéré, c'est-à-dire les paramètres connus approximativement et variables d'entrée sujettes à de la variabilité. Cependant l'analyse de sensibilité n'est pas seulement utile à la suite d'une analyse d'incertitude. Dans de nombreuses situations, l'analyse de sensibilité permet de mieux comprendre le comportement d'un modèle complexe. Elle permet en particulier d'identifier les paramètres dont il faut réduire l'incertitude en priorité ou ceux dont l'influence sur la sortie du modèle est négligeable.

Il existe de nombreuses méthodes d'analyse de sensibilité (Monod *et al.*, 2006 ; Saltelli *et al.*, 2008). Elles diffèrent d'abord par la définition même de la sensibilité. Ainsi, la sensibilité locale désigne essentiellement les dérivées du modèle autour de valeurs ponctuelles des paramètres. La sensibilité glo-

bale, par contre, porte sur la variabilité des sorties lorsque plusieurs facteurs d'entrée varient dans des gammes de variation plus importantes représentant l'incertitude ou la variabilité d'un phénomène. Nous nous intéressons ici aux méthodes d'analyse de la sensibilité globale car elles permettent de traiter un grand nombre de problèmes pratiques.

5.3.2 Exemple basé sur des simulations de Monte-Carlo : reliquat d'azote minéral dans le sol

Nous reprenons ici l'exemple traité dans la section 5.2.3. L'objectif est de réaliser une analyse de sensibilité afin de comprendre et de comparer le rôle que jouent la dose d'azote et les paramètres R_{min} et X_{min} sur le reliquat d'azote dans le sol simulé par le modèle plateau-plus-linéaire.

Nous réutilisons les 5 000 valeurs de reliquat simulées avec le modèle à partir des 5 000 valeurs des paramètres R_{min} et X_{min} générées aléatoirement avec les programmes décrits dans la section 5.2.3. Il est toujours utile de commencer par des représentations graphiques des valeurs simulées. Le programme R présenté ci-dessous construit deux diagrammes de dispersion entre R_{min} et X_{min} d'une part, le reliquat d'azote d'autre part. Ces diagrammes sont bien adaptés à des variables quantitatives, mais pas à la dose d'azote qui ne prend que deux valeurs distinctes dans notre jeu de données simulées (50 et 200 kg/ha). L'influence de la dose sur le reliquat peut par contre être correctement analysée en réalisant un diagramme en boîte (boxplots), mieux adapté aux variables qualitatives ou discrétisées en un petit nombre de valeurs. Il est également possible d'analyser l'effet de la dose en utilisant deux couleurs distinctes dans les diagrammes de dispersion.

```
# Code R: Graphiques des valeurs de reliquat récolte

donneesN <- data.frame( Rmin=rep(Rmin.vec,2),
                        Xmin=rep(Xmin.vec,2),
                        Dose=rep( c(50,200), c(5000,5000) ),
                        Reliquat=NA)

for (j in 1:5000) {
  donneesN$Reliquat[j] <- Mod(50, Rmin.vec[j], A.vec, Xmin.vec[j])
  }

for(j in 1:5000){
  donneesN$Reliquat[5000+j,3] <- Mod(200, Rmin.vec[j], A.vec,
  Xmin.vec[j])
  }

par(mfrow=c(1,3))
```

```
scatter.smooth(donneesN$Rmin, donneesN$Reliquat, cex=0.1,
            col=c("blue","red")[as.factor(donneesN$Dose)],
            xlab="Rmin", ylab="Reliquat N (kg/ha)")

scatter.smooth(donneesN$Xmin, donneesN$Reliquat, cex=0.1,
            col=c("blue","red")[as.factor(donneesN$Dose)],
            xlab="Xmin", ylab="Reliquat N (kg/ha)")

boxplot(Rel.vec ~ dose, data=donneesN,
            xlab="Dose (kg/ha)", ylab="Reliquat N (kg/ha)")
```

Le résultat de ces instructions est présenté dans la figure 5.5. On observe une forte corrélation linéaire entre R_{min} et le reliquat azote, alors que la relation entre X_{min} et le reliquat est moins marquée. Globalement, les nuages de points associés aux deux doses se recouvrent fortement. Pour une valeur R_{min} fixée, les deux doses donnent néanmoins des distributions de points bien distinctes.

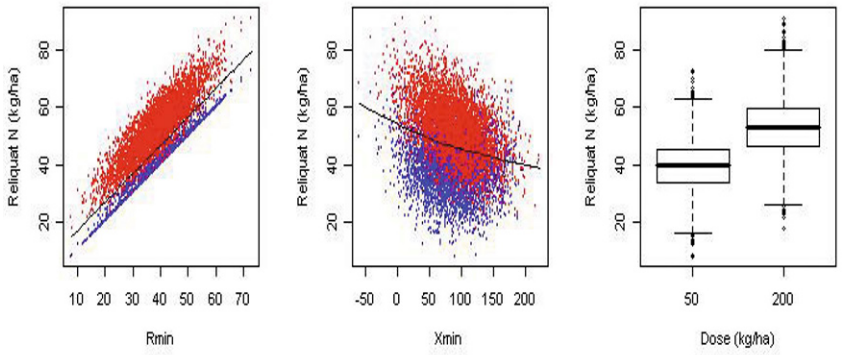

Figure 5.5 – Diagramme de dispersion reliant les valeurs des paramètres X_{min} et R_{min} aux valeurs de reliquat d'azote simulées par le modèle (rouge pour la dose de 200 kg/ha et bleu pour la dose de 50 kg/ha) et diagramme en boîte montrant la distribution des valeurs simulées de reliquat azote pour les deux doses d'engrais.

