MIA-Jouy
INRA
domaine de Vilvert
F-78352 Jouy-en-Josas Cedex

Mise en oeuvre du modèle linéaire mixte pour deux exemples types à l'aide de la procédure `mixed` de SAS et la fonction `lme` de R

Exemple d'un réseau d'essais :
le rendement des variétés de maïs transgéniques

Exemple d'une expérimentation pluriannuelle :
la fourchaison des chênes

Jessica Schohn
Sylvie Huet
Hervé Monod

Rapport technique 2004-1, 73pp

Table des matières

Introduction

L'intérêt de ce document est de présenter les démarches statistiques à adopter à travers le *modèle linéaire mixte* dans le cadre des *dispositifs aléatoires en blocs complets*. Nous privilégierons leur application aux réseaux d'essais d'évaluation variétale et à l'expérimentation pluriannuelle en nous appuyant respectivement sur l'exemple du rendement des variétés de maïs transgéniques sur le réseau national, et sur l'exemple de la fourchaison des chênes de Normandie.

Le premier chapitre est consacré dans un premier temps à leur description, puis à leur modélisation en rappelant qu'un modèle est mixte dans le sens où il regroupe des effets déterministes et des effets aléatoires.

Dans le chapitre 2, nous présentons en détails les méthodes d'analyse utilisées, notamment l'ajustement du modèle et les techniques d'estimation des paramètres, la vérification des hypothèses, puis les différents tests statistiques relatifs à la comparaison et à l'exploitation des modèles.

Enfin, appliquant les méthodes évoquées, nous exposons les résultats des analyses de variance des deux exemples cités, et détaillons leurs interprétations, les analyses étant obtenues à l'aide des outils informatiques `SAS` pour le premier exemple, et `R` pour le second.

Notre premier objectif a été de donner les informations nécessaires pour bien comprendre les principes de l'analyse de variance et bien interpréter ses résultats, en utilisant le modèle linéaire mixte.

Chapitre 1

Ecriture des modèles en fonction de l'expérience

1.1 Description des exemples

Les dispositifs expérimentaux classiques ont pour principes de base la répétition des traitements, la randomisation et la répartition des parcelles en blocs pour contrôler leur hétérogénéité. Les essais d'évaluation variétale et l'expérimentation pluriannuelle utilisent le plus souvent des *plans en blocs complets randomisés*, dont le principe est le suivant : dans chaque bloc, on alloue indépendamment une unité exactement à chaque traitement, et ce de façon aléatoire.

1.1.1 Exemple 1 : réseau d'essais pour évaluer des variétés de maïs transgéniques

◇ ***Les objectifs biologiques : évaluer l'apport du transgène***

Dans le cadre de l'inscription des nouvelles variétés au Catalogue Officiel, des études sont menées sur des variétés de maïs transgéniques tolérantes aux attaques de la pyrale, principal ravageur des champs de maïs en France. Ces variétés ont intégré dans leur génome un gène (de la bactérie *Bt*) codant une protéine toxique pour certaines larves d'insectes car déclenchant la paralysie de leur système digestif.

L'analyse de l'effet de l'insertion du gène de tolérance a porté sur le rendement en grain. Afin d'évaluer et de mettre en évidence cet effet, il s'agit d'étudier les différences de rendement entre les variétés transgéniques et non transgéniques, en particulier entre les formes transgénique et non transgénique d'une même variété, pour différents niveaux d'infestation par la pyrale, et en distinguant les zones de précocité. (*Schohn*, 2003)

◇ ***Les données : de type réseau d'essais***

Afin d'intégrer la variabilité entre lieux, des **réseaux d'expérimentation variétale** sont organisés : des **essais** semblables sont implantés en différents lieux et sur plusieurs années, un seul essai apportant une information insuffisante.

Cette étude se base sur les résultats des essais variétaux de maïs réalisés dans le cadre du CTPS de 1997 à 2002, soit au total 135 essais.

Dans chaque site expérimental, les génotypes sont répartis, chaque année, en deux essais distincts. Plus précisément, un essai est défini par son caractère protégé/infesté :

- ⋆ un essai est **protégé** à l'aide d'insecticides pour empêcher l'attaque des pyrales et ainsi vérifier le caractère de l'équivalence entre les variétés transgéniques et non transgéniques ;
- ⋆ un essai est **infesté** par les pyrales pour vérifier le caractère tolérant de la variété transgénique.

◇ *Le dipositif expérimental*

L'unité expérimentale est la **parcelle** constituée de 4 rangs de 5m de long et 0.80m d'écartement dans le cas du maïs. Toutes les plantes d'une même parcelle sont semées à la même date et sont issues de la même variété. Le **rendement en grain** est mesuré par la récolte des deux rangs centraux de la parcelle, pour éviter des biais parfois liés à des interférences entre variétés voisines (*effets de compétition*).

Afin de construire un dispositif expérimental tenant compte de possibles hétérogénéités du terrain, il est nécessaire de mettre en place des **blocs**. Dans un plan en blocs **complets**, les parcelles sont réparties en r blocs de taille v, et les blocs sont formés de façon à ce que les parcelles d'un même bloc soient *a priori* aussi homogènes que possible.

Chaque variété est présente dans chaque bloc exactement une fois, et les variétés sont réparties sur les parcelles par un tirage aléatoire : on dit que la répartition des variétés dans les blocs est **randomisée**, indépendamment entre blocs. Ceci évite que des sources d'hétérogénéité non contrôlées ne viennent biaiser les comparaisons entre variétés. Ces dispositifs sont connus sous le nom de *plans en blocs complets randomisés* : ils sont bien adaptés à de petits nombres de variétés (< 15).

◇ *Les facteurs et les variables à prendre en compte*

Il existe deux types de facteurs caractérisant un réseau d'essais.

LES FACTEURS ENVIRONNEMENTAUX :

- ⋆ l'année : le réseau national d'essais considéré s'étend de 1997 à 2002 ;
- ⋆ la zone de précocité : l'expérimentation a porté sur des variétés de deux groupes de précocité : un groupe de variétés très précoces à demi-précoces (A-B-C1-C2), et un groupe de variétés demi-tardives à très tardives (D-E) ;
- ⋆ l'essai : il correspond à un lieu d'expérimentation (une commune) sur le réseau national (ce sont pratiquement les mêmes tous les ans pour une zone de précocité donnée) ;
- ⋆ l'indice d'infestation : évalué à partir de la casse des épis et des dires des experts, il caractérise le niveau d'infestation de chaque essai, et prend les valeurs 1 (attaque quasi-inexistante), 2 (attaque moyenne) ou 3 (attaque forte);
- ⋆ le bloc : tous les essais ont été réalisés avec 4 blocs complets.

Notons que pour être efficaces, les plans doivent maximiser la variabilité inter-bloc et minimiser la variabilité intra-bloc, de sorte à rendre les blocs le plus homogènes possible.

- ⋆ la dénomination de la variété : elle représente le fonds génétique de la variété (les génotypes en étude sont tous différents et issus de divers programmes de sélection européens) ;
- ⋆ la forme variétale : elle permet d'étudier plus largement l'effet transgénique sur l'ensemble des données. On distingue quatre types de variétés, les deux premières sont non transgéniques et les deux suivantes sont transgéniques :

les témoins servent de référence entre essais et sur plusieurs années ;

les variétés initiales sont toujours accompagnées d'une forme OGM à même fonds génétique ;

les variétés modifiées portent un transgène et sont comparées à leur forme initiale ;

les variétés nouvelles, sans variété initiale identifiée, ne peuvent être comparées qu'à des variétés à fonds génétiques différents.

Le rendement en grain (mesuré en quintaux/hectare) est relevé pour chaque variété de maïs présente dans chacun des blocs d'un essai : il constitue la variable à expliquer ou réponse.

1.1.2 Exemple 2 : expérimentation pluriannuelle sur la fourchaison des chênes de Normandie

◇ ***Les objectifs biologiques : estimer le nombre de fourches en fonction de la vigueur de l'arbre***

Cette étude est destinée à répondre aux questions de forestiers de l'INRA de Nancy, qui cherchent à identifier les conditions favorables à la production de chênes de qualité en Normandie. Le critère de qualité étudié est la fourchaison.

L'objectif est d'analyser l'effet des densités de plantation initiales et des mesures de la vigueur des arbres sur les effectifs de fourches présentes, et plus précisément de modéliser l'influence du niveau de compétition sur la forme des chênes (présence ou non de fourches), avec des niveaux de compétition imposés par trois densités. (*Castelli*, 2003)

◇ ***Les données : de type répétées***

En agriculture, les conditions sont très déterminantes et les variations pluriannuelles souvent très importantes. Afin d'étudier la dynamique et de suivre la fourchaison sur un grand nombre d'années, plusieurs mesures sont réalisées sur le même individu : on parle de **mesures répétées**.

Pour apprécier l'effet de l'âge sur la fourchaison, les essais doivent alors être répétés plusieurs années de suite : on réalise pour cela des **essais pluriannuels**.

◇ ***Le dipositif expérimental***

Le dispositif expérimental a été installé en 1981 dans la forêt domaniale de Lyons-la-Forêt en Normandie. Trois espacements à la plantation des chênes sont testés et correspondent à des **densités** par hectare différentes. L'étude est réalisée sur 4 blocs, identifiés sur la base de différences de sol. Chaque **bloc** est partagé en 3 placeaux représentant chacun une densité de plantation différente (affectation au hasard), et les chênes sont plantés en rectangles selon une distance entre rangées constante de 2.5m et une distance à l'intérieur des rangées de 0.75 à 3m selon la densité.

Pour éviter de mélanger les densités et de fausser les données, chaque placeau est entouré d'une zone d'isolement de 4 à 6 mètres de largeur constituée de 2 à 5 lignes de plants et sur laquelle les données ne sont pas prélevées.

◇ ***Les facteurs et les variables à prendre en compte***

- ⋆ l'arbre : nous disposons d'un échantillon de 514 plants de chênes, tous âgés d'un an lors de la plantation et de même origine (pépinière forestière proche de l'essai) ;
- ⋆ le bloc : les arbres sont répartis en 4 blocs sur une surface totale de 1.3 ha ;
- ⋆ le placeau et la densité : chaque bloc est partagé en 3 placeaux représentant une densité différente : 1333, 2667 et 5333 tiges par hectare ;
- ⋆ l'âge de l'arbre (l'année) : les mesures réalisées s'étalent sur 20 ans et permettent de suivre les chênes aux âges de 4 (1984), 10 (1990), 14 (1994), 17 (1997) et 20 ans (2000).

La variable à expliquer, le nombre de fourches, est une variable de comptage relevée à chacune de ces années sur chaque arbre. Elle est fonction des variables qualitatives définies ci-dessus et de la variable quantitative, ou **covariable**, circonférence.

Plus précisément, la variable vigueur peut être considérée de plusieurs façons :

- soit par rapport à la circonférence de l'arbre : plus un arbre est gros, plus il est vigoureux ;
- soit en considérant la variation de circonférence et de hauteur entre les dates t et $t+1$: les arbres les plus vigoureux sont ceux pour lesquels la croissance en variation est la plus forte.

Pour notre exemple, la circonférence de l'arbre représentera sa vigueur et sera mesurée à 1m30.

Nous avons observé deux types de corrélation influençant la fourchaison :

- une corrélation positive entre la fourchaison et la vigueur de l'arbre : plus l'arbre est vigoureux, plus le nombre de fourches est important ;
- une corrélation négative entre la vigueur de l'arbre et la densité : les arbres ont tendance à être plus gros sur un sol où la densité est plus faible.

1.2 Ecriture des modèles statistiques

1.2.1 Le modèle linéaire mixte

Dans une analyse de variance, le but est d'expliquer une variable quantitative à travers un modèle qui décrit une relation entre une **variable réponse** (variable à expliquer) et des facteurs ou des variables explicatifs. Le choix d'un modèle nécessite donc de préciser les facteurs et les variables, ainsi que les éventuelles interactions dont on souhaite tenir compte pour expliquer la réponse.

