REVUE DE STATISTIQUE APPLIQUÉE

FRANÇOIS HUSSON

Construire un modèle stochastique à partir d'un modèle déterministe

Revue de statistique appliquée, tome 49, n° 4 (2001), p. 5-27

<http://www.numdam.org/item?id=RSA_2001__49_4_5_0>

Rev. Statistique Appliquée, 2001, **XLIX** (4), 05-27

CONSTRUIRE UN MODÈLE STOCHASTIQUE À PARTIR D'UN MODÈLE DÉTERMINISTE

François Husson

Laboratoire de Mathématiques appliquées, ENSA de Rennes,
65 rue de Saint Brieuc, 35042 Rennes Cedex

RÉSUMÉ

Très souvent en modélisation, les modèles construits sont déterministes dans le sens où ils prédisent toujours les mêmes résultats (*i.e.* les mêmes valeurs pour les variables de sortie) pour un jeu de variables explicatives donné. Pourtant, il est intéressant de connaître la variabilité des variables de sortie. Dans de nombreux cas simples, une méthodologie est disponible pour construire un modèle stochastique, mais ce n'est pas le cas lorsque le modèle déterministe est complexe (ce qui est fréquent en agronomie ou écologie par exemple). Dans cet article, on décrit plusieurs méthodes permettant de construire un modèle stochastique à partir d'un modèle déterministe et on choisit une méthode qui peut s'appliquer à un modèle déterministe complexe. On applique alors cette méthode à un exemple issu de l'agronomie.

Mots-clés : *Modèle non linéaire à effets mixtes, variabilité, modèle stochastique*

ABSTRACT

Very often in modeling, the built models are deterministic because they always predict the same values for the output variables when the input variables are given. However, it is interesting to know the variability of the outputs. In many simple cases, a methodology is available to build a stochastic model, but it is not the case when the deterministic model is complex (which is frequent in agronomy or ecology for example). In this article, one describes several methods allowing to build a stochastic model from a deterministic model and one chooses a method which can be applied to a complex deterministic model. One then applies this method to an example resulting from agronomy.

Keywords : *Non linear mixed model, variability, stochastic model*

0. Introduction

La plupart des modèles agronomiques donnent des résultats déterministes : ils prédisent toujours les mêmes résultats pour un jeu de variables explicatives donné. La connaissance des distributions des prédictions du modèle est pourtant précieuse pour connaître la variabilité de ces résultats et pour calculer par exemple des intervalles de confiance des prédictions. Un modèle déterministe n'est pas suffisant et l'introduction d'aléatoire dans le modèle est nécessaire, d'où la construction de modèle que l'on

appelle modèle stochastique. Un tel modèle donne à la fois la prédiction et la variabilité de ses variables de sortie.

Si nous appliquons notre définition des modèles déterministes et des modèles stochastiques au modèle linéaire, le modèle déterministe est $Y = X\beta$ et le modèle stochastique est $Y = X\beta + \varepsilon$ avec ε une variable aléatoire qui suit, par exemple, une loi normale. Donc lors de la construction de modèle linéaire, on construit directement un modèle stochastique. Par contre, lorsqu'on construit un modèle agronomique, le modèle est déterministe. En effet, les modèles agronomiques sont souvent des modèles très complexes qui sont itératifs (une itération par jour) et qui utilisent de très nombreuses équations. Des variables prédictrices, telles que des variables météorologiques, des variables décrivant l'état du sol ou encore la variété, sont utilisées en entrée par le modèle. Le modèle permet ensuite de prédire plusieurs variables telles que le rendement, la quantité de matière sèche, la date de récolte, etc. Ces modèles agronomiques étant complexes, les calculs (et donc les prédictions) ne peuvent être effectués que par des ordinateurs. On parle alors de simulations (une simulation étant l'ensemble des calculs, de quelques jours avant le semis à la récolte, qui permet de prédire l'ensemble des variables de sortie). Notons de plus que les modèles agronomiques utilisent généralement énormément de paramètres, et donc qu'il est impossible d'estimer tous les paramètres du modèle. Les paramètres sont estimés par des expériences antérieures lors de la construction de sous-modèles qui sont utilisés dans le modèle (par exemple, un sous-modèle permettant de simuler l'absorption d'eau par la plante a été paramétré, puis est inclus dans le modèle de croissance de la plante). Ainsi, lorsque le modèle déterministe est construit, les paramètres du modèle ont été déterminés (nous noterons ces paramètres $\hat{\beta}$) et sont considérés comme fixes (et non comme des variables aléatoires).

Dans cet article, on considère que plusieurs mesures sont effectuées sur un même individu, que ces mesures peuvent être effectuées sur plusieurs variables et qu'elles peuvent être effectuées à des intervalles de temps irréguliers d'un individu à l'autre. Par exemple, si on modélise la croissance d'une plante, on peut avoir 5 mesures de poids puis une mesure du rendement sur un même individu. On définit le modèle déterministe par

$$y = f\left(\beta, X_i^{d\{j\}}\right) \tag{1}$$

où y est un vecteur correspondant à toutes les prédictions disponibles dans le modèle (par exemple : plusieurs quantités de matière sèche, la date de récolte, le rendement, ...) et $X_i^{d\{j\}}$ est le vecteur des prédicteurs du $i^{\text{ème}}$ individu (ou traitement) correspondant à l'ensemble des variables explicatives utilisées par le modèle jusqu'à la date $d\{j\}$ du prélèvement de la mesure j, f est une fonction (linéaire ou non linéaire) des paramètres β et du vecteur des prédicteurs $X_i^{d\{j\}}$. On considère que le modèle déterministe a été construit et paramétré et donc que des expériences antérieures ont permis d'estimer les paramètres β (par $\hat{\beta}$). Les résidus du modèle déterministe sont donc calculables :

$$\hat{e}_i^{\{j\}} = y_i^{\{j\}} - f\left(\hat{\beta}, X_i^{d\{j\}}\right)$$

où $y_i^{\{j\}}$ est la $j^{\text{ème}}$ mesure (observations) sur le $i^{\text{ème}}$ individu, et $\hat{e}_i^{\{j\}}$ le résidu du modèle déterministe correspondant à cette mesure.

C'est la distribution de y qui nous intéresse mais pour connaître cette distribution, il suffit de connaître la distribution des résidus. Notre stratégie consiste à considérer que le modèle déterministe explique déjà une part importante de la variabilité des observations. On s'intéresse alors à la variabilité des résidus du modèle sans distinguer si cette incertitude provient d'une mauvaise connaissance des variables d'entrée, d'une mauvaise estimation des paramètres, de certaines équations du modèle qui ne sont pas tout à fait justes, d'erreurs d'échantillonnage ou d'erreurs de mesures.

Si les résidus du modèle déterministe sont indépendants et identiquement distribués, les prédictions du modèle stochastique se calculent simplement en ajoutant une variable aléatoire f_i au modèle déterministe :

$$g\left(\hat{\beta}, X_i^{d\{j\}}, \xi, \sigma^2\right) = f\left(\hat{\beta}, X_i^{d\{j\}}\right) + f_i^{\{j\}} \quad \text{avec} \quad f_i \sim \mathcal{N}(\xi, \sigma^2 I_{n_i}) \qquad (2)$$

avec σ^2 la variance des résidus (la loi des f_i est très souvent Normale).

Cependant, si les résidus du modèle déterministe ne sont pas indépendants, cette modélisation ne convient pas. Or, si plusieurs mesures sont effectuées sur un même individu, les résidus risquent d'être autocorrélés. Ils sont même presque systématiquement autocorrélés lorsqu'une même variable est mesurée à différents instants.

Pour montrer l'importance de l'étude présentée dans cet article, on peut souligner que l'analyse de données à mesures répétées est un problème très fréquent en statistique et trouve de nombreuses applications dans des disciplines très variées comme la pharmacologie ou l'agronomie (des concentrations de médicament ou des poids de plante sont des mesures répétées dans le temps).

Des modélisations stochastiques (que nous détaillerons par la suite) sont disponibles dans de nombreux cas simples (résidus du modèle déterministe indépendants, modèle linéaire, modèle non linéaire simple) mais si le modèle déterministe est un modèle non linéaire complexe, comme les modèles dynamiques que l'on rencontre fréquemment en agronomie ou écologie par exemple, alors on ne sait pas comment rendre stochastique le modèle. L'objectif de cet article est de présenter puis de comparer différentes méthodes permettant de construire un modèle stochastique à partir de modèles déterministes linéaires, non linéaires, statiques, dynamiques, etc. Pour cela, nous évaluerons différentes modélisations disponibles pour les cas simples et nous verrons si elles peuvent être adaptées à des modèles complexes.