La figure 5.5 semble indiquer que R_{min} a un effet plus important que X_{min}. Il est cependant nécessaire de confirmer cette impression visuelle à l'aide de critères quantitatifs. Une approche possible consiste à calculer des indices de sensibilité basés sur des techniques de régression linéaire. Ceci est d'autant plus justifié ici que les graphiques montrent des relations assez simples entre facteurs d'entrée et variable de sortie. Le programme R ci-dessous permet d'ajuster aux données simulées un modèle linéaire comprenant les effets quadratiques de X_{min} et R_{min}, ainsi que l'interaction de ces deux facteurs avec la dose.

```
# Code R: ajustement linéaire

don.lm3 <- lm(Rel.vec ~  dose + poly(Xmin.vec,2) +
                poly(Rmin.vec,2) + dose:Xmin.vec +
                dose:Rmin.vec, data=donnees2)
summary(don.lm3)
```

Le modèle s'ajuste parfaitement aux données : le coefficient de détermination R^2 est supérieur à 0.99. Les résultats d'estimation des coefficients de régression sont présentés dans le tableau 5.1. Pour interpréter ces résultats, il est important de ne pas oublier que les données sont simulées à partir du plateau-plus-linéaire. Il n'y a donc pas d'erreur de mesure au sens habituel du terme et les tests de signification des effets n'ont ici aucun intérêt. Par contre les statistiques du t de Student sont utiles pour comparer les termes entre eux, puisqu'elles sont normalisées par rapport à la gamme de variation des facteurs d'entrée. On vérifie ainsi que R_{min} a un effet purement linéaire et que l'effet de la dose est quasiment indépendant de la valeur de ce paramètre. En revanche, l'effet de la dose dépend fortement de la valeur de X_{min}.

Variable explicative	Valeur estimée	Écart type	t de Student
dose	$1.55e-01$	$(3.8e-04)$	403.93
X_{min} (effet linéaire)	$6.43e+01$	$(1.1e+00)$	58.23
X_{min} (effet quadratique)	$4.17e+01$	$(5.7e-01)$	73.36
R_{min} (effet linéaire)	$8.90e+02$	$(1.1e+00)$	805.84
R_{min} (effet quadratique)	$1.26e-02$	$(5.7e-01)$	0.02
Interaction dose : X_{min}	$-7.14e-04$	$(1.9e-06)$	−366.58
Interaction dose : R_{min}	$-9.29e-06$	$(8.5e-06)$	−1.09

Tableau 5.1 – Coefficients de régression estimés avec les reliquats d'azote simulés par Monte-Carlo (5 000 simulations par dose) à l'aide du modèle plateau-plus-linéaire.

L'influence des principaux facteurs identifiés ci-dessus peut être synthétisée par des indices de sensibilité. Dans le cadre de la régression, on utilise couramment les coefficients de régression standardisés (SRC, pour *standardised regression coefficients*), égaux à $\widehat{b}_i(\sigma_i/\sigma_Y)$, où $\widehat{b}_i, \sigma_i, \sigma_Y$ désignent respectivement le coefficient de régression estimé pour la variable explicative i et les écarts types de cette variable et de la variable de sortie dans le jeu de données simulées. L'encadré ci-dessous effectue le calcul des SRC en utilisant la librairie `sensitivity` de R, à partir du modèle de régression précédent dont on a éliminé les termes négligeables. Ces coefficients mesurent l'importance des facteurs d'entrée de l'analyse d'incertitude initiale, soit ici la dose d'engrais et les deux paramètres X_{min} et R_{min}.

```
#Code R : utilisation de la librairie sensitivity
```

```
library(sensitivity)

modmat <- model.matrix(Rel.vec ~  dose + poly(Xmin.vec,2) +
                  Rmin.vec + dose:Xmin.vec, data=donnees2)

don.src <- src(modmat[,-1], donnees2$Rel.vec)

print(round(don.src$SRC, 2))
```

Les coefficients de régression standardisés sont présentés dans le tableau 5.2. Les valeurs les plus élevées sont celles obtenues pour la dose d'engrais et le paramètre R_{min}. Ces coefficients confirment les résultats de la figure 5.5 mais apportent deux éléments supplémentaires : i) l'effet de la dose est un peu plus fort que celui de R_{min}, ii) il existe une forte interaction entre la dose et X_{min} ce qui indique que l'effet de la dose d'engrais dépend de X_{min}.

Variable explicative	Coefficient
Dose	0.99
X_{min} (effet linéaire)	0.05
X_{min} (effet quadratique)	0.04
R_{min}	0.76
Interaction dose : X_{min}	- 0.54

Tableau 5.2 – Coefficients de régression standardisés estimés avec les reliquats d'azote simulés par Monte-Carlo (5 000 simulations par dose).

5.3.3 Exemple basé sur des simulations planifiées : influence du parcellaire sur les flux de gènes

Le modèle plateau-plus-linéaire utilisé ci-dessus pour simuler le reliquat d'azote est très simple. Bien souvent, les modèles utilisés par les agronomes sont beaucoup plus complexes et peuvent inclure plusieurs dizaines voire plusieurs centaines de paramètres et de variables d'entrée. Le temps calcul requis pour réaliser les simulations peut par ailleurs être non négligeable avec certains modèles. Dans cette situation, il est important d'utiliser des méthodes rigoureuses de planification et d'analyse des simulations (Saltelli *et al.*, 2008).