Par ailleurs, un facteur peut être considéré avoir des **effets fixes** ou des **effets aléatoires**. De nombreux modèles statistiques peuvent s'exprimer comme des modèles linéaires intégrant à la fois ces deux types d'effets. Ces modèles, qualifiés de **modèles linéaires mixtes**, sont souvent mieux adaptés que les modèles à effets fixes uniquement.

Nous reprenons ci-dessous les principales notions sur les composants d'un modèle mixte.

◇ ***La nature des facteurs et des variables***

Une variable peut être :

- ⋆ **qualitative** lorsque ses valeurs prennent un nombre limité de modalités, à chacune desquelles est associé un effet propre sur la réponse : on parle alors de **facteur** ;
- ⋆ **quantitative** lorsque ses valeurs sont numériques ordonnées, sur une échelle discrète ou plus souvent continue, avec des effets sur la réponse fonction de ces valeurs numériques : on parle alors souvent de **covariables** lorsqu'il y a également des facteurs.

◇ ***Les relations entre facteurs***

Il existe deux principales relations entre facteurs :

- ⋆ soit chacun des facteurs a un sens indépendamment de l'autre : les facteurs sont dits **croisés** (c'est le cas général) ;
- ⋆ soit un facteur donné représente un découpage plus fin des indices associés à un autre facteur : le facteur est dit **hiérarchisé** au second facteur ou **emboîté** dans ce dernier.

◇ ***Les paramètres du modèle***

Les termes du modèle peuvent être :

- ⋆ à effets **fixes** lorsque leurs modalités sont étudiées en tant que telles : ces effets constituent des paramètres réels inconnus mais fixés ;

- ⋆ à effets **aléatoires** lorsqu'ils ne sont que des représentants d'un ensemble d'occurences possibles et que l'on souhaite généraliser les résultats observés à l'ensemble de la population dont l'échantillon est issu. Ces effets constituent des réalisations de variables aléatoires :
  - supposées suivre une distribution normale d'espérance nulle et de variance inconnue (par exemple σ_0^2 si on est dans le cas d'un facteur),
  - supposées indépendantes entre elles et indépendantes pour des niveaux différents d'un facteur donné ;

  et permettent de modéliser la structure de corrélation entre les observations d'une même classe d'un facteur donné ;

- ⋆ d'origine **non contrôlée** : les erreurs résiduelles constituent une composante aléatoire. Elles sont caractérisées par leur **variance résiduelle** σ^2 dont la valeur influence la qualité des estimations et la longueur des intervalles de confiance, et par conséquent, les conclusions pratiques qui en résultent.

Remarque 1 *Dans l'exemple du réseau d'essais, la variance résiduelle sur les différents essais intègre les erreurs de mesure, et, par suite de la randomisation, les différences entre parcelles non expliquées par les effets blocs. La variance résiduelle est donc d'autant plus faible que les parcelles sont semblables à l'intérieur des blocs.*

Un **modèle linéaire fixe** est un modèle pour lequel tous les facteurs sont supposés fixes : il permet de modéliser les modalités des facteurs représentées dans l'expérience.

Dans un tel modèle, les erreurs résiduelles sont la seule composante aléatoire : la variance des observations est donc égale à celle des erreurs, et les observations sont supposées indépendantes.

Un modèle comprenant les deux types de facteurs est appelé un **modèle linéaire mixte** : il permet de modéliser des modalités considérées comme un échantillon issu d'une population plus large.

Dans le modèle mixte à un facteur aléatoire, on suppose que la variance des observations est égale à $\sigma_0^2 + \sigma^2$, et que la covariance entre deux observations est égale à :

$$\begin{cases} \sigma_0^2 & \text{pour des observations appartenant à une même classe du facteur aléatoire} \\ 0 & \text{pour des observations appartenant à différentes classes du facteur aléatoire.} \end{cases}$$

1.2.2 La modélisation du rendement des variétés de maïs

La grande variabilité due aux effets de l'année et de l'interaction de l'année et de la commune nous a amenés à réaliser l'analyse année par année, cette approche étant relativement pertinente au vu des données car les mêmes variétés transgéniques sont très rarement étudiées plusieurs années de suite. De plus, une analyse par année apporte beaucoup car on observe un comportement différent d'une année à l'autre. Dans la suite, nous avons choisi de modéliser le rendement des variétés situées dans les zones de précocité A-B-C1-C2 en 2000.

Les hétérogénéités entre parcelles sont modélisées par les effets blocs. Mais les analyses réalisées précédemment soulignaient la faible variabilité entre les blocs d'un même essai, ce qui nous a conduit à considérer les moyennes de rendement variétales dans chaque essai comme variables réponses de nos modèles.(*Schohn*, 2003)

◇ ***Les essais réalisés***

Les observations sont regroupées en 12 essais, indicés par les niveaux d'infestation 1, 2 ou 3 selon la répartition suivante (remarquons la présence de 4 couples d'essais protégés/infestés) :

	niveau d'infestation		
essais	**1**	**2**	**3**
protégés	Bretenières Clermont Ferrand Magny les Hameaux Villampuy	Lance Oucques	
infestés	Lance Oucques Villampuy	Clermont Ferrand Selommes	Guyancourt

TAB. 1.1 – *Essais présents en 2000 dans les zones de précocité ABC1C2.*

Les effets lieux sont importants avec des rendements moyens qui s'élèvent selon les essais entre 100 et 140 quintaux/hectare.

◇ ***Les variétés expérimentées***

On distingue dix variétés, dont six sous leurs formes initiales/modifiées (deux autres n'étant étudiées qu'en essais protégés, nous choisissons de les supprimer du jeu de données pour le graphique de la Figure 1.1), une variété nouvelle et un témoin. Tous les essais ont été réalisés avec les six variétés de type initiales/modifiées ; la variété nouvelle et le témoin sont présents uniquement dans les essais protégés.

Dans notre étude préalable, nous avions comparé le rendement moyen des formes initiales et modifiées en essais protégés et infestés pour chacune des six variétés.

Nous avions souligné des différences de comportement, à savoir que le rendement moyen des formes modifiées est supérieur à celui des formes initiales en essais infestés par les pyrales (sauf pour une variété), et proche de celui-ci en essais protégés.

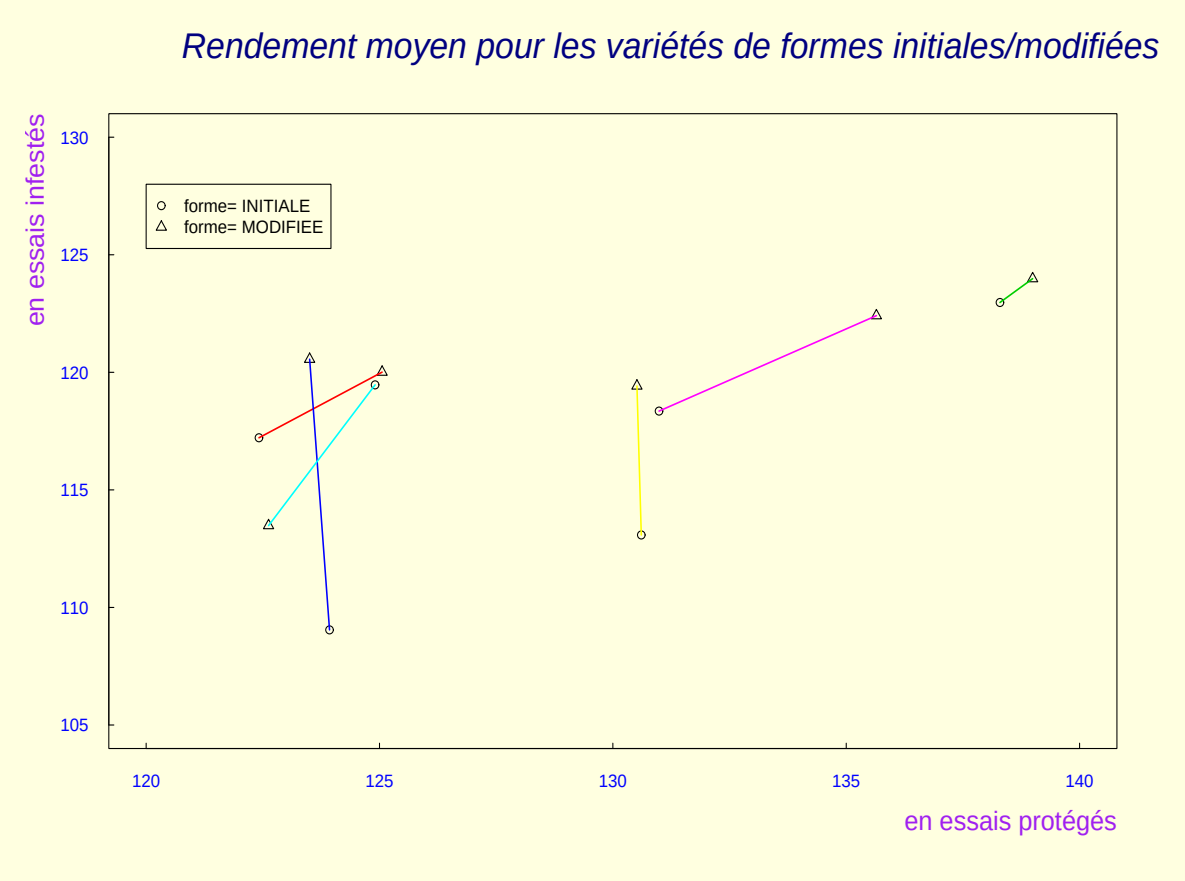

FIG. 1.1 – *Comparaison du rendement moyen en essais protégés et infestés pour chaque forme variétale, la forme variétale étant identifiée par forme et la variété par couleur.*

Notons R_{fdje} le rendement observé, pour une année donnée, sur le génotype de forme f, de dénomination (fonds génétique) d et dans l'essai e d'indice d'infestation j. Dans notre modèle, nous considérons deux paires de facteurs emboîtés : le fonds génétique de la variété est emboîté dans la forme variétale, et l'essai est emboîté dans l'indice d'infestation.

Dans notre modèle, les effets des essais sont liés à la variabilité du terrain d'expérimentation. Avec une disposition différente des parcelles, ou sur un site même très similaire, ces effets seraient différents. Les essais sont supposés représenter un échantillon aléatoire de sites et de conditions agronomiques, et leurs effets seront donc considérés comme des variables aléatoires, similaires à des termes d'erreur et supposées suivre des lois normales, indépendantes, d'espérance nulle et de variance à estimer. Naturellement, cette variance est d'autant plus élevée que le facteur à effets aléatoires essai recouvre des hétérogénéités entre ses modalités.

Par opposition aux effets des essais, les effets des variétés sont supposés avoir des valeurs fixées, a priori quelconques et reconductibles dans des conditions expérimentales similaires : le facteur variété est un facteur à effets fixes.

L'analyse des essais variétaux sur les moyennes entre blocs est basée sur le modèle suivant :

$$R_{fdje} = \mu + \underbrace{\alpha_f + \beta_{fd}}_{\Gamma_{fd}} + \underbrace{\gamma_j + A_{je}}_{\Delta_{je}} + \underbrace{(\alpha\beta)_{fj} + (\beta\gamma)_{fdj} + (\alpha A)_{fje}}_{(\Gamma\Delta)_{fdje}} + \epsilon_{fdje} \qquad (1.1)$$

où :

- μ représente la moyenne générale du modèle
- Γ_{fd} représente l'effet **Génotype**, décomposé en un effet de la forme α_f et un effet du fonds génétique β_{fd} ; notons que les effets s'interprètent ici comme des écarts à la moyenne
- Δ_{je} représente l'effet **Environnement**, décomposé en un effet de l'indice d'infestation γ_j auquel appartient l'essai et un effet de l'essai A_{je}
- $(\Gamma\Delta)_{fdje}$ représente l'**interaction Génotype** $\times$ **Environnement**, décomposée en un effet de la forme variétale $(\alpha\beta)_{fj}$ et un effet du fonds génétique $(\beta\gamma)_{fdj}$ en fonction du niveau d'infestation, et un effet de la forme selon l'essai $(\alpha A)_{fje}$
- ϵ_{fdje} représente les erreurs résiduelles sur les différents essais : elles sont supposées indépendantes, d'espérance nulle et de même variance σ^2

et :

- $\mu, \alpha_f, \beta_{fd}, \gamma_j, (\alpha\beta)_{fj}, (\beta\gamma)_{fdj}$ sont des effets fixes
- $A_{je}, (\alpha A)_{fje}$ sont des effets aléatoires centrés de variances respectives σ_A^2 et σ_B^2 (en général, les majuscules désignent les effets aléatoires).