Le choix de la modélisation dépend essentiellement du modèle déterministe lui-même et de ses résidus. Dans un premier temps, nous verrons comment tester l'indépendance, le non biais, la multinormalité des résidus. Nous décrirons ensuite quelques modèles stochastiques et nous verrons sous quelles conditions ils peuvent être utilisés. En particulier, on s'intéressera aux modèles non linéaires à effets mixtes et on décrira alors deux méthodes qui permettent d'estimer les paramètres de tels modèles. Enfin, on construira en guise d'exemple un modèle stochastique de croissance du colza.

1. Étude des résidus du modèle déterministe

Afin de choisir la méthode la plus adaptée pour construire un modèle stochastique, on étudie le comportement du modèle déterministe et celui de ses résidus : le modèle est-il biaisé, les résidus pour un même individu sont-ils indépendants ou autocorrélés, les résidus suivent-ils une loi multinormale, les erreurs d'échantillonnage sont-elles relativement importantes et expliquent-elles, à elles seules, les résidus du modèle déterministe ?

1.1. Étude du biais du modèle

L'étude du biais d'un modèle est simple : il suffit de tester, variable par variable, si la moyenne des résidus est égale à 0. Pour cela on utilise des tests de conformité (test t). Pour des variables continues (telle que la quantité de matière sèche), on est amené à réaliser autant de tests qu'il y a de dates de mesure. Si toutes les variables sont sans biais, alors le modèle sera dit sans biais. Notons qu'il est toujours possible de se ramener à un modèle sans biais par translation.

1.2. Étude de l'indépendance des résidus

L'indépendance des résidus est une hypothèse plus intéressante car c'est une condition indispensable dans de nombreux tests. Or, si plusieurs mesures sont réalisées sur un même individu, les résidus du modèle déterministe forment une série chronologique et il est intéressant de tester si c'est un bruit blanc ou non. La statistique de Durbin-Watson ou le test de portemanteau (Seber et Wild, 1989, p.322) permettent de tester l'hypothèse que les coefficients de corrélation entre observations successives sont nuls si les observations sont réalisées à intervalles de temps réguliers.

Cependant, bien souvent, il est difficile de faire beaucoup de mesures et des mesures à intervalles de temps réguliers (en raison de contraintes climatiques, de prélèvements destructifs,...). Les tests classiques d'autocorrélations ne peuvent pas être utilisés et on doit recourir à des tests non paramétriques. On va construire un test, certes moins puissant que les tests d'autocorrélations classiques, mais qui permet de tester l'indépendance d'observations quelle que soit la fréquence des mesures.

Pour que les résidus soient indépendants, il est nécessaire que les signes des résidus, classés par ordre chronologique, soient indépendants (cette condition n'est pas suffisante). On va alors tester l'hypothèse «*H : les occurrences des signes des résidus sont indépendantes (à l'ordre 1)*» . Pour cela, il est nécessaire que les résidus soient centrés. On se base alors sur le test de Wald-Wolfowitz qui permet de tester que, dans un échantillon de «A» et «B», les lettres sont distribuées au hasard (Tomassone *et al.*, 1993, p.220). Cependant, ce test permet uniquement de tester l'hypothèse individu par individu, donc on est amené à le modifier pour avoir un test global en trois étapes :

1. on calcule, individu par individu, le nombre de changements de signes observé et le nombre de changements de signes maximal qui aurait pu être observé (=le nombre d'observations moins 1);

2. on somme sur tous les individus pour obtenir le nombre de changements de signes global ($nobs$) et le nombre de changements de signes maximal global ($nmax$);
3. on calcule alors la probabilité d'observer un nombre de changements de signes inférieur ou égal à $nobs$ si l'hypothèse H est vraie et de cette probabilité, on conclut.

Soit t la variable aléatoire correspondant au nombre total de changements de signe; on note $P_H(t \leqslant nobs)$ la probabilité d'observer un nombre de changements de signes inférieur ou égal à $nobs$ sous l'hypothèse H :

$$P_H(t \leqslant nobs) = \sum_{r=0}^{nobs} P_H(t = r).$$

$P_H(t = r)$ est la probabilité d'avoir exactement r changements de signes. Les résidus étant indépendants d'un individu à l'autre, $P_H(t = r)$ ne dépend que de $nmax$ et pas du nombre d'individus. Par conséquent, pour calculer $P_H(t = r)$, on peut considérer qu'on a un seul individu avec $1 + nmax$ résidus (et donc $nmax$ changements de signes au plus), d'où :

$$P_H(t \leqslant nobs) = \sum_{r=0}^{nobs} C_{nmax}^{r}(1/2)^{nmax}.$$

De cette probabilité, on déduit l'indépendance ou non des signes des résidus, et par suite des résidus du modèle déterministe. Comparé aux tests classiques d'autocorrélations, ce test est moins puissant car il ne tient pas compte de l'amplitude des écarts entre deux valeurs successives et donc n'utilise qu'une partie de l'information présente dans les observations. Cependant, en pratique, ce test est très souvent suffisant pour mettre en évidence les corrélations entre résidus successifs.

Notons que si $P_H(t \leqslant nobs)$ est faible, alors les résidus pour un même individu ont tendance à être du même signe. Cela signifie que le modèle déterministe à tendance à surestimer (ou sous-estimer) systématiquement les observations d'un même individu.

1. 3. Étude de la multinormalité des résidus

Pour déterminer si la distribution des résidus du modèle déterministe est multinormale, on peut utiliser le test de multinormalité proposé par Mardia (1974). Ce test est fondé sur les statistiques du coefficient d'asymétrie et du coefficient d'aplatissement multivarié. Ce test de multinormalité est disponible dans le logiciel SAS (version 6.12).

1.4. Distinction des erreurs d'échantillonnage et des erreurs de prédiction du modèle

Si des répétitions sont disponibles pour un même traitement, il est intéressant de distinguer les erreurs de prédiction du modèle des erreurs d'échantillonnage dans le calcul des résidus du modèle. En effet, la variabilité des résidus peut se décomposer en deux termes : la variance des erreurs d'échantillonnage et la variance des erreurs de prédiction du modèle. Si la variance des erreurs de prédiction du modèle est faible comparée à la variance des erreurs d'échantillonnage, alors le modèle déterministe est bon et le modèle stochastique peut être construit avec l'équation (2).

La présence de répétitions permet également de tester si une combinaison des variables explicatives (ce qu'on appelle traitement) a un effet significatif sur les résidus. Des résultats de ce test dépend la modélisation stochastique que l'on choisira de mettre en place.

2. Modèles stochastiques

On va présenter quelques modèles stochastiques après les avoir séparés en deux classes :

1. Les modèles avec une variabilité additive sur les variables de sortie

$$y = \phi(\beta, X) + e \quad \text{avec} \quad e \sim \mathcal{N}(\xi, \Sigma).$$

2. Les modèles dynamiques stochastiques

$$\frac{dy}{dt} = \phi(y, \beta, X) + e \quad \text{avec} \quad e \sim \mathcal{N}(\xi, \Sigma).$$

Lors de la présentation de chacun des modèles, on précisera sous quelles conditions ils sont utilisables et quelles sont les difficultés rencontrées pour estimer les paramètres d'un tel modèle lorsque le modèle déterministe (*i.e.* la fonction ϕ) est complexe.

2.1. Modèles stochastiques additifs

Les modèles stochastiques additifs se construisent en ajoutant aux prédictions du modèle déterministe un terme d'erreur. C'est ce terme d'erreur qui rend le modèle stochastique. Plusieurs modélisations de la variabilité et des corrélations des résidus des variables de sortie du modèle sont envisageables.