Le modèle MAPOD (Angevin *et al.*, 2008) a été développé comme un outil d'aide à la décision pour comparer des scénarios et évaluer les conditions de coexistence entre cultures de maïs OGM et non OGM. C'est un modèle spatio-temporel, qui prend en entrée une carte des champs de maïs OGM et non OGM (figure 5.6), les variétés, dates et densités de semis sur chaque parcelle, et des données climatiques en particulier directions et vitesses du vent pendant la

floraison. À partir de ces données d'entrée, MAPOD calcule la dynamique de floraison sur chaque parcelle, la dispersion du pollen et les taux de maïs OGM à la récolte sur chaque parcelle de maïs non OGM. Contrairement à l'exemple de la section 5.3.2, ce modèle fait intervenir un nombre élevé de paramètres et de variables d'entrée, avec des relations complexes entre ces éléments et un temps calcul important qui limite le nombre de simulations possibles.

Dans l'article de Viaud *et al.* (2008), une analyse de sensibilité est menée sur MAPOD dans le but de mieux comprendre l'influence de la géométrie et de la répartition des champs de maïs sur les taux de pollinisation croisée. La méthode utilisée est celle des plans factoriels, suivis d'une analyse de variance. Le premier facteur d'entrée, `carte`, est constitué de six parcellaires, avec des tailles et des formes de parcelles diversifiées. La répartition du maïs est décrite par trois facteurs : `pmaïs`, pourcentage de surface couverte en maïs (20% ou 70%), `pgm`, pourcentage de surface de maïs couverte en OGM (10% ou 50%), `agreg`, mode de répartition des parcelles de maïs OGM et non OGM (aléatoire, agrégats séparés, agrégats mixtes). D'autres facteurs ont été introduits afin de pouvoir détecter des interactions entre les facteurs liés au paysage d'une part et les quantités et distances de dispersion du pollen d'autre part : `rpoll`, ratio entre les quantités de pollen produites par les variétés OGM et non OGM (0.33 ou 3.75), `dh`, différence de hauteur entre épis et panicules (0,8 ou 1,75 m), `ventv` et `ventd`, vitesse et direction moyennes du vent pendant la floraison (10 modalités en tout). Pour simplifier l'interprétation, les floraisons ont été supposées synchrones entre les variétés OGM et non OGM, bien que le décalage de floraison joue un rôle primordial sur les taux de pollinisation croisée.

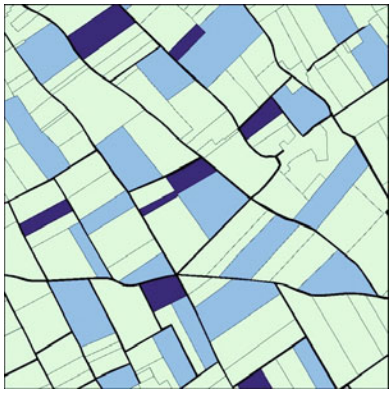

Figure 5.6 – Exemple de parcellaire agricole utilisé comme entrée du modèle MAPOD. Les parcelles cultivées en maïs sont représentées en bleu clair (variétés non OGM) et bleu foncé (variétés OGM).

Le plan factoriel complet contient une fois chacune des combinaisons possibles entre les facteurs d'entrée, soit $6\times2\times2\times3\times2\times2\times10 = 2880$ combinaisons. Ce plan est généré par la première commande du programme R présenté dans

l'encadré ci-dessous, définissant ainsi l'ensemble des scénarios à simuler à partir des combinaisons des facteurs d'entrée. Selon la façon dont le modèle étudié est programmé, les simulations peuvent être conduites à partir de R ou à partir d'un programme extérieur. Dans notre cas, l'ensemble des simulations a été lancé à partir de R grâce à des appels vers le langage de programmation de MAPOD. Nous ne détaillerons pas plus ce volet qui est forcément très dépendant de l'application considérée, mais il est utile de noter que, après la phase de simulations, le plan d'expérience créé par `expand.grid` a été complété par des colonnes contenant les sorties des simulations, ici le logarithme du taux de pollinisation croisée, en moyenne sur l'ensemble des parcelles non OGM du parcellaire.