Cette modélisation des effets aléatoires a pour conséquence de modéliser les différents niveaux de variabilité des observations et introduit une corrélation entre les variétés étudiées dans un même essai.

Plus précisément, on a :

$$\begin{aligned} \mathrm{E}(R_{fdje}) &= \mu + \Gamma_{fd} + \gamma_j + (\alpha\beta)_{fj} + (\beta\gamma)_{fdj} \\ \mathrm{var}(R_{fdje}) &= \sigma_A^2 + \sigma_B^2 + \sigma^2 \end{aligned}$$

$$\begin{aligned}
\mathrm{cov}(R_{fdje}, R_{fd'je}) &= \sigma_A^2 + \sigma_B^2 \text{ si } d \neq d' \text{ (même essai et même forme variétale)} \\
\mathrm{cov}(R_{fdje}, R_{f'd'je}) &= \sigma_A^2 \text{ si } f \neq f' \text{ (même essai et formes variétales différentes)} \\
\mathrm{cov}(R_{fdje}, R_{f'd'j'e'}) &= 0 \text{ si } j \neq j' \text{ ou } (j = j' \text{ et } e \neq e') \text{ (essais différents)}
\end{aligned}$$

1.2.3 La modélisation de la fourchaison des chênes en fonction de leur vigueur

Nous voulons étudier la relation entre le nombre de fourches et la vigueur de l'arbre. Soit F_{bdsa} le nombre de fourches observé sur le sujet (l'arbre) s, dans le placeau de densité d du bloc b et pour l'année (l'âge des arbres) a. La vigueur de l'arbre est représentée par la circonférence mesurée à 1m30 et notée C_{bdsa}.

Dans cette modélisation, le sujet est emboîté dans le placeau, qui correspond au croisement des facteurs bloc et densité, et l'âge est croisé avec tous les autres facteurs.

Nous avons dans un premier temps modélisé la relation du nombre de fourches en fonction de la circonférence de l'arbre par une fonction linéaire dont les paramètres dépendaient de la densité et de l'âge, en introduisant un effet aléatoire du bloc et de l'arbre sur la moyenne générale, et un effet aléatoire de l'arbre sur la pente de la régression en fonction de la circonférence. L'analyse de la variance nous a amenés à conclure que la fourchaison dépendait de la densité au travers de la vigueur de l'arbre. Cherchant à simplifier le modèle en gardant les effets significatifs, nous avons alors supposé une relation linéaire sans effet de la densité, en introduisant uniquement un effet de l'âge sur la circonférence.

Le modèle d'analyse de variance retenu est le suivant :

$$\boxed{F_{bdsa} = \gamma_a^0 + \gamma_a^1 C_{bdsa} + A_b + B_{bds}^0 + B_{bds}^1 C_{bdsa} + \epsilon_{bdsa}} \tag{1.2}$$

où :

γ_a^0	représente la moyenne générale et dépend de l'âge des arbres
γ_a^1	représente l'effet de l'âge sur la relation entre le nombre de fourches et la circonférence de l'arbre C_{bdsa}
$A_b + B_{bds}^0$	représente l'effet du bloc A_b auquel appartient l'arbre et l'effet de l'arbre B_{bds}^0 indépendamment de l'année
B_{bds}^1	représente l'effet spécifique de l'arbre sur la relation entre le nombre de fourches et la circonférence C_{bdsa} de l'arbre
ϵ_{bdsa}	représente l'erreur résiduelle de variance σ^2

et :

γ_a^0, γ_a^1	sont des effets fixes
$A_b, B_{bds}^0, B_{bds}^1$	sont des effets aléatoires centrés de variances respectives σ_A^2, $\sigma_{B^0}^2$ et $\sigma_{B^1}^2$, et de covariance entre B_{bds}^0 et B_{bds}^1 $\sigma_{B^0B^1}$

Cette modélisation des effets aléatoires a pour conséquence de modéliser l'hétérogénéité de la variance des observations en fonction de la circonférence de l'arbre. Elle introduit également une corrélation entre les observations effectuées sur un même arbre plusieurs années, et une corrélation entre les arbres d'un même bloc. Précisément, on obtient :

$$\begin{aligned}
\mathrm{E}(F_{bdsa}) &= \gamma_a^0 + \gamma_a^1 C_{bdsa} \\
\mathrm{var}(F_{bdsa}) &= \sigma_A^2 + \sigma_{B^0}^2 + \sigma_{B^1}^2 C_{bdsa}^2 + 2\sigma_{B^0B^1} C_{bdsa} + \sigma^2 \\
\mathrm{cov}(F_{bdsa}, F_{bdsa'}) &= \sigma_A^2 + \sigma_{B^0}^2 + \sigma_{B^1}^2 C_{bdsa} C_{bdsa'} + \sigma_{B^0B^1}(C_{bdsa} + C_{bdsa'}) \text{ si } a \neq a' \\
&\quad \text{(même arbre et années différentes)} \\
\mathrm{cov}(F_{bdsa}, F_{bd's'a}) &= \sigma_A^2 \text{ si } d \neq d' \text{ ou } (d = d' \text{ et } s \neq s') \text{ (même bloc et arbres différents)} \\
\mathrm{cov}(F_{bdsa}, F_{b'ds'a}) &= 0 \text{ si } b \neq b' \text{ (blocs différents)}
\end{aligned}$$

1.2.4 L'écriture matricielle

Supposons que les observations soient réparties en M classes (elles sont regroupées en essais dans l'exemple 1, et en blocs dans l'exemple 2), telles que les données de deux classes distinctes soient indépendantes. On note n_i le nombre d'observations dans chacune des classes i.

Un modèle linéaire mixte décrit le vecteur y_i des observations dans la classe i de la façon suivante (*Pinheiro & Bates*, 2000, p58) :

$$\begin{aligned}
y_i &= X_i\beta + Z_i b_i + \epsilon_i,\ i = 1, ..., M \\
b_i &\sim N(0, \Sigma_i) \\
\epsilon_i &\sim N(0, \sigma^2.I)
\end{aligned}$$

où :
- y_i est un vecteur à n_i composantes ;
- β est le vecteur des p paramètres inconnus associés aux effets fixes ;
- b_i est le vecteur aléatoire gaussien centré et de variance Σ_i, formé des p_i' coordonnées indépendantes associées aux effets aléatoires pour la classe i ;
- X_i est la **matrice d'incidence** connue de dimension $n_i \times p$ (supposée de plein rang pour simplifier) pour la partie fixe du modèle ;
- Z_i est la **matrice d'incidence** connue de dimension $n_i \times p_i'$ pour la partie aléatoire du modèle ;

Considérons l'exemple du réseau d'essais, et considérons la classe i correspondant à un essai infesté avec 12 unités : 6 parcelles recevant les 6 variétés de types initiales/modifiées.

Le modèle (1.1) inclut 5 facteurs à effets fixes et 2 facteurs à effets aléatoires. Les écritures du vecteur β et de la matrice d'incidence X_i étant immédiates mais relativement longues, nous nous limiterons à donner plus précisément le vecteur b_i de composantes les modalités des variables A_{je} (facteur essai) et $(\alpha A)_{fje}$ (interaction forme$\times$ essai), et la matrice Z_i pour l'essai i considéré.

$$b_i = \begin{pmatrix} A_{je} \\ \hline (\alpha A)_{1je} \\ (\alpha A)_{2je} \end{pmatrix} = \begin{pmatrix} \text{essai i} \\ \hline \text{variété initiale} \times \text{essai i} \\ \text{variété modifiée} \times \text{essai i} \end{pmatrix}$$

$$Z_i = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

La matrice de variance-covariance Σ_i de b_i est alors donnée par la formule suivante :

$$\Sigma_i = \begin{pmatrix} \sigma_A^2 & 0 & 0 \\ 0 & \sigma_B^2 & 0 \\ 0 & 0 & \sigma_B^2 \end{pmatrix}$$

– l'erreur résiduelle ϵ_i est indépendante de b_i, centrée et de variance $\sigma^2.I$, où I est la matrice identité d'ordre n_i. Pour effectuer des tests ou calculer des intervalles de confiance, on suppose de plus que les erreurs suivent toutes une distribution normale.

De la formule ci-dessus, on déduit facilement que $\mathrm{E}(y_i) = X_i\beta$, et on note également Ψ_i sa matrice de variance-covariance :

$$\Psi_i = Z_i \Sigma_i Z_i' + \sigma^2.I$$

Cette matrice doit être **symétrique** et **définie positive** (toutes ses valeurs propres doivent être strictement positives).

Les paramètres à estimer sont constitués des p paramètres du vecteur β pour les effets fixes, des q **paramètres de variance-covariance** ou **composantes de la variance** définissant les matrices Σ_i, et que l'on note sous forme d'un vecteur θ, et de σ^2 pour les effets aléatoires résiduels.

Chapitre 2

Méthodes d'analyse utilisées

2.1 L'ajustement du modèle

Les estimateurs de variance minimale des paramètres fixes du modèle s'obtiennent par la méthode des moindres carrés généralisés. Ces formules font cependant intervenir les valeurs des composantes de la variance, qui sont a priori inconnues et qu'il faut donc estimer à partir des observations. Pour cela, on utilise des méthodes d'estimation des paramètres adaptées aux modèles linéaires mixtes. Elles font appel à des techniques basées sur des fonctions de vraisemblance et reposant sur les hypothèses de normalité des vecteurs b_i et ϵ_i. On les appelle le maximum de vraisemblance et le maximum de vraisemblance restreint, en abrégé `ML` et `REML` (de l'anglais "*M*aximum *L*ikelihood" et "*RE*stricted *M*aximum *L*ikelihood").

Ces méthodes consistent à choisir comme estimateurs des paramètres inconnus les valeurs qui maximisent la **fonction cible**, qui correspond respectivement à la **fonction de log-vraisemblance** et à la **fonction de log-vraisemblance réduite**. L'optimisation de ces fonctions est réalisée par des algorithmes itératifs, par exemple l'*algorithme de Newton-Raphson*, qui nécessite, à chaque itération, le calcul du gradient de la fonction cible et de sa dérivée.

2.1.1 La méthode `ML`

La fonction de vraisemblance, notée L, associée au modèle est définie de la façon suivante :

$$L(\beta, \theta, \sigma^2|y) = p(y|\beta, \theta, \sigma^2),$$

où p est la densité de probabilité du vecteur y des observations, associée au modèle (1.1) et aux valeurs de paramètres β, θ, σ^2.

De façon intuitive, la vraisemblance peut être interprétée comme la probabilité d'obtenir l'échantillon observé, si les valeurs des paramètres inconnus sont égales à β, θ, σ^2. La méthode d'estimation `ML` consiste à rechercher les valeurs $\widehat{\beta}$, $\widehat{\theta}$, $\widehat{\sigma^2}$ qui maximisent simultanément la vraisemblance.

2.1.2 La méthode `REML`

Dans de nombreux cas cependant, les simulations montrent que les estimations du maximum de vraisemblance des paramètres θ ont tendance à sous-estimer les valeurs de ces paramètres (le maximum de vraisemblance est dit **biaisé**). Comme la vraie difficulté réside dans l'estimation des composantes de la variance, il s'agit en quelque sorte de "concentrer"

la vraisemblance sur cette estimation. De nombreux statisticiens et analystes ont alors recours à une méthode basée sur la vraisemblance dite restreinte ou résiduelle (*Pinheiro & Bates*, 2000, p75-76).

La fonction de vraisemblance restreinte, notée L_R, inclut un terme dépendant des effets fixes du modèle, et est définie de la façon suivante :

$$L_R(\theta, \sigma^2|y) = \int L(\beta, \theta, \sigma^2|y) d\beta$$

Ceci revient à supposer une distribution a priori uniforme sur les paramètres des effets fixes β, et à intégrer L par rapport à cette distribution. Maximiser la vraisemblance restreinte est par ailleurs équivalent à maximiser $L(\theta, \sigma^2|Qy)$, où Q est une matrice orthogonale à la matrice X des effets fixes et de rang maximal.