2.1.1. Modèle de base, ξ et Σ quelconques

Avant d'étudier des modèles stochastiques complexes, on peut adapter le modèle standard (2) à notre étude. On note :

$$y_i = f(\hat{\beta}, X_i) + \hat{e}_i = \hat{y}_i + \hat{e}_i$$

où y_i est le n_i-vecteur des observations, $f(\hat{\beta}, X_i)$ est le n_i-vecteur des prédictions du modèle (ce n_i-vecteur des prédictions peut aussi se noter $\hat{y}_i$), $\hat{\beta}$ est le vecteur estimé des paramètres du modèle, X_i est la matrice des variables prédictrices et $\hat{e}_i$ est le n_i-vecteur des résidus du modèle pour le traitement i. On a ici $\hat{\beta}$ et non β car comme cela a été souligné en introduction, le modèle utilise les estimations $\hat{\beta}$ des paramètres β obtenues par des expériences antérieures. Le biais du modèle s'écrit $\xi_i = E(\hat{e}_i) = E(y_i - f(\hat{\beta}, X_i))$. Si les vecteurs $\hat{e}_i$ sont indépendants et suivent une loi multinormale, on a $\hat{e}_i \sim \mathcal{N}(\xi_i, \Sigma_i)$ où ξ_i et Σ_i dépendent de i uniquement par leur dimension. Cela signifie qu'on peut écrire $\xi_i = D_i'\xi$ et $\Sigma_i = D_i'\Sigma D_i$ où ξ est un vecteur $(N \times 1)$ et Σ est une matrice $(N \times N)$ de variances-covariances quelconque (N étant le nombre de mesures possible) qui n'est diagonale que dans le cas particulier où les composantes du vecteur $\hat{e}_i$ sont indépendantes (*i.e.* si les résidus du modèle déterministe sont indépendants d'une mesure à l'autre) et D_i est une matrice $(N \times n_i)$ composée de 1 et de 0. Les matrices D_i sont particulières car on trouve une et une seule composante qui vaut 1 par colonne et toutes les autres composantes valent 0. Les n_i composantes qui valent 1 ont pour numéro de ligne les numéros des mesures qui sont effectuées sur l'individu i. Par exemple si, lors d'une expérience, $N = 5$ mesures peuvent être effectuées sur des individus et que sur l'individu i_0, les mesures 1, 3 et 5 sont manquantes, on écrira :

$$\xi_{i_0} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \xi = \begin{bmatrix} \xi_2 \\ \xi_4 \end{bmatrix},$$

et

$$\Sigma_{i_0} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \Sigma \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \sigma_{22}^2 & \sigma_{24} \\ \sigma_{24} & \sigma_{44}^2 \end{bmatrix}.$$

Chaque composante de ξ et Σ peut être estimée, respectivement, par la moyenne empirique et par les variances ou covariances empiriques des résidus du modèle. Cependant, les estimations de toutes les composantes de ξ et Σ sont souvent difficiles à obtenir car il est très fréquent que des données soient manquantes et que l'on effectue un nombre n_i de mesures sur le traitement i qui est très inférieur à N. Par exemple, pour des mesures de poids journalières, N correspond au nombre de jours pendant la culture et comme ces mesures sont coûteuses et destructrices, elles sont effectuées beaucoup moins souvent.

Cette modélisation convient aux problèmes ayant peu de variables de sortie et où les sorties sont des variables à temps fixe, *i.e.* mesurée une seule fois (par exemple, le rendement, la date de levée). En effet, si des variables sont à temps discret (*i.e.* mesurées plusieurs fois dans l'étude comme par exemple le poids), alors le nombre de sortie du modèle est important, ce qui augmente la taille de ξ et Σ. Par exemple, si une variable est observée à 200 instants, le vecteur ξ sera de taille 200×1 et la matrice Σ de taille 200×200 d'où l'estimation de 20300 paramètres indépendants. Le modèle stochastique utilise alors beaucoup de paramètres ce qui rend les estimations peu stables et la modélisation mauvaise d'un point de vue prédictif.

2.1.2. Modélisation de ξ et Σ

Si on veut décrire des variables à temps discret, les estimations empiriques des composantes de ξ et Σ sont très difficiles à obtenir. On peut alors modéliser le vecteur des espérances et la matrice de covariances des résidus afin de diminuer le nombre de paramètres à estimer. Faivre *et al.* (1991) proposent des formes fonctionnelles pour les espérances, variances et covariances des résidus. Par exemple, si les espérances sont linéaires en fonction du temps (de j) et si les covariances sont linéaires en $j - j'$, on peut proposer les formes suivantes :

$$\forall j : E(\hat{e}_j) = \theta_1 + \theta_2 \times j, \;\; V(\hat{e}_j) = \theta_3 + \theta_4$$
$$\text{et } Cov(\hat{e}_j, \hat{e}_{j'}) = \theta_4 + \theta_5 \times (j' - j) \quad (\forall j' > j).$$

θ_4 est un paramètre commun à la variance et à la covariance et représente (en pratique) la variance associée à l'erreur de mesure. Le nombre de paramètres à estimer est considérablement réduit par rapport à la méthode précédemment décrite, mais ce nombre reste tout de même très important si on veut modéliser plusieurs variables à temps discret. On peut noter que l'estimation des paramètres θ n'est pas simple car elle doit être effectuée sous la contrainte de positivité de la matrice de covariances $V(\hat{e}_i)$. On peut noter que des modèles de type ARMA ou ARIMA permettent de modéliser des variables continues et peuvent être utilisés ici sur les résidus.

Cette méthode convient aux problèmes dans lesquels les observations se résument à une seule variable à temps discret et éventuellement quelques variables à temps fixe.

Cette modélisation a un défaut majeur qui est commun à toutes les modélisations de cette classe : les modèles stochastiques ainsi construits génèrent de la variabilité sur les prédictions des variables pour lesquelles des observations sont disponibles et à partir desquelles ξ et Σ ont été estimés. Par contre, pour les autres variables de sortie du modèle, aucune variabilité n'est modélisée.

2.2. Modèle autorégressif d'ordre 1

Afin de rendre aléatoires toutes les variables (intermédiaires et de sortie) du modèle, le terme de variabilité peut être inséré directement dans les équations du modèle plutôt qu'ajouté aux résultats déterministes de celui-ci.

2.2.1. *Modèle autorégressif additif*

Un modèle dynamique stochastique se définit par :

$$\frac{dy}{dj} = \phi(y, \beta, X, j) + e^{\{j\}}$$

où $\frac{dy}{dj}$ est le vecteur des accroissements de y dans l'intervalle de temps dj, $\phi(y, \beta, X, j)$ le vecteur des prédictions de l'accroissement du modèle déterministe et $e^{\{j\}}$ l'erreur au temps j. Ce modèle peut s'écrire comme un modèle autorégressif fonctionnel d'ordre 1. Si on écrit ce modèle pour le traitement i, on a :

$$y_i^{\{j\}} = f\left(y_i^{\{j-1\}}\right) + e_i^{\{j\}},$$

ou bien en faisant apparaître un terme d'accroissement :

$$y_i^{\{j\}} = y_i^{\{j-1\}} + \Delta_i^{\{j\}} + e_i^{\{j\}}$$

où $\Delta_i^{\{j\}}$ est le p-vecteur des accroissements (p correspond au nombre de variables observées et donc au nombre d'équations rendues aléatoires) entre l'instant $j-1$ et l'instant j (fonction non linéaire de $y_i^{\{j-1\}}$) calculé par le modèle déterministe et $e_i^{\{j\}}$ le p-vecteur des aléas à l'instant j sur le traitement i de loi $\mathcal{N}(\xi, \Sigma)$.

Cette méthode permet de modéliser des variables continues pour lesquelles la variabilité du résidu est du même ordre de grandeur à chaque instant de la simulation, et donc pour lesquelles la variabilité est indépendante de l'accroissement. Dans le cas contraire, les aléas risquent d'être le moteur du modèle en début de culture (dans le sens où les aléas auront le plus d'influence sur les sorties du modèle) et les aléas n'auront pas d'effet sur les résultats du modèle et la variabilité du modèle stochastique sera proche de 0 en fin de culture. Typiquement, un modèle de type exponentiel (très utilisé pour modéliser la croissance des plantes : par exemple, le poids d'une plante en début de culture est une fonction exponentielle du nombre de jours depuis le semis) ne pourra pas être rendu stochastique par cette méthode.

Dans la définition d'un modèle autorégressif d'ordre 1, les aléas sont des bruits blancs, mais on peut généraliser ces modèles en utilisant une des trois hypothèses suivantes :

- les $e_i^{\{j\}}$ sont des p-vecteurs aléatoires indépendants (cas des bruits blancs),
- $e_i^{\{j\}} = \alpha_i$ où α_i est un p-vecteur aléatoire,
- $e_i^{\{j\}} = \alpha_{ik}$ pour $j_{k-1} < j < j_k$ où les α_{ik} sont des p-vecteurs aléatoires indépendants deux à deux ($\forall i, k$).