L'analyse de la variance (commande `aov`) permet de calculer les parts de variabilité de la variable de sortie dues aux différents facteurs d'entrée et à leurs interactions. Pour faciliter l'interprétation, la formule du modèle d'anova ci-dessous regroupe en un seul terme les facteurs liés à la répartition du maïs, ceux liés aux caractères variétaux et ceux liés au vent. Les interactions entre ces groupes de facteurs sont incluses jusqu'à l'ordre 3. Comme l'analyse porte sur des résultats de calcul déterministes, les statistiques de Fisher et les probabilités associées n'ont aucun sens ici. À la place, on calcule des indices de sensibilité variant de 0 à 1 en divisant les sommes de carrés de chaque terme par la somme de carrés totale. Ces indices de sensibilité sont les analogues, pour des plans factoriels, des indices de sensibilité définis dans Saltelli *et al.* (2008). Voir aussi Ginot *et al.* (2006).

```
# Code R: calcul d'indices de sensibilite

plan <- expand.grid(carte=c("A1","O1","P1","S1","S4","T1"),
          pmais=c("20%","70%"), pGM=c("10%","50%"),
          aggr=c("ale","sep","mix"),
          rpoll=c(0.33,3.75), dh=c(0.8,1.75),
          vent=1:10)

plan <- mapod.f(plan)

# mapod.f = fonction qui lance une simulation de MAPOD pour chaque
# ligne de ''plan'')

# Le taux moyen de pollinisation croisée (taux de grains OGM dans les
# parcelles non OGM du paysage) est calcule apres chaque
# simulation et stocke dans la colonne mean de "plan''
# et son logarithme dans la colonne logmean.

plan.aov <- aov(logmean ~
       (carte+(pmais~:pgm~:agreg)+(dh~:rpoll)+(vent))^3, data=plan)
```

```
aov.summ <- summary(plan.aov)[[1]]
aov.ss <- aov.summ[,"Sum Sq"]
indices <- data.frame(names=rownames(aov.summ),
                                    IS=aov.ss/sum(aov.ss))
```

Les résultats sont présentés dans le tableau 5.3, à l'exception de deux interactions négligeables d'ordre 3. Logiquement le groupe de facteurs qui a le plus d'effet sur la variabilité du taux moyen est la répartition du maïs (surface et organisation), mais ce résultat était attendu. L'analyse met aussi en évidence l'importance des effets variétés et surtout l'importance des interactions entre les facteurs. Ainsi, l'effet du fonds de carte semble négligeable si l'on considère uniquement son effet principal, alors que ce facteur intervient fortement au travers d'interactions avec la répartition du maïs et le régime de vent. Nous référons à Viaud *et al.* (2008) pour une analyse beaucoup plus détaillée des résultats de ces simulations.

Terme du modèle	Indice de sensibilité (%)
Carte	1
Répartition	35
Variété	29
Vent	5
Carte × Répartition	13
Carte × Variété	< 1
Carte × Vent	1
Répartition × Variété	1
Répartition × Vent	1
Variété × Vent	< 1
Carte × Répartition × Variété	< 1
Carte × Répartition × Vent	14

Tableau 5.3 – Indices de sensibilité de l'analyse effectuée sur MAPOD.

5.4 Exercices

Exercice 5.1. Considérons le modèle $y(z_1, z_2) = z_1 + 2z_2$, où z_1 et z_2 sont deux facteurs incertains. Supposons que l'incertitude de z_1 et z_2 soit décrite par deux lois gaussiennes indépendantes, $z_1 \sim N(20, 16)$ et $z_2 \sim N(60, 64)$. Déterminer la loi décrivant l'incertitude de $y(z_1, z_2)$.

Exercice 5.2. Générer 10 séries de 100 valeurs dans une loi gaussienne de moyenne 10 et de variance 16 avec R. Estimer l'espérance de la loi gaussienne

avec chaque série. Calculer la racine carrée de l'erreur quadratique moyenne des estimations.

Exercice 5.3. Considérons le modèle de Magarey *et al.* (2005) qui prédit la durée requise d'humidité W pour qu'un champignon puisse infecter une plante lorsque la température est égale à T. Ce modèle est défini par $W = W_{\min}/f(T)$ et :

$$f(T) = \left(\frac{T_{\max} - T}{T_{\max} - T_{\text{opt}}}\right)\left(\frac{T - T_{\min}}{T_{\text{opt}} - T_{\min}}\right)^{(T_{\text{opt}} - T_{\min})/(T_{\max} - T_{\text{opt}})}$$

Il inclut cinq paramètres : $T_{\min}$, T_{opt}, $T_{\max}$, $W_{\min}$, $W_{\max}$. Des expérimentations ont été réalisées pour estimer ces paramètres pour un champignon x et une espèce de plante y, mais les valeurs de ces paramètres demeurent incertains. Les gammes de valeurs possibles pour les paramètres sont données dans le tableau 5.4.

	Min	Max
$T_{\min}$ (°C)	10	15
$T_{\max}$ (°C)	32	35
T_{opt} (°C)	25	30
$W_{\min}$ (h)	12	14
$W_{\max}$ (h)	35	48

Tableau 5.4 – Gammes de valeurs pour les paramètres du modèle de Magarey *et al.* (2005).

Questions :

- Définir des distributions uniformes pour décrire l'incertitude des paramètres.
- Tirer 1 000 valeurs de paramètres dans ces distributions avec R.
- Créer une fonction R incluant le modèle et utiliser cette fonction pour calculer les valeurs de W pour les 1 000 valeurs de paramètres avec $T = 20\,°\text{C}$.
- Décrire la distribution de W ainsi obtenue en traçant un histogramme et en calculant la médiane ainsi que les 1-er et 3-ième quartiles.
- Calculer la probabilité que W soit inférieure à 20 h.

Exercice 5.4. On reprend le modèle décrit dans l'exercice 5.3 dans le but de réaliser une analyse de sensibilité.

- Utiliser l'instruction R `expand.grid` pour définir un plan d'expérience factoriel complet combinant les valeurs minimales et maximales des cinq paramètres du modèle.
- Calculer W pour toutes les combinaisons de paramètres du plan d'expérience.
- Réaliser une analyse de variance avec l'instruction R `aov` pour estimer les indices de sensibilité des paramètres.

– Identifier le paramètre le plus influent à partir des résultats.