La méthode `REML` procède en deux étapes :

1. d'abord l'estimation des composantes de la variance (le terme "restreint" vient du fait que cette estimation utilise une partie restreinte des données, celle qui est orthogonale aux facteurs à effets fixes) ;
2. puis l'estimation des effets fixes par moindres carrés généralisés, en fixant les variances à leurs valeurs estimées dans la première étape : ce sont ces estimations que l'on appellera *moyennes ajustées*.

2.1.3 L'estimation des paramètres

Les méthodes `ML` et `REML` permettent ainsi d'obtenir les *estimateurs du maximum de vraisemblance* $\widehat{\sigma}^2$ et $\widehat{\theta}$. Ces estimateurs sont asymptotiquement sans biais, de distribution gaussienne dont la variance est donnée par l'*inverse de l'information de Fisher*. Les estimations de β et b_i s'obtiennent ensuite en résolvant les équations du modèle mixte (*Henderson*, 1984), à savoir :

$$\begin{pmatrix} X_i'X_i & X_i'Z_i \\ Z_i'X_i & Z_i'Z_i + \sigma^2\Sigma^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widetilde{b_i} \end{pmatrix} = \begin{pmatrix} X_i'y_i \\ Z_i'y_i \end{pmatrix}$$

On estime alors les effets fixes et aléatoires en utilisant les équations ci-dessus avec $\widehat{\sigma}^2$ et $\widehat{\Sigma}$ comme s'il s'agissait des vraies valeurs :

$$\begin{cases} \widehat{\beta} = (X_i'\widehat{\Psi_i}^{-1}X_i)^- X_i'\widehat{\Psi_i}^{-1}y_i \\ \widetilde{b_i} = \widehat{\Sigma}Z_i'\widehat{\Psi_i}^{-1}(y_i - X_i\widehat{\beta}) \end{cases}$$

La première équation est connue sous le nom d'*équation de Gauss Markov* : $\widehat{\beta}$ est l'estimateur optimal de matrice de variance-covariance asymptotique $(X_i'\Psi_i^{-1}X_i)^-$ parmi les estimateurs linéaires sans biais de β.

Notons que pour les effets aléatoires, qui ne sont pas des paramètres proprement dits, on parle souvent de "prédicteurs" des effets aléatoires b_i plutôt que d'estimateurs, et on note $\widetilde{b_i}$ plutôt que $\widehat{b_i}$. Les prédicteurs $\widetilde{b_i}$ sont obtenus en minimisant la somme de carrés des résidus : ils sont appelés *BLUP* (de l'anglais "*B*est *L*inear *U*nbiased *P*redictor").

Les résultats que nous présenterons sont basés sur les estimations du maximum de vraisemblance restreint pour l'exemple du maïs, et du maximum de vraisemblance pour l'exemple des chênes.

2.2 La validation du modèle : ajustés et résidus

2.2.1 Les ajustés du modèle

Les valeurs **ajustées** $\widetilde{y}_i$ et $\widehat{y}_i$ correspondent aux prédicteurs linéaires optimaux et sans biais des valeurs prévues dans la population.

Ils sont donnés par les estimations :

$$\begin{cases} \widetilde{y}_i = X_i\widehat{\beta} + Z_i\widetilde{b}_i \text{ de } E(y_i \mid b_i) = X_i\beta + Z_i b_i \\ \widehat{y}_i = X_i\widehat{\beta} \text{ de } E(y_i) = X_i\beta \end{cases}$$

La matrice de variance-covariance $\widetilde{\Psi}_i$ de $\widetilde{y}_i$ est définie par la formule suivante :

$$\widetilde{\Psi}_i = W_i \Phi_i W_i'$$

où W_i est la matrice $(X_i\, Z_i)$, et

$$\Phi_i = \frac{1}{\widehat{\sigma}^2} \begin{pmatrix} X_i'X_i & X_i'Z_i \\ Z_i'X_i & Z_i'Z_i + \widehat{\sigma}^2\widehat{\Sigma}^{-1} \end{pmatrix}^-$$

est la matrice de variance-covariance de $\begin{pmatrix} \widehat{\beta} \\ \widetilde{b}_i \end{pmatrix}$.

2.2.2 Les résidus du modèle

Les **résidus** $\widetilde{\epsilon}_i$ et $\widehat{\epsilon}_i$ associés à la classe i représentent les différences entre les valeurs observées et les valeurs ajustées du modèle. Ils sont donnés par les formules suivantes :

$$\begin{cases} \widetilde{\epsilon}_i = y_i - \widetilde{y}_i \\ \widehat{\epsilon}_i = y_i - \widehat{y}_i \end{cases}$$

et dépendent des erreurs aléatoires, des erreurs de mesures, et des erreurs dues à un mauvais choix de modèle.

2.2.3 Les résidus standardisés du modèle

Les **résidus standardisés** $\widetilde{\epsilon}_i{}^*$ sont le rapport des résidus $\widetilde{\epsilon}_i$ et des écart-types résiduels estimés de $\widetilde{y}_i$, qui correspondent aux racines carrées des termes diagonaux de la matrice de variance-covariance $\widetilde{\Psi}_i$ de $\widetilde{y}_i$:

$$\widetilde{\epsilon}_i{}^* = \frac{y_i - \widetilde{y}_i}{\sqrt{\widetilde{\Psi}_i}}$$

Remarque 2 *Avec* R*, les résidus standardisés sont de la forme* $\frac{y_i - \widetilde{y}_i}{\widehat{\sigma}}$*, car, conditionnellement à* $\widetilde{b}_i$*,* $\mathrm{E}(y_i|b_i) = \mathrm{E}(\widetilde{y}_i|b_i) = X_i\beta + Z_i b_i$ *et* $\mathrm{var}(\widetilde{y}_i|b_i) \approx \widehat{\sigma}^2$.

2.2.4 Les représentations graphiques des résidus

L'analyse de variance doit nécessairement comporter une analyse des résidus à travers une étude graphique, qui permet de valider ou non le modèle retenu.

Parmi les différents graphiques possibles, il est indispensable de représenter les résidus standardisés $\widetilde{\epsilon_i}^*$ en fonction des ajustés $\widehat{y}_i$ ou $\widetilde{y}_i$. En particulier, les résidus permettent de détecter des valeurs aberrantes et de vérifier s'il n'existe pas de sous ou sur-estimation systématique d'un groupe d'observations, ce qui nous amènerait à modifier le modèle. On cherche également à contrôler la variance des résidus: si elle croît avec les ajustés, une transformation permettant de stabiliser la variance est nécessaire. L'absence de structure évidente permet d'accepter le modèle.

Il est recommandé de représenter également les résidus en fonction des principaux facteurs ou covariables de l'essai afin de vérifier que la distribution des résidus est raisonnablement homogène entre les différents niveaux de ces variables. Pour les facteurs (qualitatifs), ces graphiques peuvent être effectués sous la forme de **boîte à moustaches** ou **boxplots**.

Pour vérifier que la distribution des erreurs ne s'écarte pas excessivement d'une loi normale, on utilise soit un histogramme des résidus (réduits de préférence), soit un graphique des quantiles empiriques des résidus en fonction des quantiles théoriques d'une loi normale.

En principe, les mêmes vérifications devraient être effectuées sur les prédicteurs des effets aléatoires de chaque terme aléatoire du modèle. Ces vérifications sont souvent négligées, car en pratique, pour un nombre fini d'observations, on accorde moins d'importance à ce que les variables b_i soient rigoureusement symétriques et normalement distribuées.

2.3 Comparaison de modèles et tests d'hypothèses

En pratique, l'analyse d'un jeu de données dans le cadre du modèle mixte se décompose très souvent en :

1. une étape de choix et de validation du modèle ;
2. une étape d'interprétation des résultats obtenus avec le modèle retenu.

Pour la première étape, le point de départ est le modèle issu de la réflexion sur les facteurs et covariables du jeu de données, comme nous l'avons vu au Chapitre 1. Cependant, cette étape laisse souvent place à plusieurs options sur des termes non essentiels à l'interprétation mais pouvant avoir un effet sur les observations : faut-il retenir tel facteur, telle covariable, ou telle interaction ? Les *tests de modèles emboîtés* et les *critères d'Akaike* décrits ci-dessous offrent un support pour trancher entre ces options. Cela n'exclut pas une part de choix subjectifs, dont il ne faut toutefois pas abuser.
Par ailleurs, le choix du modèle doit également s'appuyer sur l'étude des résidus, afin de garantir que le modèle décrit convenablement les données, en particulier pour leur partie "non expliquée" par le modèle. Il est indispensable qu'une étude des résidus soit effectuée sur le modèle qui sera retenu, mais il est aussi recommandé d'en effectuer le plus tôt possible dans la phase de choix de modèle.

Au cours de l'étape d'interprétation du modèle retenu, l'enjeu des tests n'est plus de choisir un cadre d'analyse, mais de répondre aux questions scientifiques pour lesquelles les données ont été recueillies. Les hypothèses à tester peuvent alors porter sur l'ensemble des

effets associés à un terme du modèle, mais aussi sur des fonctions plus précises de ces effets. Les tests d'hypothèse, comme le *test de Wald* pour la **significativité** des effets aléatoires, et les *tests de Student* et *de Fisher* pour la significativité des effets fixes, peuvent alors être utilement complétés par des calculs d'*intervalles de confiance*, des *comparaisons de moyennes* et des *études de contrastes*.

Nous présentons d'abord les principales méthodes pour comparer des modèles et tester des hypothèses, puis nous proposerons des éléments pour choisir entre ces méthodes.

2.3.1 Le test de rapport de vraisemblance pour comparer deux modèles emboîtés

Un modèle statistique est dit **emboîté** dans un autre s'il constitue un sous-modèle de celui-ci, c'est-à-dire s'il s'en déduit par une contrainte linéaire sur les paramètres. Les *tests de rapport de vraisemblance* permettent de comparer des modèles emboîtés, en testant la significativité de la contrainte linéaire qui les différencie.

◇ *Pour les effets fixes*

La statistique du test de rapport de vraisemblance repose sur l'écart entre les vraisemblances maximisées obtenues sous chacun des modèles. Si $\widehat{L}$ et $\widehat{L'}$ sont respectivement les vraisemblances maximisées sous le modèle général et sous le modèle réduit, alors la statistique LRT (de l'anglais "*L*ikelihood *R*atio *T*est") du test de rapport de vraisemblance est définie par :

$$LRT = 2\log(\widehat{L}/\widehat{L'}) = 2[\log(\widehat{L}) - \log(\widehat{L'})]$$

Si k' et k sont les nombres des paramètres estimés sous le modèle général et sous le modèle réduit, alors, sous l'hypothèse que le modèle réduit soit correct, la statistique LRT est approchée par une loi du χ^2 à $k' - k$ degrés de liberté.

Pour tester que le modèle général est significativement différent du modèle réduit, on calcule donc la probabilité qu'une loi du χ^2 à $k' - k$ degrés de liberté dépasse la valeur observée de LRT. Si la probabilité est inférieure au niveau choisi pour le test (par exemple 5%), la différence est significative et il faut donc conserver le modèle général. Sinon, on peut envisager de ne retenir que le modèle réduit.

◇ *Pour les composantes de la variance*

Pour tester les paramètres de variance-covariance d'un modèle, on peut construire un test similaire en comparant les vraisemblances. Si les deux modèles ne diffèrent que par les effets aléatoires, alors on peut également utiliser des statistiques basées sur la vraisemblance restreinte. Mais les *Likelihood Ratio Test* ne suivent pas les χ^2 attendus et les tests du χ^2 ont tendance à être conservatifs (on ne rejette pas l'hypothèse nulle aussi souvent que nécessaire) (*Pinheiro & Bates*, 2000, p83-84).