La première hypothèse signifie qu'une valeur différente de l'aléa est ajoutée à chaque instant. La deuxième hypothèse qu'un même aléa est calculé et ajouté à chaque instant de la simulation. Ces deux hypothèses étant très radicales, Seber et Wild (1989) ont proposé la troisième hypothèse qui est intermédiaire. La simulation est découpée en plusieurs phases (m). Au début de chaque phase, une valeur de l'aléa est choisie

puis ajoutée chaque instant à la prédiction déterministe. L'hypothèse de Seber et Wild (1989) est très générale puisque la première hypothèse proposée correspond à *m = le nombre de fois où l'équation est utilisée* tandis que la seconde correspond à $m = 1$. La distribution des α_{ik} est normale de moyenne ξ et de matrice de covariance Σ; ξ et Σ peuvent être estimés, à partir des données, par moindres carrés.

White et Brisbin (1980) ont montré que, même si les α_{ik} sont indépendants et changent à chaque instant, les sorties d'un modèle stochastique ainsi construit sont autocorrélées.

2.2.2. *Modèle autorégressif multiplicatif*

Plutôt que d'ajouter un aléa aux équations du modèle, on peut multiplier l'accroissement par un aléa. Ainsi, l'équation aléatoire du modèle s'écrira à l'instant j :

$$y_i^{\{j\}} = y_i^{\{j-1\}} + \Delta_i^{\{j\}} \times e_i^{\{j\}} \quad \text{avec} \quad e_i^{\{j\}} \sim \mathcal{N}(\xi, \Sigma)$$

$y_i^{\{j\}}$, $y_i^{\{j-1\}}$ et $e_i^{\{j\}}$ sont des p-vecteurs et $\Delta_i^{\{j\}}$ est une matrice diagonale $p \times p$ (p étant le nombre de variables observées). Les espérances et variances de e_i, ξ et Σ de dimension $p \times 1$ et $p \times p$, sont à nouveau les paramètres à estimer à partir des données. Cette modélisation revient à ajouter, à chaque équation, des aléas ayant des variances qui varient dans le temps et qui dépendent de l'accroissement :

$$y_i^{\{j\}} = y_i^{\{j-1\}} + \mu_i^{\{j\}}$$

avec

$$\mu_i^{\{j\}} \sim N\left(\Delta_i \, \xi, \Delta_i \, \Sigma \, \Delta_i'\right).$$

Ainsi, la variabilité créée à l'instant j dépend de la prédiction du modèle et le rapport de l'écart-type de chaque composante de l'aléa sur l'accroissement reste constant pour tout j. De tels modèles stochastiques créent une variabilité proportionnelle aux résultats du modèle déterministe. C'est une situation particulièrement fréquente pour les modèles de croissance de plante.

Comme pour le modèle additif, on utilise le modèle de Seber et Wild avec $e_i^{\{j\}} = \alpha_{ik}$, pour $j_{k-1} < j < j_k$, les p-vecteurs α_{ik} étant indépendants de moyenne ξ et de variance Σ. Les estimations de ξ et Σ peuvent être calculées comme pour le modèle additif.

Ces deux modélisations autorégressives (additive et multiplicative) posent cependant des problèmes d'ordre calculatoire : si le nombre d'observations n'est pas le même d'un traitement à l'autre et si les observations ne sont pas réalisées aux mêmes instants d'un traitement à l'autre, comment déterminer le nombre de phases m et les différentes phases ? Par contre, si les phases peuvent être facilement définies, les deux méthodes décrites dans cette section permettent de modéliser la variabilité de variables continues. De plus, la variabilité étant ajoutée directement dans les équations du modèle, toutes les sorties du modèle deviennent aléatoires, qu'elles soient à temps discret ou à temps fixe.

3. Modèle à paramètres aléatoires - Modèle mixte

Plutôt que d'ajouter un aléa aux résultats de certaines équations du modèle, il est toujours possible de considérer que ce sont les paramètres du modèle qui sont des variables aléatoires. Tous les paramètres d'un modèle déterministe sont constants, mais on peut considérer que certains paramètres suivent une loi de probabilité et qu'avant toute simulation (*i.e.* toute exécution du modèle sur ordinateur), une valeur du paramètre est choisie au hasard dans sa loi de probabilité. Le paramètre reste fixe pendant toute une simulation mais sa valeur est différente d'une simulation à l'autre. Ces modèles sont fréquemment utilisés en économétrie et sont appelés « modèles à paramètres aléatoires ». Cela revient à considérer que l'effet d'un traitement n'est pas totalement pris en compte dans le modèle, soit à cause de variables explicatives qui ne sont pas utilisées par le modèle, soit plutôt parce que dans certaines équations il y a un effet individuel non pris en compte : un individu se développe plus (ou moins) rapidement qu'attendu à un certain stade de croissance à cause d'aptitudes individuelles de la plante ou d'un micro-environnement favorable (zone plus fertile, micro-climat,...). Ces particularités individuelles peuvent se traduire par un changement de paramètre de croissance spécifique à chaque individu; le paramètre β_i propre à l'individu i s'écrit

$$\beta_i = \beta + b_i$$

où β est un effet commun à tous les traitements et b_i est un effet individuel avec $b_i \sim \mathcal{N}(0, D)$. Ceci nous amène à introduire les modèles mixtes que nous pourrions appeler modèles à effets individuels. Chaque paramètre individuel ne nous intéresse pas en tant que tel mais c'est son existence et sa distribution qui nous intéresse.

Définition d'un modèle non linéaire mixte

Le vecteur des paramètres β_i peut varier d'un individu à l'autre (notons que la dimension de ce vecteur est indépendante de i puisqu'il faut toujours connaître les valeurs de tous les paramètres du modèle pour pouvoir utiliser le modèle) :

$$y_i^{\{j\}} = f\left(\beta_i, X_i^{d\{j\}}\right) + e_i^{\{j\}} \quad \text{avec} \quad \beta_i = A_i\beta + B_i b_i, \quad b_i \sim \mathcal{N}(0, D) \quad (3)$$

où β est un p-vecteur des paramètres à effets fixes, b_i est un q-vecteur des effets aléatoires associés au traitement i, les matrices A_i et B_i de tailles respectives $r \times p$ et $r \times q$ (avec $r \leqslant p$ et $r \leqslant q$) sont des matrices d'effets fixes et d'effets aléatoires, et D est une matrice de covariances de dimension $q \times q$. Les modèles mixtes permettent à la fois des corrélations non nulles et des variances différentes mais supposent toujours la normalité des paramètres.

Davidian et Giltinan (1995) indiquent que deux grandes classes de procédures ont été proposées pour estimer les paramètres de tels modèles qu'ils nomment « modèles non linéaires hiérarchiques ». Si le nombre de mesures par individu est trop faible et ne permet pas d'estimer les paramètres de la régression spécifique à chaque individu, les méthodes proposées sont toutes basées sur la linéarisation du modèle, et la méthode proposée par Lindstrom et Bates (1990) est tout à fait adéquate. Si, par

contre, suffisamment de données sont disponibles, ce qui est souvent le cas pour les études de croissance, il est préférable de travailler individu par individu. Davidian et Giltinan (1993 et 1995) proposent une méthode pour estimer les paramètres du modèle stochastique dans ce cas.

3.1. Méthode proposée par Lindstrom et Bates

Entre 1980 et 1985, de nombreux auteurs (Sheiner et Beal, 1980; Berkey, 1982; Racine-Poon, 1985; Steimer *et al.*, 1985) ont étudié les modèles non linéaires à effets aléatoires pour décrire la fonction de réponse moyenne ainsi que la variabilité intra et inter-individus. En 1990, Lindstrom et Bates proposent un modèle général non linéaire mixte pour des données à mesures répétées :

$$y_i = f_i(\beta_i) + e_i, \qquad \beta_i = A_i\beta + B_i b_i \tag{4}$$

avec

$$y_i = \begin{bmatrix} y_i^1 \\ y_i^2 \\ \vdots \\ y_i^{n_i} \end{bmatrix}, \quad e_i = \begin{bmatrix} e_i^1 \\ e_i^2 \\ \vdots \\ e_i^{n_i} \end{bmatrix}, \quad f_i(\beta_i) = \begin{bmatrix} f\left(\beta_i, X_i^{d\{1\}}\right) \\ f\left(\beta_i, X_i^{d\{2\}}\right) \\ \vdots \\ f\left(\beta_i, X_i^{d\{n_i\}}\right) \end{bmatrix},$$

et où $e_i \mid b_i \sim \mathcal{N}(0, R_i)$ avec R_i une matrice qui dépend de i uniquement par sa dimension.