Exercice 5.5. Les indices de sensibilité totaux de la teneur en protéines du blé simulée par un modèle de culture ont été calculés pour 13 paramètres de ce modèle. Leurs valeurs sont présentées sur la figure 5.7, avec les intervalles de confiance. D'après ce graphique, quelles sont les paramètres qu'il est important d'estimer avec précision pour simuler correctement la teneur en protéines ?

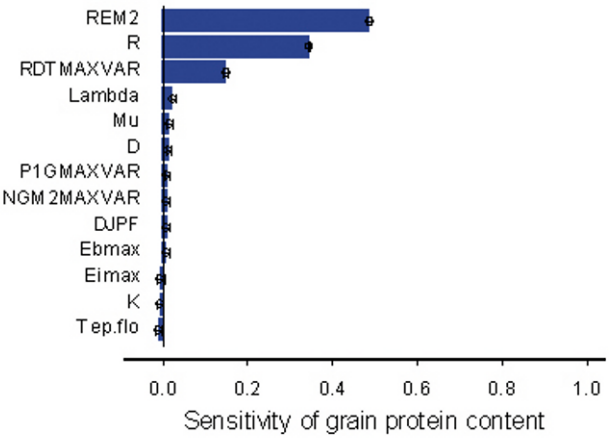

Figure 5.7 – Indices de sensibilité des paramètres d'un modèle de culture pour la teneur en protéines du blé

Chapitre 6

Recommandations

6.1 Quelques conseils

Les méthodes décrites et illustrées dans ce livre permettent de traiter un grand nombre de problèmes pratiques. Nous présentons ci-dessous quelques conseils pour appliquer ces méthodes :

- **Faire des exercices pratiques avec vos données**. Les programmes présentés dans le texte peuvent être facilement adaptés pour analyser d'autres bases de données. Il suffit en général de modifier le nom des variables. Parfois des modifications plus importantes seront cependant nécessaires, par exemple pour estimer les paramètres de certains modèles non linéaires.
- **Explorer et vérifier ses données avant toute analyse statistique**. L'un des grands intérêts de R est de mettre à disposition sous le même environnement, aussi bien de nombreux outils de description et de représentation graphique des données, que des méthodes de modélisation statistique plus élaborées. Il est vivement conseillé d'effectuer d'abord des représentations graphiques de ses données (par exemple des diagrammes de dispersion, des histogrammes, etc.) afin de détecter d'éventuels données aberrantes ou des comportements inattendus.
- **Ne pas se laisser fasciner par les modèles complexes**. La modélisation est un exercice grisant. Certains modélisateurs construisent des versions de plus en en plus complexes de leurs modèles en ajoutant de nouvelles variables ou en complexifiant les équations. Bien souvent, ces modèles deviennent des objets à la fois difficiles à manipuler et très peu prédictifs. Il est important de comparer un modèle complexe avec des versions plus simples en utilisant des critères quantitatifs.
- **Essayer de comprendre les limites des méthodes utilisées**. Les logiciels d'analyse de données permettent d'appliquer facilement un grand nombre de méthodes et de générer de nombreux résultats en très peu de temps. Cette facilité ne doit pas faire oublier que toute méthode a

des limites et qu'une bonne interprétation des résultats nécessite de bien connaître la méthode utilisée.
- **Comparer les méthodes**. Il est rare qu'une seule méthode statistique ou qu'un seul type de modèle puisse être appliqué pour traiter un problème donné. Il est généralement très utile d'utiliser plusieurs approches pour analyser un jeu de données. Chacune de ces approches donnera un éclairage différent au problème considéré.
- **Se former**. La statistique est une discipline qui évolue rapidement. De nouvelles méthodes sont produites régulièrement et il ne faut pas hésiter à se documenter et à suivre des formations.

6.2 Pour continuer

Il existe bien sûr d'autres types de modèles et d'autres méthodes statistiques que ceux présentés dans cet ouvrage. Parmi les sujets que nous avons peu évoqués, deux grandes familles de méthodes statistiques nous paraissent particulièrement utiles dans le cadre de l'analyse des risques agro-environnementaux : l'analyse des séries chronologiques et la statistique spatiale. La première famille de méthodes permet d'analyser les relations entre des données mesurées à des dates différentes. Les techniques d'analyse des séries chronologiques permettent ainsi de mettre en évidence l'existence de tendances et de cycles et de réaliser des prédictions d'événements futurs. Voir par exemple le livre de Brockwell et Davis (2002).

La statistique spatiale a un objectif différent : l'analyse des relations entre données obtenues dans diverses zones géographiques, en prenant en compte les distances et les relations de voisinage entre observations. Les méthodes de statistique spatiale permettent de mettre en évidence des corrélations entre mesures spatialisées et de tenir compte de ces corrélations pour réaliser des prédictions aux endroits où aucune mesure n'a été réalisée. Ces méthodes n'ont été que brièvement évoquées ici et le lecteur intéressé pourra se reporter à des ouvrages spécialisés comme Banerjee *et al.* (2004).

Bibliographie

[1] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 1990.

[2] H. Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60 :255–265, 1973.

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

[4] J.-M. Azaïs and J.-M. Bardet. *Le Modèle Linéaire par l'exemple - Régression, Analyse de la Variance et Plans d'Expérience Illustrés avec R, Sas et Splus*. Dunod, 2006.

[5] R.A. Bailey. *Design of Comparative Experiments*. Cambridge University Press, 2008.