2.3.2 Le test de Wald pour tester la présence significative d'un effet aléatoire

Afin de tester la significativité des composantes de la variance θ, on peut avoir recours au *test de Wald*, qui est basé sur la loi asymptotique de l'estimateur de θ : lorsque le nombre

d'observations est grand, la loi de $\widehat{\theta}$ est approchée par une loi gaussienne d'espérance θ et de matrice de variance-covariance Ω , où $\Omega/2$ est l'inverse de la *matrice hessienne* (matrice des dérivées secondes) de la fonction de vraisemblance calculée en θ. La statistique du test de l'hypothèse $\theta_1 = 0$ par exemple correspond au rapport du paramètre de variance-covariance et de son écart-type estimé, soit $Z = \widehat{\theta}_1/s_1$, où s_1 est la racine carrée du terme $\widehat{\Omega}_{11}$. La loi de Z est approchée par une loi gaussienne centrée réduite.

Remarque 3 *SAS et R permettent de construire des intervalles de confiance respectivement pour les paramètres de variance-covariance et pour les écarts-types des effets aléatoires.*

2.3.3 Les tests de Student et de Fisher pour tester la présence significative d'un ou plusieurs effets fixes

Lorsque deux modèles emboîtés diffèrent par leur partie fixe, le test de rapport de vraisemblance ne peut être défini qu'en utilisant le maximum de vraisemblance. Même si la statistique *LRT* peut être calculée pour ces modèles, *Pinheiro* et *Bates* (2000) ne la recommandent pas car il a été démontré que ce test tendait à être anti-conservatif (on rejette trop souvent l'hypothèse nulle) (*Pinheiro* & *Bates*, 2000, p87-88). Nous préférerons alors les tests d'hypothèses basés sur les *statistiques de Student* (T) lorsqu'il s'agit d'estimer des coefficients individuels, et plus généralement sur les *statistiques de Fisher* (F) lorsqu'il s'agit d'estimer des combinaisons linéaires de coefficients.

Ces tests amènent à considérer des combinaisons linéaires estimables de la forme $K\beta$, et les hypothèses testées sont de la forme :

$$K\beta = 0$$

◇ *Le test de Student*

Lorsque l'on cherche à étudier la significativité individuelle de chaque paramètre ou d'une fonction de paramètres du modèle (autrement dit, K est un vecteur d'une ligne et $K\beta$ est une combinaison linéaire des paramètres), on construit la statistique T en divisant l'estimation par son écart-type estimé sous l'hypothèse considérée :

$$T = \frac{K\widehat{\beta}}{\sqrt{K\widehat{\mathrm{var}}(\widehat{\beta})K'}}$$

Sous les hypothèses de normalité des b_i et ϵ_i, la loi de T est approchée par une *loi de Student*, dont le degré de liberté $\widehat{\nu}$ est calculé selon plusieurs méthodes possibles (voir plus bas).

◇ *Le test de Fisher*

Lorsque l'on cherche à étudier la significativité globale de plusieurs paramètres ou fonctions de paramètres (autrement dit, K est une matrice de rang q supérieur à 1), on

construit alors la statistique F suivante, dont la distribution est approchée par une *loi de Fisher* de degrés de liberté le rang de K pour le numérateur et $\widehat{\nu}$ pour le dénominateur :

$$F = \frac{(K\widehat{\beta})'(K\widehat{\mathrm{var}}(\widehat{\beta})K')^{-1}K\widehat{\beta}}{\mathrm{rank}(K)}$$

Calcul de $\widehat{\nu}$:

La variance de $K\widehat{\beta}$, égale à $K\widehat{\mathrm{var}}(\widehat{\beta})K'$, est une fonction linéaire connue des composantes de la variance θ et de σ^2, qui, en revanche, sont inconnus et doivent donc être estimés. Lorsque cette variance ne dépend que de la variance résiduelle σ^2, $\widehat{\nu}$ est le nombre de degrés de liberté résiduels. Lorqu'elle dépend d'autres paramètres de variance, il n'existe pas d'expression exacte pour $\widehat{\nu}$, et plusieurs approches sont possibles pour en déterminer une valeur raisonnable. La plus fréquente est exposée dans *Pinheiro & Bates* (2000, p91).

Lorsque, en pratique, on teste les différents termes fixes d'un modèle à plusieurs facteurs ou covariables, on distingue deux types de tests de Fisher :

le test de Fisher "séquentiel" teste l'ajout successif des termes du modèle. On parle alors de *sommes de carrés de type* I : la somme de carrés de type I d'un terme donné est **ajustée** pour tous les termes précédents du modèle.

le test de Fisher "marginal" teste la nullité d'un sous-ensemble des paramètres du modèle contre le modèle complet. On parle alors de *sommes de carrés de type* III : la somme de carrés de type III d'un terme donné est ajustée pour tous les autres termes du modèle et ne dépend donc pas de l'ordre des termes dans le modèle.

Lorsque l'analyse de variance est **orthogonale**, c'est-à-dire lorsque le jeu de données est **équirépété** ou **équilibré**), les tests de Fisher de types I et III sont équivalents. Dans le cas contraire, il est conseillé d'utiliser les tests de type III.

Les T et F-statistiques permettent de construire des tests d'hypothèses relatifs aux effets fixes d'un modèle mixte. Elles figurent dans les tableaux d'analyse de variance avec la p-valeur associée au test : une p-valeur plus petite qu'une valeur seuil traduira un effet significatif.

2.3.4 Le critère d'Akaike

Le *critère d'Akaike* est un critère de choix de modèles basé sur le principe de parcimonie. Il est défini de la façon suivante :

$$AIC = -2\widehat{L} + 2n_{par}$$

où $\widehat{L}$ est le maximum de la log-vraisemblance et n_{par} est le nombre de paramètres intervenant dans le modèle.

Lorsque l'on souhaite comparer deux modèles ou plus, il suffit de calculer leur critère AIC et de retenir le modèle pour lequel le critère est le plus petit.

Le critère AIC est connu pour avoir tendance à retenir des modèles surparamétrés lorsque le nombre de données n'est pas suffisamment grand. Il en existe plusieurs variantes

qui corrigent plus ou moins ce défaut, mais nous ne les détaillerons pas ici.

Remarque 4 *Si les modèles à comparer ont la même partie fixe, le critère d'Akaike peut être construit à partir du maximum de la log-vraisemblance restreinte.*

Remarque 5 *La méthode REML étant la méthode par défaut sous SAS, les modèles comparés doivent alors avoir la même partie fixe.*

2.3.5 En résumé...

1. Pour tester un terme fixe, ou de manière équivalente, pour comparer deux modèles emboîtés qui diffèrent uniquement par leur partie fixe, on peut utiliser :
   - le test de rapport de vraisemblance,
   - le test de Student (si on teste un seul degré de liberté) ou de Fisher (si on teste un nombre de degrés de liberté > 1) ;

   Selon *Pinheiro & Bates* (2000, p87-88), nous conseillons d'utiliser la seconde possibilité.

2. Pour tester un terme aléatoire, ou de manière équivalente, pour comparer deux modèles emboîtés qui diffèrent uniquement par leur partie aléatoire, il n'existe pas de méthode précise et unique. Nous pourrons utiliser les intervalles de confiance basés sur le test de Wald. Mais si vraiment le test des paramètres de la variance est un des objectifs de l'étude, il faut consulter les articles comme *Self & Liang* (1987). Une autre façon de tester la significativité des variances des effets aléatoires est par le calcul des intervalles de confiance comme décrit dans *Pinheiro & Bates* (2000, p92-93).

3. Pour comparer des modèles non nécessairement emboîtés l'un dans l'autre, on peut utiliser :
   - le critère d'Akaike basé sur la vraisemblance,
   - le critère d'Akaike basé sur la vraisemblance restreinte seulement si les modèles ont la même partie fixe.

2.4 L'exploitation du modèle : intervalles de confiance, moyennes ajustées et méthode des contrastes

Pour exploiter les résultats de l'analyse, on calcule des intervalles de confiance et des statistiques de test pour diverses fonctions linéaires des paramètres et des effets ajustés. Nous précisons quelques-unes de ces techniques ci-dessous. Les méthodes de calcul sont toutes similaires à celles exposées pour tester $K\widehat{\beta}$.

2.4.1 Les intervalles de confiance

Les *intervalles de confiance* pour les effets fixes et les effets aléatoires permettent d'obtenir une information plus fiable sur les paramètres que leur simple estimation.

Ils sont construits à partir des estimations du maximum de vraisemblance et des T-statistiques décrites ci-dessus de la manière suivante, pour une combinaison linéaire I des paramètres et effets aléatoires :

$$I \begin{pmatrix} \widehat{\beta} \\ \widetilde{b_i} \end{pmatrix} \pm t_{\widehat{\nu}_I, \alpha/2} \sqrt{I \Phi_i I'}$$

où $t_{\widehat{\nu}_I, \alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la *loi de Student*.

Remarque 6 *En général, on construit les intervalles de confiance pour les paramètres de variance-covariance des effets aléatoires.*

2.4.2 Les moyennes ajustées et les comparaisons multiples

Après avoir rejeté l'hypothèse d'égalité de l'ensemble des moyennes d'un ou plusieurs facteurs, lorsque la p-valeur est inférieure au risque α choisi, il est utile de poursuivre l'analyse en calculant les *moyennes estimées* ou *moyennes ajustées* pour chaque modalité des facteurs fixes, et en recherchant lesquelles sont différentes et lesquelles ne le sont pas.

Les moyennes ajustées diffèrent des moyennes brutes dans le cas où le jeu de données n'est pas orthogonal. Chacune d'entre elles est donnée par la formule $M\widehat{\beta}$, où M est la matrice des coefficients associés aux modalités des facteurs testés, et a pour matrice de variance-covariance la matrice $M\widehat{\mathrm{var}}(\widehat{\beta})M'$.

Il existe également différentes méthodes permettant de comparer ces moyennes deux à deux, et de tester si les différences entre moyennes pour des modalités d'un facteur ou d'une interaction de facteurs sont significatives ou non.

2.4.3 La méthode des contrastes

La *méthode des contrastes* permet d'obtenir des estimations de moyennes des facteurs et des fonctions linéaires des paramètres du modèle. Dans notre premier exemple, nous cherchons à savoir si, en moyenne, les variétés modifiées sont meilleures que les variétés initiales.

L'interprétation d'un contraste se fait en construisant un test d'hypothèse de la forme :

$$C \begin{pmatrix} \beta \\ b_i \end{pmatrix} = 0,$$

et en utilisant la statistique F suivante, dont la loi est approchée par une *loi de Fisher* de degrés de liberté le rang de L et $\widehat{\nu}$:

$$F = \frac{\begin{pmatrix} \widehat{\beta} \\ \widetilde{b_i} \end{pmatrix}' C'(C\Phi_i C')^{-1} C \begin{pmatrix} \widehat{\beta} \\ \widetilde{b_i} \end{pmatrix}}{\mathrm{rank}(C)}$$

Chapitre 3

Mise en oeuvre

Cette partie présente l'analyse de variance et l'interprétation des résultats d'un modèle jugé pertinent pour chacun des exemples introduits précédemment. Nous ne détaillerons donc pas toute la démarche de choix de modèle. Ces analyses ont été réalisées en utilisant deux outils différents : le logiciel `SAS` pour l'exemple du maïs d'une part, puis le logiciel `R` pour l'exemple des chênes. Les représentations graphiques se feront principalement sous `R`, car ce logiciel permet une grande souplesse pour les représentations graphiques.

La démarche proposée pour interpréter les résultats d'un dispositif en blocs complets randomisés est la suivante :

- analyse de variance, permettant notamment de vérifier le contrôle de l'hétérogénéité ;
- avec le modèle mixte, estimation des paramètres de variance ou des écarts-types des effets aléatoires, et tests de Fisher qui permettent de quantifier la significativité des facteurs ;
- examen des résidus ;
- estimation des moyennes ajustées pour les facteurs fixes ;
- éventuellement estimation et tests des effets des facteurs fixes et aléatoires.

L'analyse de variance, des moyennes ajustées et des contrastes sont traitées dans l'exemple du maïs, et nous avons choisi d'insister plus particulièrement sur la question du choix du modèle dans l'exemple des arbres.