Pour des modèles non linéaires, il est très souvent difficile de calculer la vraisemblance de la distribution marginale de y_i car la fonction f n'est pas linéaire en b_i. Lindstrom et Bates (1990) indiquent que l'approximation de (4) qui consiste à supposer que les espérances des effets aléatoires b_i sont nulles peut être mauvaise. Ils préfèrent linéariser le modèle (4) puis maximiser la vraisemblance du modèle linéaire approché. Pour cela, ils approchent $y_i - f_i(A_i\beta + B_i b_i)$ au voisinage de $\hat{b}_i$ par

$$y_i - f_i(A_i\beta + B_i b_i) \approx y_i - [f_i(A_i\beta + B_i\hat{b}_i) + \hat{Z}_i b_i - \hat{Z}_i\hat{b}_i]$$

où

$$\hat{Z}_i = \hat{Z}_i(\theta) = \left.\frac{\partial f_i}{\partial b_i'}\right|_{\hat{\beta},\hat{b}_i}.$$

$\hat{Z}_i$ est une fonction de θ (où θ est un vecteur contenant les éléments des matrices R_i et de la matrice D) car $\hat{\beta}$ et $\hat{b}_i$ dépendent de θ. Alors :

$$y_i - f_i(A_i\beta + B_i\hat{b}_i) + \hat{Z}_i\hat{b}_i - \hat{Z}_i b_i \mid b_i \sim \mathcal{N}(0, R_i)$$

et l'approximation de la distribution conditionnelle de y_i s'écrit

$$y_i \mid b_i \sim \mathcal{N}(f_i(A_i\beta + B_i\hat{b}_i) - \hat{Z}_i\hat{b}_i + \hat{Z}_i b_i, R_i).$$

Ainsi, la distribution marginale de y_i peut être approchée par

$$y_i \sim \mathcal{N}(f_i(A_i\beta + B_i\hat{b}_i) - \hat{Z}_i\hat{b}_i, \hat{V}_i)$$

où $\hat{V}_i = \hat{V}_i(\hat{\theta}) = R_i + \hat{Z}_i D \hat{Z}_i'$.

La distribution de y_i est approchée par une distribution multinormale, il est donc possible de calculer la vraisemblance de cette distribution approchée et d'estimer β et θ par maximum de vraisemblance ou par maximum de vraisemblance restreint.

Lindstrom et Bates proposent un algorithme itératif en deux étapes pour trouver ces estimateurs du maximum de vraisemblance :

1. étape des pseudo-données : à partir des estimations $\hat{\theta}$ de θ, minimiser en β et b_i, $i = 1, \ldots, I$ (où I est le nombre de traitements),

$$\sum_{i=1}^{I} \Big(\log|\hat{D}| + b_i' \hat{D}^{-1} b_i + \log|\hat{R}_i| + (y_i - f_i(A_i\beta + B_i b_i))' \hat{R}_i^{-1} (y_i - f_i(A_i\beta + B_i b_i)) \Big).$$

   Les résultats de ces estimations sont notés $\hat{b}_i$ et $\hat{\beta}_0$;

2. étape des effets linéaires mixtes : calculer les estimations $\hat{\beta}$ et $\hat{\theta}$ de β et θ qui minimisent

$$ML(\beta, \theta) = \sum_{i=1}^{I} \Big(\log|\hat{V}_i| + (y_i - f_i(A_i\beta + B_i\hat{b}_i) + \hat{Z}_i\hat{b}_i)' \hat{V}_i^{-1} (y_i - f_i(A_i\beta + B_i\hat{b}_i) + \hat{Z}_i\hat{b}_i) \Big)$$

   où $\hat{V}_i$ et $\hat{Z}_i$ dépendent de $\hat{\beta}_0$ et $\hat{b}_i$, les estimations de l'étape 1.

L'algorithme consiste à itérer ces deux étapes jusqu'à convergence.

La méthode de Lindstrom et Bates est très séduisante d'un point de vue théorique puisque l'estimation est effectuée de manière globale et maximise directement une vraisemblance approchée sur tous les individus. Des critères (comme celui d'Akaike) fondés sur la vraisemblance approchée peuvent être utilisés pour choisir entre plusieurs modèles. En effet, il suffit de calculer la vraisemblance approchée des modèles en compétition, évaluée aux valeurs de convergence des paramètres. Sous SAS, un programme construit et estime les paramètres de tels modèles non linéaires mixtes. En revanche, en pratique, la méthode de Lindstrom et Bates nécessite de trouver les développements de Taylor du modèle. Aussi, dès que le modèle devient complexe, des problèmes de stabilité et de précision dans les calculs des dérivés numériques du modèle rendent difficile la linéarisation. Pour des modèles aussi complexes qu'un modèle agronomique, cette méthode est très difficile à utiliser.

3.2. Méthode proposée par Davidian et Giltinan

Davidian et Giltinan (1995) définissent le modèle non linéaire mixte exactement comme en (3). Par rapport à Lindstrom et Bates (1988 et 1990), ils considèrent que pour tout i, $B_i = Id_q$ (donc avec les notations associées à la formule (3), $r = q$), Id_q étant la matrice identité d'ordre q et que :

$$E\left(e_i^{\{j\}}|\beta_i\right) = 0 \ , \qquad Cov\left(e_i^{\{j\}}|\beta_i\right) = R_i(\beta_i).$$

Les R_i peuvent s'écrire $R_i(\beta_i) = \sigma^2 S_i(\beta_i, \gamma)$ où S_i est une forme fonctionnelle commune à tous les traitements (et qui dépend de i uniquement pour sa dimension) et γ est un vecteur de paramètres (éventuellement réduit à un scalaire) des S_i. Les structures de variance S_i sont, dans beaucoup d'applications, fonction de la moyenne des $f\left(\beta_i, X_i^{d\{j\}}\right)$ (Davidian et Giltinan, 1993). Beal et Sheiner (1988) proposent de prendre $S_i(\beta_i, \gamma) = \left(f\left(\beta_i, X_i^{d\{j\}}\right)\right)^\gamma$ (ici, γ est un scalaire) lorsqu'il y a une seule variable de sortie. Carroll et Ruppert (1988, chapitre 3) discutent plus généralement du choix des structures de variance.

La méthode proposée par Davidian et Giltinan (1993 et 1995) revient à trouver, pour chaque traitement, une estimation β_i^* des paramètres de la régression, puis à calculer les estimateurs de la moyenne et de la matrice de covariances de ces paramètres estimés (β_i^*), par une méthode qu'ils appellent «Global two Stage» .

Plusieurs méthodes sont envisageables pour estimer les β_i à partir du traitement i uniquement. Davidian et Giltinan (1995, chap 2) montrent qu'il est préférable d'utiliser les méthodes des moindres carrés généralisés plutôt que le maximum de vraisemblance en raison de leur meilleure robustesse aux mauvaises prédictions du modèle. L'algorithme suivant permet donc d'estimer les paramètres, traitement par traitement :

1. à partir de I régressions non pondérées indépendantes (par exemple par moindres carrés ordinaires), trouver les estimations $\hat{\beta}_i$ pour chaque traitement $i = 1, \ldots, I$;
2. utiliser les résidus provenant de ces ajustements préliminaires pour estimer chaque matrice R_i. La forme fonctionnelle S_i étant définie, les paramètres σ^2 et γ peuvent être estimés par maximum de vraisemblance (Davidian et Giltinan, 1995, p127). On construit alors la matrice de poids

$$\hat{W}_i = S_i^{-1}(\hat{\beta}_i, \hat{\gamma});$$

3. utiliser les matrices de poids de l'étape 2 et réestimer les β_i par I minimisations séparées : pour le traitement i, $i = 1, \ldots, I$, minimiser en β_i

$$(y_i - f(\beta_i, X_i))' \hat{W}_i (y_i - f(\beta_i, X_i)) .$$

On utilise alors ces nouvelles estimations de β_i et on retourne à l'étape 2. Cet algorithme est itératif et converge souvent après quelques itérations (Carroll, Wu *et al.*, 1988). Les dernières estimations de β_i sont notées β_i^*.