[6] S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, 2004.

[7] A. Barbottin, D. Makowski, M. Le Bail, M.-H. Jeuffroy, Ch. Bouchard, and C. Barrier. Comparison of models and indicators for categorizing soft wheat fields according to their grain protein contents. *European Journal of Agronomy*, 29 :159–183, 2008.

[8] J. Bernier, É. Parent, and J.-J. Boreux. *Statistique pour l'Environnement. Traitement Bayésien de l'Incertitude*. Editions TEC&DOC Lavoisier, 2nd edition, 2000.

[9] J.-J. Boreux, É. Parent, and J. Bernier. *Pratique du Calcul Bayésien*. Collection Statistique et probabilités appliquées. Springer-Verlag France, 2010.

[10] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Springer-Verlag, 2nd edition, 2002.

[11] S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7 :434–455, 1998.

[12] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference*. Springer-Verlag, 2nd edition, 2002.

[13] B.S. Cade, J.W. Terrel, and R.L. Schroeder. Estimating the effects of limiting factors with regression quantiles. *Ecology*, 80 :311–323, 1999.

[14] B.P. Carlin and T.A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, 3rd edition, 2008.

[15] J.M. Chambers. *Software for Data Analysis. Programming with R.* Springer-Verlag, 2008.

[16] B. Chauvel, J.-P. Guillemin, and N. Colbach. Evolution of herbicide-resistant population of *Alopecurus myosuroides Huds.* in a long-term cropping system experiment. *Crop Protection*, 28 :343–349, 2009.

[17] B. Chauvel, J.-P. Guillemin, N. Colbach, and J. Gasquez. Evaluation of cropping systems for management of herbicide-resistant population of blackgrass (*Alopecurus myosuroides Huds.*). *Crop Protection*, 20 :127–137, 2001.

[18] B. Chevassus-au Louis. *L'analyse des Risques. L'Expert, le Décideur et le Citoyen*. Editions Quae, 2007.

[19] Y.-K. Choi, W.O. Johnson, M.T. Collins, and I.A. Gardner. Bayesian inferences for Receiver Operating Characteristic curves in the absence of gold standard. *Journal of Agricultural, Biological, and Environmental Statistics*, 11 :210–229, 2006.

[20] Comifer. Calcul de la fertilization raisonnée. Technical report, Comifer, 1996.

[21] P.-A. Cornillon and E. Matzner-Lober. *Régression avec R.* Springer-Verlag France, 2011.

[22] P. Dagnelie. *Principes d'Expérimentation : Planification des Expériences et Analyse de leurs Résultats*. Presses Agronomiques de Gembloux, 2003.

[23] J.-J. Daudin, S. Robin, and C. Vuillet. *Statistique Inférentielle. Idées, démarche, exemples*. Presses Universitaires de Rennes 2, 1999.

[24] M. Davidian and D.M. Giltinan. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC, 1995.

[25] A. Dean and D. Voss. *Design and Analysis of Experiments*. Springer texts in Statistics. Springer-Verlag, 1999.

[26] J.A. Delgado, M. Shaffer, C. Hu, R.S. Lavado, J.C Wong, P. Joose, X. Li, H. Rimski-Korsakov, R. Follett, W. Colon, and D. Sotomayor. A decade of change in nutrient management : a new nitrogen index. *Journal of Soil and Water Conservation*, 61 :63A–71A, 2006.

[27] J.-B. Denis and H. Monod. Utilisation de tests statistiques. Rapport technique 2007-1, INRA, UR341 Mathématiques et Informatique Appliquées, Jouy-en-Josas, France, 2007.

[28] Y. Dodge and G. Melfi. *Premiers Pas en Simulation*. Collection Statistique et Probabilités Appliquées. Springer-Verlag France, 2008.

[29] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10 :197–208, 2000.

[30] J.-J. Droesbeke, J. Fine, and G. Saporta, editors. *Plans d'expériences. Applications à l'entreprise*. Technip, 1997.

[31] S. Ennaïfar, D. Makowski, J.-M. Meynard, and Ph. Lucas. Evaluation of models to predict take-all incidence on winter wheat as a function of cropping practices, soil, and climate. *European Journal of Plant Pathology*, 118 :127–143, 2007.

[32] D. Farragi and B. Reiser. Estimation of the area under the ROC curve. *Statistics in Medicine*, 21 :3093–3106, 2002.

[33] A. Gelman, Stern H. Carlin, J.B. and, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2003.

[34] W.R. Gilks, S Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996.

[35] V. Ginot, S. Gaba, R. Beaudouin, F. Ariès, and H. Monod. Combined use of local and ANOVA-based global sensitivity analyses for investigation of a stochastic dynamic model : Application to the case study of an individual-based model of a fish population. *Ecological Modelling*, 193 :479–491, 2006.

[36] A.C. Grundy, N.D. Boatman, and R.J. Froud-Williams. Effects of herbicide and nitrogen fertilizer application on grain yield and quality of wheat and barley. *Journal of Agricultural Science (Cambridge)*, 126 :379–385, 1996.

[37] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143 :29–36, 1982.

[38] F.E. Jr. Harrel. *Regression Modeling Strategies*. Springer-Verlag, 2001.

[39] J. Hillier, D. Makowski, and B. Andrieu. Maximum likelihood inference and bootstrap methods for plant organ growth via multi-phase kinetic models and their application to maize. *Annals of Botany*, 96 :137–148, 2005.