3.1 Exemple du rendement des variétés de maïs transgéniques avec `SAS`

3.1.1 L'ajustement du modèle

Considérons le modèle défini en (1.1) et réalisons les analyses avec `SAS` en utilisant la méthode `REML` présentée en (2.1.2).
Pour cela, nous utiliserons les notations suivantes : *rdt* pour la variable explicative rendement, et *forme*, *denom*, *indice*, *essai* respectivement pour les facteurs forme variétale, dénomination de la variété (fonds génétique), indice d'infestation et essai.

Le rendement est analysé selon le modèle abrégé suivant :

$$rdt = forme + denom + indice + essai + indice \times forme + indice \times denom + essai \times forme$$

◇ ***Les commandes*** `SAS`

Le système `SAS` est un ensemble de modules logiciels pour la gestion et le traitement statistique des données. Un **programme** `SAS` est un enchaînement d'**étapes** de gestion de données (les **data**) et d'appels de **procédures**, décrivant, dans une syntaxe souvent spécifique à chaque **module**, les traitements à réaliser sous le couvert d'**options** prises par défaut ou explicitement définies.

Le modèle linéaire est traité dans `SAS` essentiellement par les procédures `anova`, `reg`, `glm` et `mixed` pour, respectivement, les analyses de variance orthogonales, les régressions, les analyses de variance non orthogonales ainsi que les analyses de covariance et les modèles linéaires généralisés, et les modèles linéaires mixtes.

Pour chacune de ces procédures, l'instruction `MODEL Y=X` permet de définir le modèle : `Y` est la réponse analysée et `X` est un ensemble de termes (fixes) pour lesquels les paramètres sont estimés.

Remarquons que la notation $*$ peut être utilisée pour décrire des facteurs emboîtés ou des facteurs croisés. Lorsque l'on est en présence d'un facteur hiérarchisé, il faut prendre soin de le déclarer avec la syntaxe propre au logiciel que l'on utilise. En effet, il ne faut pas définir un effet propre du facteur : la décomposition est différente de celle du modèle croisé dans le sens où ce qui serait l'effet principal du facteur hiérarchisé est incorporé à l'interaction.

Ainsi, $[Y = A\ B\ A*B]$ décrit un modèle avec deux facteurs croisés A et B, alors que $[Y = A\ A*B]$ décrit un modèle avec B emboîté dans A, d'où l'absence du terme B qui n'a pas de signification isolé de A.

⋆ La procédure `glm`

Remarque 7 *Le modèle que nous analysons est un modèle linéaire mixte. En pratique, nous recommandons néanmoins de commencer par une procédure* ***`glm`*** *qui donne notamment un tableau complet de l'analyse de variance, alors que la procédure* ***`mixed`*** *ne construit ce tableau que pour les effets fixes.*

Dans un premier temps, le modèle est analysé à l'aide de la procédure `glm` et des directives suivantes :

```
PROC GLM DATA=tablesas ;
CLASS forme denom indice essai ;
MODEL rdt=
  forme forme*denom
  indice indice*essai
  indice*forme indice*forme*denom
  indice*essai*forme ;
WHERE annee=2000 ;
RUN ;
```

Interprétons ce programme :

- la première ligne déclare que la procédure va travailler sur la table *tablesas* ;
- la ligne `class` déclare que *forme*, *denom*, *indice* et *essai* sont des facteurs qualitatifs ;
- la ligne `model` déclare la variable à expliquer *rdt* et le modèle standard d'analyse de la variance à quatre facteurs à effets fixes, dont les facteurs emboîtés *denom* (`forme*denom`) et *essai* (`indice*essai`), avec interactions (`indice*forme`, `indice*forme*denom` et `indice*essai*forme`) ;
- la ligne `where` sélectionne les données relatives à l'année 2000 dans le tableau de données.

⋆ La procédure `mixed`

Nous reprenons l'exemple précédent, mais plus raisonnablement, nous supposons que les essais ont été échantillonnés parmi les trois populations définissant les niveaux d'infestation du réseau d'essais. Nous voulons donc faire de l'inférence au niveau de la population et non plus des individus. Le modèle approprié est alors le modèle linéaire mixte : nous utilisons la `proc mixed`.

Comme dans la `proc glm`, on déclare d'abord les variables qualitatives. La seule différence est que les effets aléatoires, le facteur *essai* et son interaction avec la forme variétale (*indice*×*essai*×*forme*), sont déclarés dans une ligne à part : la ligne `random`. En voici la programmation :

```
PROC MIXED DATA=tablesas METHOD=REML COVTEST CL ;
CLASS forme denom indice essai ;
MODEL rdt=
  forme forme*denom
  indice indice*forme
  indice*forme denom / OUTP=analyse ;
RANDOM indice*essai indice*essai*forme ;
WHERE annee=2000 ;
RUN;
```

Précisons les options à spécifier :

- Les options `method=`, `covtest` et `cl` de la procédure `mixed` permettent respectivement de spécifier la méthode utilisée dans l'estimation des paramètres de variance-covariance, de demander leur écart-type estimé et leur test de Wald, et de construire leurs intervalles de confiance (en effet, il peut être intéressant de calculer les intervalles de confiance de ces estimations : si l'intervalle est grand et que l'estimation $\widehat{\theta}$ est proche de 0, alors l'effet du facteur aléatoire est inexistant et on peut le retirer du modèle).
- L'option `outp=` de `model` réalise une sortie des résultats de l'analyse sur la table `SAS` *analyse* : elle contient les données initiales, les valeurs observées, les valeurs prédites avec leur écart-type estimé et les résidus.
- L'option `outpm=` a le même rôle, mais les valeurs prédites et les résidus sont calculés

en tenant compte des effets fixes seulement.

◇ *Les résultats des procédures* `glm` *et* `mixed` *et leurs interprétations*

Dans toute la suite, nous choisissons de donner les résultats avec une précision de deux chiffres après la virgule, et de 3 chiffres pour les p-valeurs.

⋆ L'analyse de variance avec la procédure `GLM`

Les tableaux issus de la procédure `glm` sont utiles tout d'abord pour vérifier que les degrés de liberté des différents termes sont compatibles avec ce que l'on attend. Le tableau de type III permet ensuite, en comparant les valeurs des statistiques F de Fisher, d'identifier rapidement les termes du modèle qui ont le plus d'influence sur la variabilité des données. Par contre, il ne faut pas utiliser ce tableau pour tester les termes du modèle, car les tests sont tous effectués par rapport à la seule variance résiduelle, ce qui n'est pas valide pour les termes qui emboîtent des termes aléatoires. Nous éviterons donc de porter attention à la colonne des probabilités.

Rappelons qu'un tableau d'analyse de variance est constitué de colonnes donnant respectivement :
- le *nombre de degrés de libertés* de chaque terme du modèle ;
- sa *somme de carrés*, c'est-à-dire la variabilité sur les données associées à ce terme (elle ne figurera pas dans nos résultats) ;
- son *carré moyen*, qui correspond à la somme de carrés sur le nombre de degrés de liberté ;
- la valeur de la statistique du F de Fisher, qui correspond au carré moyen du terme sur le carré moyen résiduel ;
- une valeur de probabilité du F observé, sous l'hypothèse d'absence d'effets de ce terme.

Pour une valeur donnée du carré moyen d'un terme du modèle, la valeur du F est d'autant plus grande que le carré moyen résiduel est petit, c'est-à-dire que l'erreur est petite. Travailler à minimiser l'erreur, c'est améliorer la capacité de l'expérience à détecter les effets des facteurs.

Le jeu de données n'étant pas équilibré, il existe plusieurs tests de Fisher pour les effets fixes : nous présentons ci-dessous les tableaux d'analyse de la variance de types I (Tableau 3.1) et III (Tableau 3.2) que la procédure `glm` permet d'obtenir.

Le principe du test de Fisher séquentiel repose sur des tests d'ajout successif des paramètres. Par exemple, pour le facteur *forme*, nous testons si un écart du à la fome variétale sur le terme constant du modèle est significatif.

Contrairement au test précécent, il n'y a pas de notion d'ordre sur les paramètres dans le test de Fisher marginal. Par exemple, pour la le facteur *indice*, nous testons la nullité de l'effet marginal du facteur dans le modèle complet.

Type 1 Tests of Fixed Effects				
Source	**Num DF**	**Mean Square**	**F Value**	**Pr>F**
forme	3	100.61	1.74	0.163
forme*denom	14	239.54	4.14	$< 10^{-4}$
indice	2	2068.86	35.79	$< 10^{-4}$
indice*essai	9	2351.27	40.67	$< 10^{-4}$
forme*indice	4	160.65	2.78	0.031
forme*denom*indice	24	93.98	1.63	0.049
forme*indice*essai	17	31.06	0.54	0.928

TAB. 3.1 – *Procédure* `glm` *: Tests de Fisher de type* I *pour les effets fixes.*

Type 3 Tests of Fixed Effects				
Source	**Num DF**	**Mean Square**	**F Value**	**Pr>F**
forme	3	329.16	5.69	0.001
forme*denom	14	144.86	2.51	0.004
indice	2	1734.60	30.01	$< 10^{-4}$
indice*essai	9	1574.84	27.24	$< 10^{-4}$
forme*indice	4	159.23	2.75	0.032
forme*denom*indice	24	94.00	1.63	0.049
forme*indice*essai	17	31.06	0.54	0.928

TAB. 3.2 – *Procédure* `glm` *: Tests de Fisher de type* III *pour les effets fixes.*

Remarque 8 *1. on a huit variétés initiales, huit variétés modifiées, une variété nouvelle et un témoin, donc le nombre de degrés de liberté total des deux premières lignes est égal à* $(8+8+1+1)-1=17$ *;*

2. de même, d'après le Tableau 1.1, le nombre de degrés de liberté pour le facteur essai est égal à $(7+4+1)-(1+2)=9$ *;*

3. nous sommes dans un cas un peu particulier : par exemple, le facteur forme n'est pas complètement croisé avec le facteur indice car la variété nouvelle et le témoin ne sont pas présents dans les essais de niveau d'infestation 3 : on a alors un nombre de degrés de liberté égal à $2\times 3+2\times 2-(1+3+2)=4$ *au lieu de* $3\times 2=6$ *;*

4. le rang de la matrice X_i *est donné par le total des nombres de degrés de liberté du Tableau 3.1 corrigé de 1, soit* $73+1=74$*.*

La variance résiduelle (57.81) des observations est estimée par le carré moyen résiduel, et l'**écart-type résiduel** (7.60) en est la racine carrée. Exprimé dans la même unité que la variable analysée et variant généralement de 3 à 8 quintaux/hectares, il traduit la précision des résultats et constitue une estimation globale de l'erreur expérimentale.

Dans notre exemple, il n'y a pas orthogonalité entre tous les facteurs. La somme de carrés de chaque terme du modèle doit donc être ajustée par rapport aux autres termes du modèle : on préfèrera alors utiliser les sommes de carrées de type III.

Nous en déduisons que :

- la valeur du F pour le facteur *forme* est passé de 1.74 à 5.69 entre les deux analyses : la décomposition en type III augmente souvent la **puissance** de l'essai, c'est-à-dire sa capacité à mettre en évidence les différences entre formes variétales ;
- l'effet du niveau d'infestation intervient fortement à travers les facteurs *indice* et *indice*×*essai* ;

- les interactions *forme*×*indice* et *forme*×*denom*×*indice* ne sont pas très significatives : les différences entre fonds génétiques d'une même forme variétale et entre formes variétales varient peu selon l'indice d'infestation ;
- la faible valeur du F de l'effet du fonds génétique (*forme*×*denom*) traduit une faible variabilité entre génotypes de même forme ;
- l'interaction entre la forme et l'essai (*forme*×*indice*×*essai*) n'est pas significative : pour un même niveau d'infestation, les différences entre formes variétales ne dépendent pas de l'essai ;

- l'essai reste hétérogène (écart-type résiduel= 7.60 q/ha), malgré le contrôle de forts effets de l'indice d'infestation ($F = 30.01$) et de l'essai ($F = 27.24$).

⋆ L'ESTIMATION PAR `REML` ET LES TESTS DE FISHER AVEC LA PROCÉDURE `MIXED`

L'estimation des composantes de la variance :

La première étape de l'analyse est l'estimation des paramètres de variance par `REML` (Tableau 3.3), indispensables pour ensuite estimer les effets des facteurs fixes.