Dans la deuxième phase de la méthode proposée, on utilise les estimations individuelles β_i^* pour estimer les paramètres de la population, β et D. Plusieurs méthodes sont envisageables pour estimer ces paramètres. On a

$$\beta_i = A_i\beta + b_i,$$

tel que $\beta_i \sim \mathcal{N}(A_i\beta, D)$; on considère ainsi uniquement les situations où toutes les composantes de β_i sont supposées aléatoires. Une première méthode (Standard Two-Stage Method), proposée par Steimer *et al.* (1984), consiste à supposer que les estimations de β_i^* sont sans erreur et donc que $\beta_i^* = \beta_i$. Dans le cas simple, où $A_i = Id_p$ et où les β_i sont indépendants et identiquement distribués, β peut être estimé par la moyenne empirique des β_i^* et D par les covariances empiriques de β_i^*. Une deuxième méthode (appelé Global Two Stage Method), tient compte de l'erreur dans les estimations de β_i^*. Davidian et Giltinan recommandent d'utiliser cette deuxième méthode. L'incorporation de l'incertitude de l'estimation de β_i^* est habituellement fondée sur la théorie asymptotique pour β_i^* à β_i donné. On peut supposer que $\beta_i^*|\beta_i$ est approximativement $\mathcal{N}(\beta_i, C_i)$, où C_i est la matrice de covariances asymptotique des β_i^*. Si les β_i^* sont estimés par moindres carrés généralisés, C_i s'écrit : $C_i = \sigma^2(X_i'(\beta_i^*)S_i^{-1}(\beta_i^*, \gamma)X_i(\beta_i^*))^{-1}$ avec $X_i(\beta_i^*)$ la linéarisation du modèle au voisinage de β_i^* ($f(\beta_i, X_i) \approx X_i(\beta_i^*)\beta_i$ et $X_i(\beta_i^*)$ joue le rôle de la matrice X dans le modèle linéaire $y = X\beta + e$). Sous l'hypothèse $\beta_i^*|\beta_i \sim \mathcal{N}(\beta_i, C_i)$, la distribution marginale de β_i^* est approximativement normale de moyenne $A_i\beta$ et de matrice de covariances $C_i + D$. Les β_i^* peuvent alors être approchés par

$$\beta_i^* \approx A_i\beta + b_i + e_i^* \tag{5}$$

où e_i^* a pour moyenne 0 et pour matrice de covariances C_i. L'approximation (5) suggère d'estimer β et D par maximum de vraisemblance à partir des pseudo-données β_i^*. Ainsi, β et D sont estimés en minimisant (le double de) l'opposée de la log-vraisemblance :

$$ML(\beta, D) = \sum_{i=1}^{I} \log|C_i + D| + \sum_{i=1}^{I} (\beta_i^* - A_i\beta)'(C_i + D)^{-1}(\beta_i^* - A_i\beta)$$

où C_i est fixée et connue pour chaque i. Les estimations de β et D satisfont alors les relations suivantes :

$$\hat{\beta} = \left(\sum_{i=1}^{I} A_i'(C_i + \hat{D})^{-1} A_i \right)^{-1} \sum_{i=1}^{I} A_i'(C_i + \hat{D})^{-1}\beta_i^* \tag{6}$$

et

$$\hat{D} = I^{-1} \sum_{i=1}^{I} c_i^* c_i^{*\prime} + I^{-1} \sum_{i=1}^{I} (C_i^{-1} + \hat{D}^{-1})^{-1} \tag{7}$$

avec

$$c_i^* = (C_i^{-1} + \hat{D}^{-1})^{-1} C_i^{-1} (\beta_i^* - A_i \hat{\beta}).$$

Une estimation de la matrice de covariances asymptotique C_i peut être obtenue en substituant les estimations de σ et γ et l'estimation finale, pour le traitement i, β_i^*, dans l'expression de la covariance asymptotique

$$C_i = \sigma^2 (X_i'(\beta_i^*) S_i^{-1}(\beta_i^*, \gamma) X_i(\beta_i^*))^{-1}.$$

Les estimateurs des équations (6) et (7) peuvent être obtenues par un algorithme de Newton-Raphson ou par un algorithme EM. Dans l'exemple, nous utiliserons l'algorithme EM suivant :

1. *étape E* : à l'étape $c+1$, on calcule pour tout i

$$\hat{\beta}_{i,(c+1)} = \left(C_i^{-1} + \hat{D}_{(c)}^{-1} \right)^{-1} \left(C_i^{-1} \beta_i^* + \hat{D}_{(c)}^{-1} \hat{\beta}_{(c)} \right);$$

2. *étape M* :

$$\hat{\beta}_{(c+1)} = I^{-1} \sum_{i=1}^{I} \hat{\beta}_{i,(c+1)},$$

$$\hat{D}_{(c+1)} = I^{-1} \sum_{i=1}^{I} \left(C_i^{-1} + \hat{D}_{(c)}^{-1} \right)^{-1} + I^{-1} \sum_{i=1}^{I} \left(\hat{\beta}_{i,(c+1)} - \hat{\beta}_{(c+1)} \right) \left(\hat{\beta}_{i,(c+1)} - \hat{\beta}_{(c+1)} \right)'.$$

On itère alors les étapes E et M jusqu'à convergence, *i.e.* jusqu'à ce que les différences entre les estimations successives de β et de D soient suffisamment petites.

Notons que l'estimateur de β (de l'équation (6)) a la forme d'un estimateur des moindres carrés généralisés, et ainsi sera un estimateur naturel de β même si les distributions de b_i et e_i^* ne sont pas exactement normales. En effet, la covariance marginale de β_i^* ne dépend pas de β et donc les estimateurs du maximum de vraisemblance et des moindres carrés généralisés de β coïncident. Par conséquent, la méthode en deux étapes est robuste à la non normalité de b_i et e_i^*.

Il est clair que les estimations de β et D sont d'autant plus robustes que les dimensions de ce vecteur et de cette matrice sont faibles, et que le nombre de mesures et le nombre d'individus est important. Le nombre de paramètres du modèle déterministe rendus aléatoires doit être faible, et le choix des paramètres qui sont rendus aléatoires est donc important. La construction d'un modèle stochastique étant

relativement longue, on ne peut généralement pas construire beaucoup de modèles puis choisir le modèle qui décrit le mieux la réalité (en se basant sur la vraisemblance, par exemple). Il est inutile de rendre aléatoire un paramètre du modèle qui a un effet négligeable sur les sorties du modèle car, dans ce cas, la variabilité créée sera négligeable. Une analyse de sensibilité permet d'éliminer les paramètres inadaptés. L'analyse de sensibilité (les deux articles fondateurs de cette technique sont ceux de Mc Kay *et al.*, 1979 et Iman et Connover, 1980) permet en effet de rechercher les variables ou les paramètres influents d'un modèle (le chapitre 2 de Husson, 1997, est une étude bibliographique sur l'analyse de sensibilité et l'analyse d'incertitude). Ensuite, on préférera rendre aléatoires des paramètres qui sont mal connus ou mal estimés plutôt que des paramètres pour lesquels l'estimation est précise.

Pour les modèles de croissance et les essais en pharmacocinétique, les variances intra-individuelle sont souvent faibles comparées à la valeur de la réponse moyenne, et le nombre d'observations par traitement est raisonnable. Ainsi, les approximations que font Davidian et Giltinan semblent raisonnables et donc cette méthode est la plus apte à décrire la variabilité. La distribution conjointe des paramètres est alors totalement déterminée puisque la loi est normale et la moyenne et la matrice de covariances sont connues. Pour effectuer des prédictions avec le modèle stochastique, il suffit de remplacer le vecteur des paramètres du modèle déterministe par un vecteur de paramètres β_i choisi au hasard dans la distribution que nous venons de définir (de moyenne $A_i\beta$ et de matrice de covariances $D + C_i$), au début de chaque simulation.