[40] S. Huet, A. Bouvier, M-A. Poursat, and E. Jolivet. *Statistical Tools for Nonlinear Regression : a Practical Guide with S-plus and R examples*. Springer Series in Statistics. Springer-Verlag, 2nd edition, 2004.

[41] G. Hughes, N. McRoberts, and F.J. Burnett. Decision-making and diagnosis in disease management. *Plant Pathology*, 48 :147–153, 1999.

[42] M.A. Jauregui and Q. Paris. Spline response functions for direct and carry-over effects involving single nutrient. *Soil Science Society of America Journal*, 49 :140–145, 1985.

[43] J.W. Jones and W.D. Graham. Application of extended and ensemble Kalman filters to soil carbon estimation. In D. Wallach, D. Makowski, and J. Jones, editors, *Working with Dynamic Crop Models*, pages 399–407. Elsevier, 2006.

[44] J.W. Jones, W.D. Graham, D. Wallach, W.M. Bostick, and J. Koo. Estimating soil carbon levels using an ensemble Kalman filter. *Transactions of the ASAE*, 47 :331–339, 2004.

[45] É. Justes, B. Mary, J.-M. Meynard, J.-M. Machet, and L. Thelier-Huche. Determination of a critical nitrogen dilution curve for winter wheat crop. *Annals of Botany*, 74 :397–407, 1994.

[46] R. Koenker and G. Basset. Regression quantiles. *Econometrica*, 46 :33–50, 1978.

[47] R. Koenker and J.A. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94 :1296–1310, 1999.

[48] R. Koenker and B.J. Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71 :265–283, 1996.

[49] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM Probability and Statistics*, 8 :114–131, 2004.

[50] A. Lacroix, N. Beaudoin, and D. Makowski. Agricultural water nonpoint pollution control under uncertainty and climate variability. *Ecological Economics*, 53 :115–127, 2005.

[51] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet, editors. *Le Logiciel R*. Springer-Verlag France, 2011.

[52] M. Le Bail, M.H. Jeuffroy, C. Bouchard, and A. Barbottin. Is it possible to forecast the grain quality and yield of different varieties of winter wheat from Minolta SPAD meter measurements ? *European Journal of Agronomy*, 23 :379–391, 2005.

[53] M. Le Bail and D. Makowski. A model-based approach for optimizing segregation of soft wheat in country elevators. *European Journal of Agronomy*, 21 :171–180, 2004.

[54] M. Lejeune. *Statistique. La Théorie et ses Applications*. Collection Statistique et Probabilités appliquées. Springer-Verlag France, 2010.

[55] B.G. Lindsay. Statistical distances as loss functions in assessing model adequacy. In M.L. Taper and R. L. Subhash, editors, *The Nature of Scientific Evidence*. The University of Chicago Press, 2004.

[56] D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS – a Bayesian modelling framework : concepts, structure, and extensibility. *Statistics and Computing*, 10 :325–337, 2000.

[57] R.D. Magarey, T.B. Sutton, and C.L. Thayer. A simple generic infection model for foliar fungal plant pathogens. *Phytopathology*, 95 :92–100, 2005.

[58] C. Maisonneuve, A. Penaud, B. Auclert, G. Arjaure, and L. Jung. *Sclerotinia sclerotiorum Lib. De Bary* : mise au point d'un outil de décision pour le traitement du colza d'hiver. In *ANPP-Cinquième conférence internationale sur les maladies des plantes*, 1997.

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

[60] D. Makowski, T. Doré, and H. Monod. A new method to analyse relationships between yield components with boundary lines. *Agronomy for Sustainable Development*, 27 :119–128, 2007.

[61] D. Makowski, M. Guérif, J. Jones, and W. Graham. Data assimilation with crop models. In D. Wallach, D. Makowski, and J. Jones, editors, *Working with Dynamic Crop Models*, pages 151–172. Elsevier, 2006.

[62] D. Makowski and M. Lavielle. Using SAEM (Stochastic Approximation of EM) to estimate parameters of models of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics*, 11 :45–60, 2006.

[63] D. Makowski, M. Taverne, J. Bolomier, and M. Ducarne. Comparison of risk indicators for sclerotinia control in oilseed rape. *Crop Protection*, 24 :527–531, 2005.

[64] D. Makowski, M. Tichit, L. Guichard, H. van Keulen, and N. Beaudoin. Measuring the accuracy of agro-environmental indicators. *Journal of Environmental Management*, 90 :S139–S146, 2009.

[65] D. Makowski and D. Wallach. How to improve model-based decision rules for nitrogen fertilization. *European Journal of Agronomy*, 15 :197–208, 2001.

[66] D. Makowski and D. Wallach. It pays to base parameter estimation on a realistic description of model errors. *Agronomy for Sustainable Development*, 22 :179–189, 2002.

[67] D. Makowski, D. Wallach, and J.-M. Meynard. Models of yield, grain protein, and residual mineral nitrogen responses to applied nitrogen of winter wheat. *Agronomy Journal*, 91 :377–385, 1999.

[68] D. Makowski, D. Wallach, and J.-M. Meynard. Statistical methods for predicting responses to applied nitrogen and for calculating optimal nitrogen rates. *Agronomy Journal*, 93 :531–539, 2001.

[69] C.E. McCulloch and S.R. Searle. *Generalized, Linear, and Mixed Models.* John Wiley & Sons, 2001.