Covariance Parameter Estimates						
Covariance Parameter	**Estimate**	**Standard Error**	**Z Value**	**Pr>|Z|**	**Lower**	**Upper**
indice*essai	155.50	75.18	2.07	0.019	72.44	539.61
forme*indice*essai	0.00	.	.	.	.	.
Residual	54.12	6.90	7.84	$< 10^{-4}$	42.78	70.67

TAB. 3.3 – *Procédure* `mixed` *: Tests de Wald pour les composantes de la variance.*

Le premier paramètre à regarder est l'écart-type résiduel de l'essai : il est passé de 7.60 à $\sqrt{54.12} = 7.36$ quintaux/hectare entre les analyses de la `proc glm` et de la `proc mixed` : ceci traduit une précision plutôt faible des essais, mais légèrement meilleure.

Nous observons que les valeurs estimées sont redondantes avec les informations fournies par l'analyse de variance avec le modèle fixe.

En particulier :

- le test de Wald montre que les effets aléatoires entre essais (*indice*×*essai*) sont significatifs, ce qui traduit une grande variabilité entre essais, et l'écart-type estimé pour ce facteur, égal à $\sqrt{155.50} = 12.50$, est très supérieur à l'écart-type résiduel ;
- par contre, l'interaction entre formes variétales et essais (*forme*×*indice*×*essai*) n'est pas significative.

Remarque 9 *La nullité de la composante traduit la nullité de l'effet : l'interaction forme×indice×essai peut donc être retirée du modèle.*

Les tests de Fisher pour les facteurs fixes :

Le tableau d'analyse de variance de type III issu de la procédure `mixed` (Tableau 3.4) permet de tester les effets fixes du modèle :

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr>F**
forme	3	17	6.64	0.004
forme*denom	14	106	2.67	0.002
indice	2	9	1.22	0.339
forme*indice	4	17	2.96	0.050
forme*denom*indice	24	106	1.73	0.031

TAB. 3.4 – *Procédure* `mixed` *: Tests de Fisher de type* III *pour les effets fixes.*

Remarque 10
1. *le nombre de degrés de liberté du dénominateur est adapté à chaque terme du modèle : par exemple, le facteur forme est emboîté dans l'interaction* `forme*indice*essai` *du Tableau 3.3, dont le nombre de degrés de liberté du numérateur vaut 17 dans le Tableau 3.2 ;*
2. *le degré de liberté résiduel est égal à la différence du nombre d'observations et du rang de la matrice* X_i*, soit* $180 - 74 = 106$.

Nous déduisons du Tableau 3.4 les résultats suivants :

- cette analyse souligne davantage l'effet du transgène et les différences entre formes variétales puisque la valeur du F pour le facteur *forme* est passée de 5.69 à 6.64 entre les analyses de la `proc glm` et de la `proc mixed` ;
- la présence de l'effet du fonds génétique (*forme*×*denom*) est mis en évidence mais dans une moindre mesure ;
- l'interaction *forme*×*indice* est significative : les différentes formes variétales se comportent différemment selon le niveau d'attaques des pyrales ;

- l'effet du niveau d'infestation n'intervient pas à travers le facteur *indice*, ce qui peut être du à la forte variabilité entre essais (*indice*×*essai*) ;

- l'interaction *forme*×*denom*×*indice* est très peu significative : on distingue peu le fonds génétique à l'intérieur d'une forme variétale et pour un niveau d'infestation donné.

Remarque 11 *Seul l'effet du niveau d'infestation n'est pas significatif ($P = 0.339 > 0.05$): il faut cependant conserver ce terme dans le modèle puisqu'il intervient dans des interactions significatives*

En conclusion, l'analyse de variance par le modèle (1.1) montre que :

- les effets variétaux, *forme* et *forme*×*denom*, sont significatifs dans l'ensemble des essais ;
- l'effet du niveau d'infestation n'intervient pas à travers le facteur *indice* ($P = 0.339$) mais à travers le facteur emboîté *indice*×*essai*, ce qui traduit une forte variabilité entre essais ;
- l'interaction Génotype×Environnement est mise en évidence à travers la significativité des termes :
  - *forme*×*denom*×*indice* : la petite valeur de la statistique de Fisher ($F = 1.73$) révèle cependant une faible variabilité des différents fonds génétiques à l'intérieur d'une forme variétale et pour un certain niveau d'infestation,
  - *forme*×*indice*, ce qui souligne un comportement différent des formes variétales selon le niveau d'infestation.

3.1.2 L'analyse des résidus

◇ ***Les commandes*** `SAS` ***et*** `R`

`R` offre une variété de graphiques remarquable. Les nombreuses options disponibles rendant la production de graphiques extrêmement flexible, nous avons choisi d'utiliser R pour nos représentations graphiques. Le package nommé `base` est en quelque sorte le coeur de `R` et contient les fonctions de base du langage pour la lecture et la manipulation des données, des fonctions graphiques...

L'option `outp` permet tout d'abord d'obtenir, dans la table `SAS` *analyse*, les résidus standardisés en divisant les résidus par leur écart-type estimé.

Il s'agit ensuite de convertir cette table en fichier texte (ASCII) en utilisant la procédure suivante dans l'éditeur `SAS` :

```
PROC EXPORT DATA= analyse
  OUTFILE= "~/fichier.txt"
  DBMS=DLM REPLACE ;
DELIMITER='00'x;
RUN;
```

`R` peut alors lire des données stockées dans des fichiers texte à l'aide la fonction `read.table`, qui a pour effet de créer un **data.frame** et est donc le moyen principal pour lire des tableaux de données, et en tracer des représentations graphiques :

```
tableR <- READ.TABLE("fichier.txt")
```

◇ ***Les résultats graphiques : les résidus standardisés en fonction des ajustés***

Les graphiques de la Figure 3.1 représentent les résidus standardisés en fonction des ajustés, en distinguant les formes initiales et les formes modifiées. Par ce type de graphiques, nous cherchons à valider notre modèle, en examinant si les résidus ne présentent pas de structure évidente.

Nous y avons également identifié, par des numéros, les variétés pour lesquelles nous obtenons des résidus standardisés supérieurs à 3 en valeur absolue.

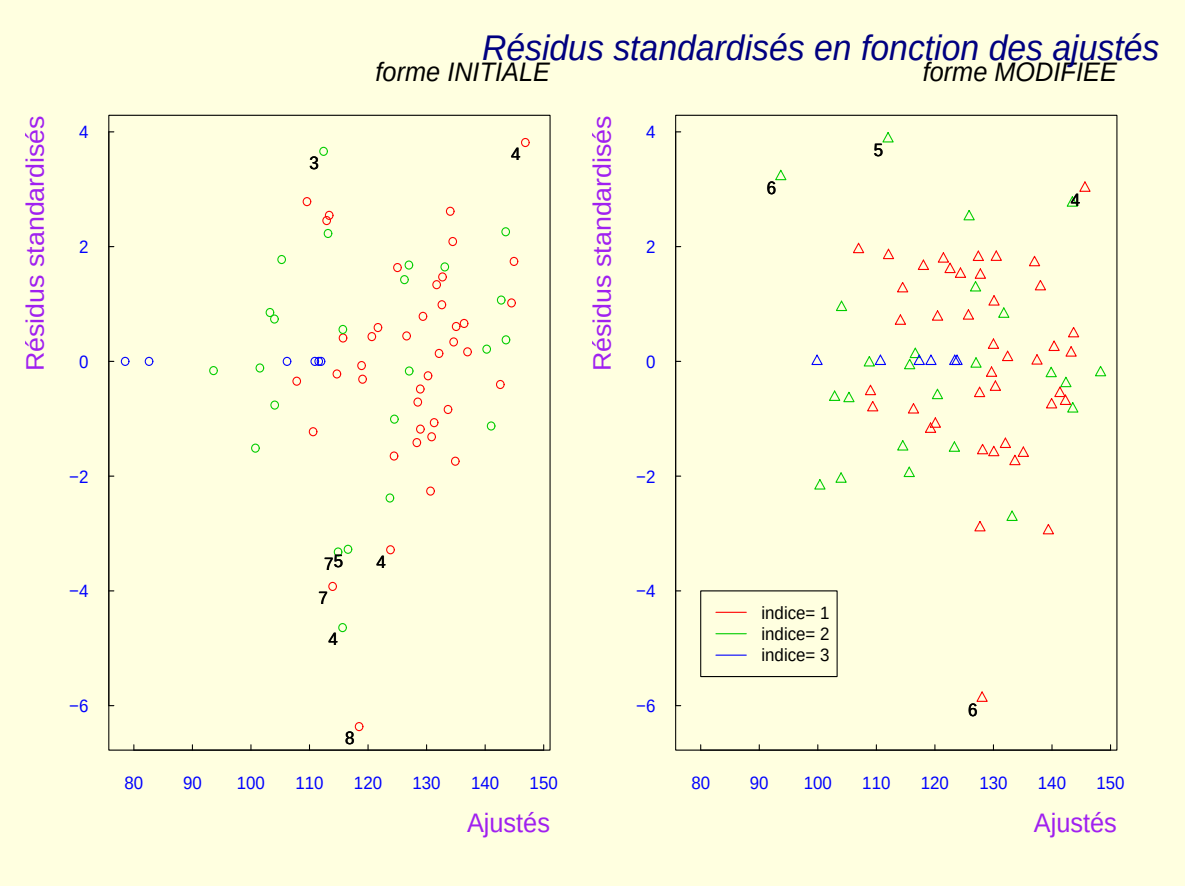

FIG. 3.1 – *Résidus standardisés, identifiés par couleur et représentés par forme variétale, en fonction des ajustés.*

Remarquons qu'en 2000, un seul essai est indicé du niveau d'infestation 3, et, du fait que le modèle comprenne une interaction fixe $forme \times denom \times indice$, les résidus sont nuls pour les données de cet essai (ils sont représentés par des cercles et triangles bleus alignés sur la droite d'ordonnée nulle).

Nous ne considérons pas les variétés de résidus élevés comme des points aberrants et choisissons de ne pas les retirer du jeu de données. Aucune structure n'est mise en évidence, ce qui nous permet de valider le modèle (1.1) retenu.

3.1.3 Les estimations des moyennes et les tests sur les contrastes pour les effets fixes

Estimer les effets des variétés permet de classer celles-ci selon leurs performances. Pour interpréter ce classement, il est nécessaire de savoir si les différences observées sont significatives ou si elles sont essentiellement dues à la variabilité des observations.

◇ ***Les commandes*** `SAS` ***et*** `R`

Dans le cas où les facteurs ont plus de deux niveaux, et dans le cas où leurs effets sont significatifs, une comparaison des moyennes est nécessaire.

Le calcul des moyennes brutes est demandé, dans le cas équilibré, par l'instruction `means`, et celui des moyennes ajustées, dans le cas deséquilibré, par l'instruction `lsmeans`. Les tests sur les contrastes se font en ajoutant à la syntaxe précédente les lignes `estimate`.

Il est possible de récupérer les estimations des moyennes et les tests sur les contrastes obtenus sous `SAS` dans des tableaux `.sas`.

```
LSMEANS forme*indice /CL ;
LSMEANS forme*denom /CL ;
LSMEANS forme*denom*indice /CL ;
ESTIMATE 'modifiée-initiale indice=1'
  forme -1 1 0 0
  forme*indice -1 0 0 1 0 0 /CL ;
ESTIMATE 'modifiée-initiale indice=2'
  forme -1 1 0 0
  forme*indice 0 -1 0 0 1 0 /CL ;
ESTIMATE 'modifiée-initiale indice=3'
  forme -1 1 0 0
  forme*indice 0 0 -1 0 0 1 /CL ;
ESTIMATE 'OGM-non OGM indice=1'
  forme -1 1 1 -1
  forme*indice -1 0 0 1 0 0 1 0 0 -1 0 0 /CL ;
ESTIMATE 'OGM-non OGM indice=2'
  forme -1 1 1 -1
  forme*indice 0 -1 0 0 1 0 0 1 0 0 -1 0 /CL ;
ESTIMATE 'OGM-non OGM indice=3'
  forme -1 1 0 0
  forme*indice 0 0 -1 0 0 1 /CL ;
ODS OUTPUT LSMEANS=moyennes ESTIMATES=contrastes ;
```

- Les lignes `lsmeans` demandent explicitement les moyennes ajustées (qui ne sont pas données par défaut) pour les facteurs *forme∗indice*, *forme∗denom* et *forme∗denom∗indice*, et les écrivent dans la table *moyennes*.
- Les lignes `estimate` demandent l'estimation des contrastes entre formes modifiée/initiale d'une variété pour chaque valeur de l'indice d'infestation, puis entre formes transgéniques/non transgénique.
- L'option `CL` permet de construire les intervalles de confiance des estimations (de niveau 0.95 par défaut).