4. Exemple

Notre objectif est de construire un modèle stochastique de croissance de colza à partir d'un modèle de colza déterministe nommé CECOL. Ce modèle est décrit dans Husson (1997) et plus succinctement dans Husson *et al.* (1998). Les variables d'entrée (X_i) du modèle sont : l'état du sol le 1er août (quantité d'eau, analyse physico-chimique du sol, culture précédente, la date de 1er août est antérieure à la date de semis car le colza est semé à l'automne), les données météorologiques journalières (pluviométrie, température minimum et maximum, rayonnement solaire), la variété de colza utilisée et enfin la densité de semis. Les variables de sortie du modèle sont très nombreuses mais beaucoup de ces variables sont difficilement observables. Les deux variables de sortie qui ont été observées et qui sont disponibles sont le rendement et la quantité de matière sèche (ce sont donc nos y_i). Parmi les variables prédites par le modèle et difficilement observables, on peut citer : la quantité d'azote dans le sol, la quantité de nitrate lessivée (nitrate qui se retrouve dans les nappes phréatiques), quantité d'azote perdu par ruissellement, quantité d'eau dans le sol, etc. Ces variables sont rarement observées (par manque de temps et en raison du coût) et pourtant elles intéressent énormément les agronomes qui aimeraient raisonner les apports d'azote (apporter suffisamment d'azote pour avoir un rendement important et ne pas en apporter trop pour ne pas polluer l'environnement).

Le modèle déterministe est complexe et calcule jour après jour, (en fonction du climat, de l'état de la plante, des échanges d'eau et d'azote sol-plante) plusieurs variables décrivant l'état de la plante (nombre de feuilles, quantité de matière sèche, indice foliaire, nombre de feuilles tombées, ...), plusieurs variables décrivant l'état

hydrique et l'état physico-chimique du sol. Les équations du modèle sont très nombreuses (le programme fait plus de 2000 lignes de Fortran), elles utilisent 90 paramètres (estimés par des expériences antérieures), elles sont souvent non linéaires et elles sont utilisées à chaque jour de la simulation.

Pour construire le modèle stochastique, les données disponibles sont issues d'une série de 70 traitements répétés chacun trois ou quatre fois pour un total de 251 individus. Pour chaque individu, le rendement et de deux à quatre prélèvements de matières sèches ont été observés.

Dans le tableau 1, on a noté le biais et les écarts-types des résidus du modèle déterministe. Le biais est important car la variété utilisée est beaucoup plus récente que les variétés qui ont été utilisées pour estimer les paramètres du modèle déterministe.

Les prélèvements de matière sèche n'étant pas effectués à des dates identiques, on a utilisé le test d'indépendance des signes des résidus que nous avons construit. Avec un nombre de changements de signes observé de $nobs = 24$ et un nombre de changements de signes maximal global de $nmax = 103$, on rejette l'hypothèse « *l'occurrence des signes des résidus sont indépendants* » (probabilité critique de $1.503\ 10^{-31}$), et, par suite, on rejette l'hypothèse « *les résidus sont indépendants* ».

Le test de multinormalité de Mardia (1974) indique que les résidus suivent une loi multinormale (probabilité critique de 0.137) et la présence de répétitions (3 ou 4 par traitement) permet de tester et rejeter (au seuil 5 %) l'hypothèse « *il n'y a pas d'effet des traitements sur les résidus (du rendement ou d'un prélèvement)* ». On considère donc qu'il y a un effet du traitement sur les résidus.

Choix des paramètres

Le nombre de paramètres dans le modèle étant très important (90), il est impossible de tous les rendre aléatoires. Ainsi, nous allons considérer que 88 paramètres sont fixes (et que les valeurs de ces paramètres sont les valeurs estimées lors des expériences antérieures) et que seulement deux paramètres sont aléatoires. On est donc amené à choisir deux paramètres parmi les 90. Une analyse de sensibilité nous a permis d'éliminer de nombreux paramètres qui n'ont pas d'influence sur le rendement et la matière sèche.

Parmi les paramètres restants, on a choisi de rendre aléatoires deux paramètres de croissance qui sont mal connus par les agronomes. Ainsi, on pourra tenir compte d'effets individuels qu'on ne peut expliquer, comme par exemple un individu qui se développe plus rapidement en raison d'aptitude individuelle de la plante ou d'un micro-environnement favorable. Ces paramètres permettent de créer de la variabilité du début à la fin de la culture.

Estimation de β et D

On a pris $A_i = Id_p$ et $B_i = Id_p$ pour tout i. Chaque traitement a son propre jeu de paramètres β_i, l'espérance du 2-vecteur β_i est β et sa matrice de covariances est la matrice D de dimension 2×2.

On met donc en place la méthode proposée par Davidian et Giltinan. Au vu du tableau 1, on ne peut pas considérer que les variances soient égales et on ne peut pas

considérer qu'elles soient fonction des espérances du modèle, donc on pose

$$R_i = \left(Id_{(n_i \times n_i)} \quad 0_{(n_i \times N - n_i)} \right) \Sigma \begin{pmatrix} Id_{(n_i \times n_i)} \\ 0_{(N-n_i \times n_i)} \end{pmatrix}.$$

avec N le nombre de mesures qu'il est possible de faire (*i.e.* ici le nombre de jour entre le semis et la récolte) et Σ une matrice diagonale telle que :

$$\widehat{\Sigma}_{jj} = \frac{1}{I_j} \sum_{i=1}^{I_j} \left(y_i^{\{j\}} - f\left(\beta_i, X_i^{d\{j\}}\right) \right)^2$$

où I_j est le nombre de traitements observés pour la $j^{\text{ème}}$ variable ($I_j = 70$ pour les deux premiers prélèvements de matière sèche et le rendement, $I_j = 15$ pour les prélèvements 3 et 4).

Pour estimer β et D, on a ensuite utilisé l'algorithme EM décrit précédemment en considérant que l'algorithme avait convergé dès que la différence entre deux valeurs successives de β et D étaient inférieures à 10^{-10}.

Pour estimer C_i, la matrice de covariances des e_i^*, on a linéarisé le modèle au voisinage de β_i^* et écrit :

$$C_i = Var(e_i^*) = (X_i'(\beta_i^*)\Sigma_i X_i(\beta_i^*))^{-1}. \tag{8}$$

On a alors testé (et accepté au seuil 5%) l'égalité des matrices C_i à partir du test de Box (voir Mardia, 1970) et donc on a estimé une matrice commune C.

On a obtenu

$$\hat{\beta} = \begin{pmatrix} 0.836 \\ 1.16 \end{pmatrix}, \ \hat{D} = \begin{pmatrix} 5.33\ 10^{-2} & 2.60\ 10^{-3} \\ 2.60\ 10^{-3} & 1.80\ 10^{-2} \end{pmatrix}$$

$$\text{et } \hat{C} = \begin{pmatrix} 3.41\ 10^{-4} & -1.61\ 10^{-3} \\ -1.61\ 10^{-3} & 8.04\ 10^{-4} \end{pmatrix}.$$

La loi conjointe des paramètres est alors totalement déterminée puisque la distribution est normale et la moyenne et la matrice de covariances sont connues. Pour faire les simulations avec le modèle stochastique, il suffit alors d'utiliser le modèle déterministe en prenant au début de chaque simulation un vecteur de paramètre β_i au hasard dans la loi de distribution multinormale de moyenne $\hat{\beta}$ et de matrice de covariances $\hat{D} + \hat{C}$.

Vérification de la construction du modèle stochastique

Pour vérifier que la variabilité engendrée par le modèle stochastique est équivalente à la variabilité des résidus du modèle déterministe, on a comparé les variances et corrélations des résidus du modèle déterministe avec les variances intra-traitement et les corrélations intra-traitement des résultats du modèle stochastique.

Pour cela, on a effectué 100 simulations (ce qui est suffisant pour avoir des résultats stables) et calculé les prédictions des matières sèches et du rendement sur chacun des 70 traitements. Le tableau 1, donne les biais et les écarts-types intra-traitement du modèle stochastique et, les biais et les écarts-types des résidus du modèle déterministe. Le tableau 2 donne la matrice des corrélations entre les résidus du modèle déterministe et la matrice des corrélations des prédictions intra-traitement du modèle stochastique.