[70] J.-M. Meynard, M. Cerf, L. Guichard, M-H. Jeuffroy, and D. Makowski. Which decision support tools for the environmental management of nitrogen ? *Agronomy for Sustainable Development*, 22 :817–829, 2002.

[71] A. Miller. *Subset Selection in Regression.* Chapman & Hall/CRC, 2nd edition, 2002.

[72] T.P. Milsom, S.D. Langton, W.K. Parkin, S. Peel, J.D. Bishop, J.D. Hart, and N.P. Moore. Habitat models of bird species' distribution : an aid to the management of coastal grazing marshes. *Journal of Applied Ecology*, 37 :706–727, 2000.

[73] H. Monod, C. Naud, and D. Makowski. Uncertainty and sensitivity analysis for crop models. In D. Wallach, D. Makowski, and J. Jones, editors, *Working with Dynamic Crop Models*, pages 55–100. Elsevier, 2006.

[74] P.A. Murtaugh. The statistical evaluation of ecological indicators. *Ecological Applications*, 6 :132–139, 1996.

[75] S.W. Myers, I. Fraser, and V.C. Mastro. Evaluation of heat treatment schedules for emerald ash borer (*Coleoptera : Buprestidae*). *Journal of Economic Entomology*, 102 :2048–2055, 2009.

[76] C. Naud, D Makowski, and MH. Jeuffroy. An interacting particle filter to improve model-based predictions of nitrogen nutrition index for winter wheat. *Ecological Modelling*, 207 :251–263, 2007.

[77] É. Parent and J. Bernier. *Le raisonnement bayésien. Modélisation et inférence.* Collection Statistique et probabilités appliquées. Springer-Verlag France, 2007.

[78] A. Pavé. *Modélisation en Biologie et en Écologie.* Aléas, 1994.

[79] M.S. Pepe. Three approaches to regression analysis of Receiver Operating Characteristic curves for continuous test results. *Biometrics*, 54 :124–135, 1998.

[80] M.S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, volume 28 of *Oxford Statistical Science Series.* Oxford University Press, 2003.

[81] J.C. Pinheiro and D.M. Bates. *Mixed-effects Models in S and S-PLUS.* Springer-Verlag, 2000.

[82] S. Primot, M. Valantin-Morison, and D. Makowski. Predicting the risk of weed infestation in winter oilseed rape crops. *Weed Research*, 46 :22–33, 2006.

[83] L. Prost, D. Makowski, and M-H. Jeuffroy. Comparison of stepwise selection and bayesian model averaging for yield gap analysis. *Ecological Modelling*, 219 :66–76, 2008.

[84] A.E. Raftery, D. Madigan, and J.A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92 :179–191, 1997.

[85] C. Robert. *Le Choix Bayésien.* Collection Statistique et Probabilités appliquées. Springer-Verlag France, 2006.

[86] C. Robert and Casella G. *Méthodes de Monte Carlo avec R.* Springer-Verlag France, 2011.

[87] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, 2008.

[88] G. Saporta. *Probabilités, Analyse des Données et Statistique*. Éditions Technip, 2nd edition, 2006.

[89] G.A.F. Seber and C.J. Wild. *Nonlinear Regression*. John Wiley & Sons, 2003.

[90] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. ROCR : visualizing classifier performance in R. *Bioinformatics*, 21 :3940–3941, 2005.

[91] C.D. Stansbury, S.J. McKirdy, A.J. Diggle, and I.T. Riley. Modeling the risk of entry, establishment, spread, containment, and economic impact of *Tilletia indica*, the cause of Karnal Bunt of Wheat, using an Australian context. *Phytopathology*, 92 :321–331, 2002.

[92] J.A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240 :1285–1293, 1988.

[93] M. Taverne, F. Dupeuble, and A. Penaud. Evaluation of a diagnostic test for sclerotinia on oilseed rape at flowering. In *Proceedings of the 11th International Rapesed Congress*, 6-10 July 2003.

[94] M. Tichit, O. Renault, and T. Potter. Grazing regime as a tool to assess positive side effects of livestock farming systems on wading birds. *Livestock Production Science*, 96 :109–117, 2005.

[95] H.M.G. van der Werf. Assessing the impact of pesticides on the environment. *Agriculture, Ecosystems and Environment*, 60 :81–96, 1996.

[96] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, 4th edition, 2002.

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

[98] D. Vose. *Risk Analysis. A Quantitative Guide*. John Wiley & Sons, 2nd edition, 2000.

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

[100] W.L. Winston. *Operations Research, Applications and Algorithms*. Duxbury, 1994.

[101] J.Y. Yang, E.C. Huffman, R. de Jong, V. Kirkwood, K.B. MacDonald, and C.F. Drury. Residual soil nitrogen in soil landscapes of canada as affected by land use practices and agricultural policy scenario. *Land Use Policy*, 24 :89–99, 2007.

[102] Z. Yuan and Y. Yang. Combining linear regression models : when and how ? *Journal of the American Statistical Association*, 100 :1202–1214, 2005.

[103] J. Yuen and G. Hughes. Bayesian analysis of plant disease prediction. *Plant Pathology*, 51 :407–412, 2002.

[104] J. Yuen, E. Twengström, and R. Sigwald. Calibration and verification of risk algorithms using logistic regression. *European Journal of Plant Pathology*, 102 :847–584, 1996.

Index