TABLEAU 1

Biais et écarts-types intra-traitement des prédictions du modèle stochastique et des résidus du modèle déterministe

Variables de sortie du modèle	Biais modèle stochastique	Biais modèle déterministe	Écart-type intratraitement des prédictions du modèle stochastique	Écart-type des résidus du modèle déterministe
1er prélèvement de MS	−0,034	−0,779	0,914	0,938
2eme prélèvement de MS	0,515	−0,440	1,199	1,350
3eme prélèvement de MS	−0,122	−0,637	1,186	0,295
4eme prélèvement de MS	−0,127	−0,889	1,159	0,593
Rendement	−0,009	0,294	0,777	0,689

On peut remarquer que les biais du modèle stochastique sont faibles (ce qui s'explique par la réestimation de deux paramètres). Les écarts-types et les covariances intra-traitement du modèle stochastique et les écarts-types et covariances des résidus du modèle déterministe sont proches pour les deux premiers prélèvements de matière sèche et pour le rendement (par exemple, l'écart-type généré par le modèle stochastique pour le deuxième prélèvement est de 1.199 à comparer à 1.35, et la corrélation générée par le modèle stochastique entre les deux premiers prélèvements est de 0.98 à comparer à 0.91). Pour les deux derniers prélèvements, le nombre de données manquantes étant important, l'ajustement est moins bon.

La modélisation stochastique est satisfaisante pour les prédictions des deux premières quantités de matière sèche et du rendement puisque ces prédictions sont sans biais et les variances et corrélations intra-traitement des modèles stochastiques sont proches des variances et corrélations des résidus du modèle déterministe. En revanche, pour le troisième et le quatrième prélèvement de matière sèche, la prédiction n'est pas bonne (à cause d'un nombre important de données manquantes). Cette méthode a de plus l'avantage de générer de la variabilité dans toutes les sorties du modèle (pour la quantité de matière sèche, non seulement les jours de prélèvements, mais tous les jours, et pour les autres variables du modèle comme la quantité d'azote dans le sol, les sorties sont aléatoires car ces variables dépendent de la matière sèche qui est aléatoire).

TABLEAU 2
Matrice de corrélations des résidus du modèle déterministe et matrice de corrélations intra-traitement des prédictions du modèle stochastique

	MS1	MS2	MS3	MS4	RDT
MS1	1.00				
MS2	0.91	1.00			
MS3	0.38	0.06	1.00		
MS4	0.66	0.68	0.30	1.00	
RDT	0.76	0.72	0.09	-0.28	1.00

	MS1	MS2	MS3	MS4	RDT
MS1	1.00				
MS2	0.98	1.00			
MS3	0.98	0.99	1.00		
MS4	0.98	0.99	1.00	1.00	
RDT	0.67	0.69	0.66	0.66	1.00

5. Discussion

Le choix d'une modélisation stochastique dépend du comportement des résidus du modèle déterministe et du modèle déterministe lui-même. Pour des modèles de croissance avec des mesures répétées, les modèles mixtes semblent les plus adaptés pour reproduire la variabilité des données. Pour estimer les paramètres de cette modélisation, deux méthodes sont envisageables. Si le modèle peut facilement être linéarisé, et si peu de mesures sont effectués pour chaque individu, alors la méthode proposée par Lindstrom et Bates (1990) est la plus adaptée. En revanche, s'il est difficile de trouver les développements de Taylor du modèle, et si suffisamment de mesures sont effectuées sur chaque individu, la méthode proposée par Davidian et Giltinan (1995) est la mieux adaptée. Pour des modèles agronomiques, on conseille donc d'utiliser une méthode fondée sur celle de Davidian et Giltinan.

Un modèle stochastique ainsi construit permet de créer de la variabilité dans toutes les variables de sortie du modèle qu'elles soient à temps discret ou à temps fixe. En effet, même si les paramètres du modèle rendus aléatoires n'interviennent pas directement dans l'équation d'une variable intermédiaire ou de sortie, cette variable deviendra aléatoire si elle dépend d'une variable rendue aléatoire.

Comme pour les modèles déterministes, l'étape finale de la construction d'un modèle est l'évaluation (Husson, *et al.*, 1998). Notons que l'évaluation est beaucoup plus complexe que pour un modèle déterministe car elle doit porter à la fois sur les espérances, les variances et même les covariances des variables de sortie. On peut alors utiliser le test d'égalité de deux distributions empiriques (test de Smirnov) : la distribution des prédictions du modèle stochastique et la distribution des observations.

Un tel modèle stochastique permet de prédire la distribution de y à X fixé. Cependant, les vraies valeurs des paramètres du modèle stochastique, β et D, sont inconnues et on utilise les estimations $\hat{\beta}$ et $\hat{D}$ pour calculer la distribution des prédictions $f((\hat{\beta}, \hat{D}), X)$. Cette imprécision sur les valeurs des paramètres β et D se répercute ensuite sur la distribution des sorties du modèle. Ainsi, si on calcule

l'intervalle de confiance d'une prédiction d'une variable de sortie du modèle, on sous-estimera son amplitude puisqu'on ne tient pas compte de l'incertitude engendrée par la non-connaissance de β et D.

Remerciements

Je remercie l'Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME) ainsi que le Centre Technique Interprofessionnel des Oléagineux Metropolitains (CETIOM) qui ont participé aux financements de ces travaux.

Références

BEAL, S. L. et SHEINER, L. B. (1988). *Heteroscedastic non linear regression.* Technometrics, **30**, 327-338.

BERKEY, C. S. (1982). *Bayesian approach for a nonlinear growth model.* Biometrics **38**, 953-961.

CARROLL, R. J. et RUPPERT, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, London.

CARROLL, R. J., WU, C. F. J. et RUPPERT, D. (1988). *The effect of estimating weights in weighted least squares.* Journal of the American Statistical Association, **83**, 1045-1054.

DAVIDIAN, M. et GILTINAN, D. M. (1993). *Some simple methods for estimating intra-individual variability in nonlinear mixed effects models.* Biometrics **49** 59-73.

DAVIDIAN, M. et GILTINAN, D. M. (1995). *Nonlinear models for repeated measurement data.* Chapman and Hall, London.

FAIVRE, R., GOFFINET, B. et WALLACH, D. (1991). *Utilisation de Donnees Intermédiaires pour Corriger la Prédiction de Modèles Mécanistes.* Biometrics. **47**, 1-12.

HUSSON, F. (1997) *Validation et introduction d'aléas dans un modèle déterministe complexe : application à un modèle de croissance.* PhD thesis, Institut National Agronomique Paris-Grignon, France.

HUSSON, F., WALLACH, D. et VANDEPUTTE, B. (1998) *Evaluation of CECOL a model of winter rape, Brassica Napus L.*. European Journal of Agronomy, **8**, 205-214.

IMAN, R. L. et CONOVER, W. J. (1980). Small sample sensitivity analysis techniques for computer models, with an application to risk assessmnet (with discussions) *Communications in Statistics, Part A-Theory and Methods*. **9**, 1749-1874

LINDSTROM, M. J. et BATES, D. M. (1988). *Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data* Journal of the American Statistical Association, **83**, 1014-1022.

LINDSTROM, M. J. et BATES, D. M. (1990). *Nonlinear mixed effects models for repeated measures data.* Biometrics, **46**, 673-687.

MARDIA, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57, 3**, 519-530.

MARDIA, K. V. (1974). *Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies.* Sankhya B, **36**, 115-128.

Mc KAY, M. D., CONNOVER, W. J. et BECKMAN, R. J. (1979). A comparison of three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. *Technometrics*, **21**, 239-245.

RACINE-POON, A. (1985). *Bayesian approach to nonlinear random effects models.* Biometrics, **41**, 1015-1023.

SEBER, G. A. F. et WILD, C. J. (1989). *Nonlinear Regression.* Wiley, New-York.

SHEINER, L. B. et BEAL, S. L. (1980). *Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model : Routine clinical pharmacokinetic data.* Journal of pharmacokinetics and Biopharmaceutics, **8**, 553-571.

STEIMER, J. L., MALLET, A., GOLLMARD, J. L. et BOISIVIEUX, J. F. (1984). *Alternative approches to estimation of population pharmacokinetic parameters : Comparison with the nonlinear mixed effect model.* Drug Metabolism Reviews, **15**, 265-292.

STEIMER, J. L., MALLET, A. et MENTRÉ, F. (1985). *Estimation interindividual pharmacokinetic variability.* Variability in Drug Therapy (eds. M. Rowland *et al.*). Raven Press, New-York.

TOMASSONE, R., DERVIN, C. et MASSON, J. P. (1993). *Biométrie : Modélisation de phénomènes biologiques.* Masson, Paris.

WHITE, G. C. et BRISBIN, L. B. (1980). *Estimation and comparison of parameters in stochastic growth models for barn owls.* Growth, **44**, 97-111.