– La ligne `ods` permet l' exportation des moyennes ajustées et des tests sur les contrastes de la fenêtre `output` vers les tables `SAS` respectives *moyennes* et *contrastes*.

Remarque 12 *Ce que l'on souhaite en général, c'est tester les différences entre moyennes : l'option `pdiff` de l'instruction `lsmeans` permet d'obtenir des comparaisons deux par deux.*

On répète ensuite les mêmes procédures que précédemment pour rendre les données disponibles sous `R` et en permettre le tracé des graphiques : on convertit les tables *moyennes* et *contrastes* en fichiers texte et on utilise la fonction `read.table`.

◇ ***Les résultats et leurs interprétations***

Avant de parler de contrastes, il est préférable de parler de moyennes ajustées : les contrastes consistent en général à regarder les choses plus finement, notamment des combinaisons linéaires de ces moyennes.

⋆ L'ANALYSE DES MOYENNES

Le tableau des `lsmeans` (Tableau 3.5) permet de comparer les moyennes ajustées par `REML` des différentes formes variétales pour chaque niveau d'infestation (*forme*×*indice*). Les autres moyennes, moins intéressantes car faisant intervenir les différents génotypes, sont illustrées par les représentations graphiques.

Remarque 13 *Comme nous l'avons vu pour le tracé de la Figure 1.1, afin de permettre l'estimation des moyennes, nous ne considérerons à partir d'ici que les six variétés de types initiales/modifiées présentes dans les essais protégés et infestés, la variété nouvelle et le témoin.*

Least Squares Means						
Effect	**forme**	**indice**	**Estimate**	**Standard Error**	**Lower**	**Upper**
forme*indice	initiale	1	127.44	4.82	117.28	137.60
forme*indice	initiale	2	119.73	6.37	106.29	133.17
forme*indice	initiale	3	100.28	12.74	73.40	127.17
forme*indice	modifiée	1	128.17	4.82	118.01	138.33
forme*indice	modifiée	2	120.82	6.37	107.37	134.26
forme*indice	modifiée	3	115.73	12.74	88.85	142.61
forme*indice	nouvelle	1	122.21	6.08	109.38	135.05
forme*indice	nouvelle	2	113.26	8.29	95.77	130.75
forme*indice	témoin	1	116.95	6.08	104.11	129.78
forme*indice	témoin	2	115.46	8.29	97.97	132.95

TAB. 3.5 – *Estimation des moyennes ajustées.*

Les résultats du Tableau 3.5 nous permettent de confirmer de quelle façon le niveau de rendement dépend du niveau d'infestation : plus les pyrales attaquent et plus le rendement moyen diminue.

Nous présentons dans la suite trois types de graphiques illustrant les différentes moyennes ajustées.

La Figure 3.2, tout d'abord, reprend les données du Tableau 3.5. Ceci nous permet d'observer le comportement global des quatre formes variétales pour chaque valeur de l'indice d'infestation.

Nous y avons également tracé les intervalles de confiance pour chaque moyenne, et avons relié les données d'une même forme variétale entre elles.

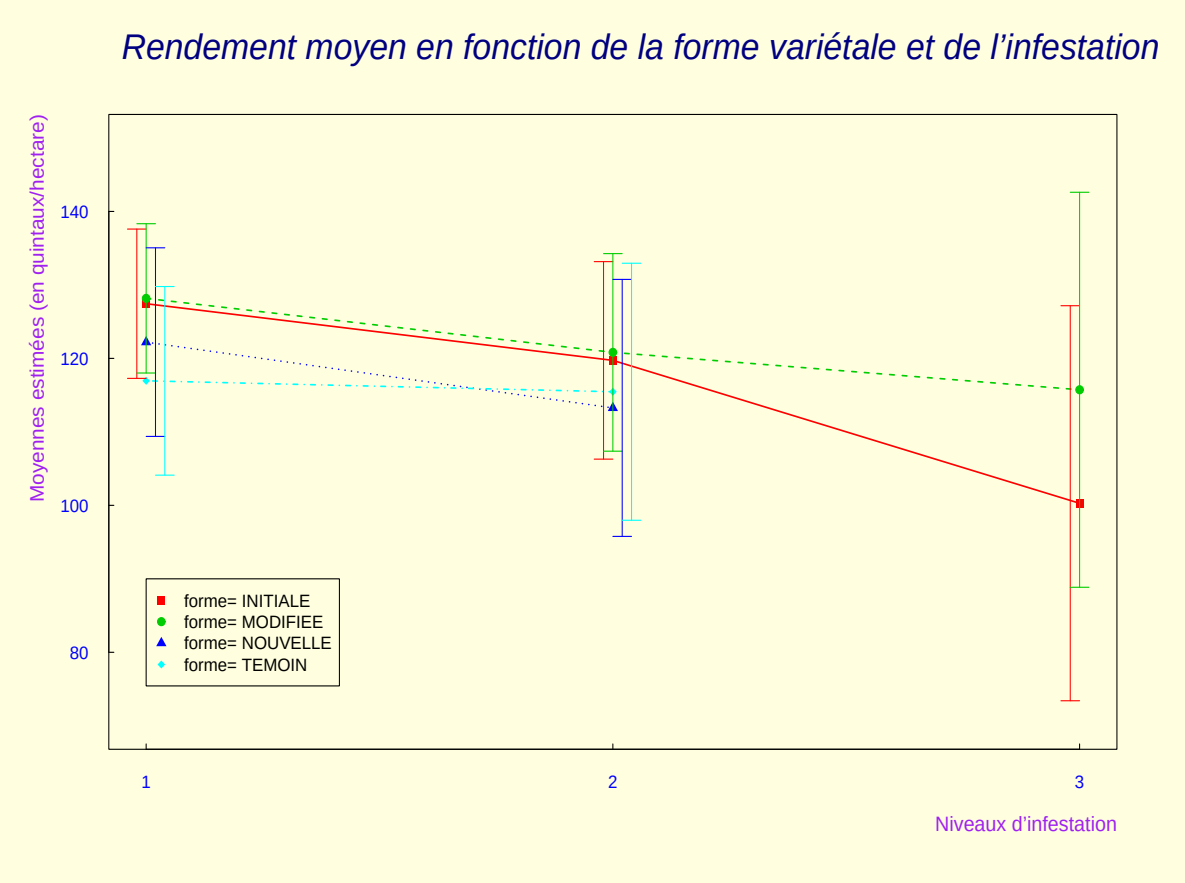

FIG. 3.2 – *Moyennes estimées du rendement, identifiées par couleur et par forme, en fonction du niveau d'infestation.*

L'année 2000 met parfaitement en évidence les trois aspects suivants :

- le rendement moyen diminue avec le niveau d'infestation pour l'ensemble des formes variétales ;
- les formes modifiées restent supérieures aux formes initiales en présence ou non de pyrales ;
- les formes modifiées et initiales se comportent de façon quasi-identique en essais pas ou peu infestés et l'écart de rendement s'amplifie pour le niveau d'infestation 3.

Les Figures 3.3 et 3.4 illustrent, elles, les moyennes ajustées obtenues pour chaque fonds génétique (au nombre de 14) en fonction du niveau d'infestation toujours, et en distinguant pour chacune des six variétés sa forme initiale et sa forme modifiée. Ces graphiques nous permettent notamment d'apprécier les différences de rendement moyen entre variétés, mais également entre formes issues d'une même variété.

Nous avons également relié les valeurs des moyennes pour les formes initiales d'une part, et les formes modifiées d'autre part.

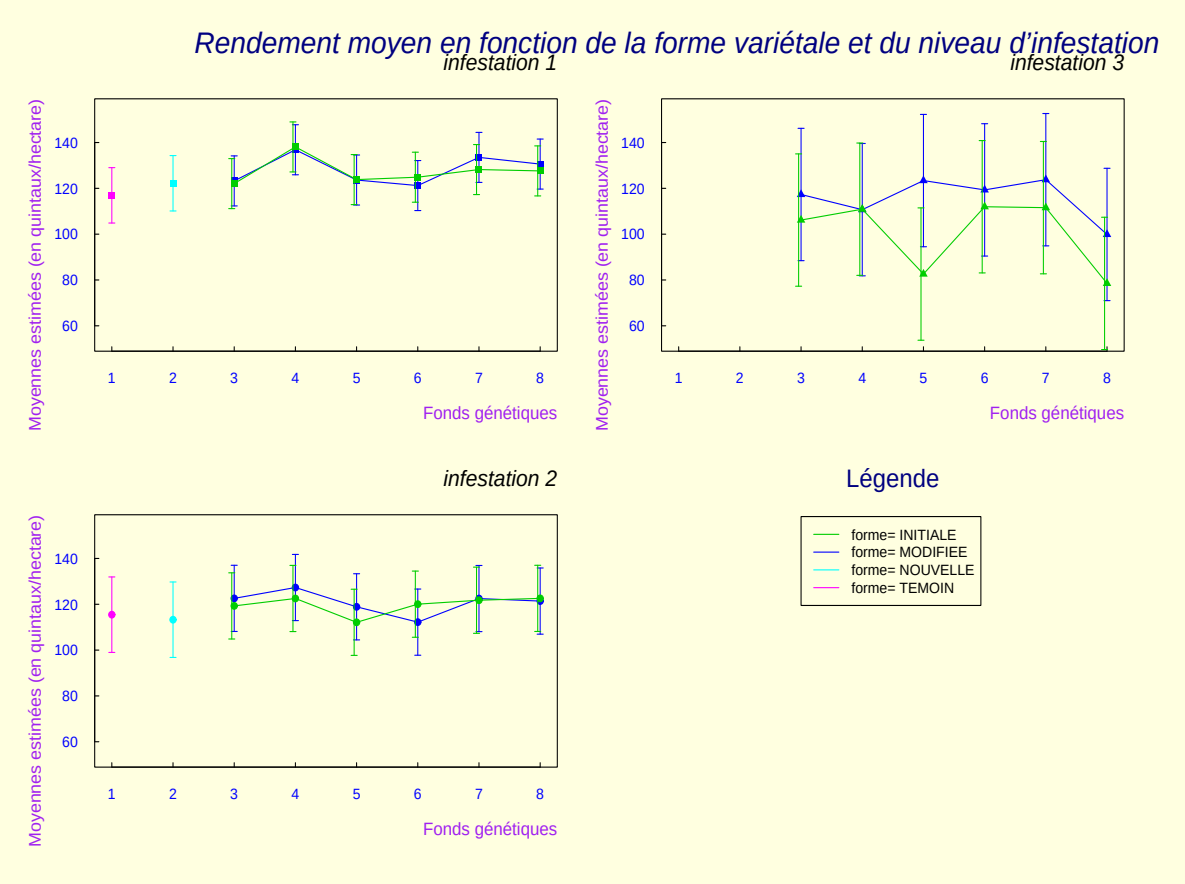

FIG. 3.3 – *Moyennes estimées du rendement, identifiées par couleur et par forme et représentées par niveau d'infestation, en fonction du fonds génétique.*

Les graphiques de la Figure 3.3 soulignent le fait que l'écart de rendement entre les formes initiales et modifiées est d'autant plus prononcé que le niveau d'infestation des pyrales est important.

La seule variété, pourtant, pour laquelle cet écart semble être significatif, d'après l'option `pdiff` de `lsmeans`, est la variété 5 du graphique 3.

La Figure 3.4 est la superposition des graphiques de la Figure 3.3.

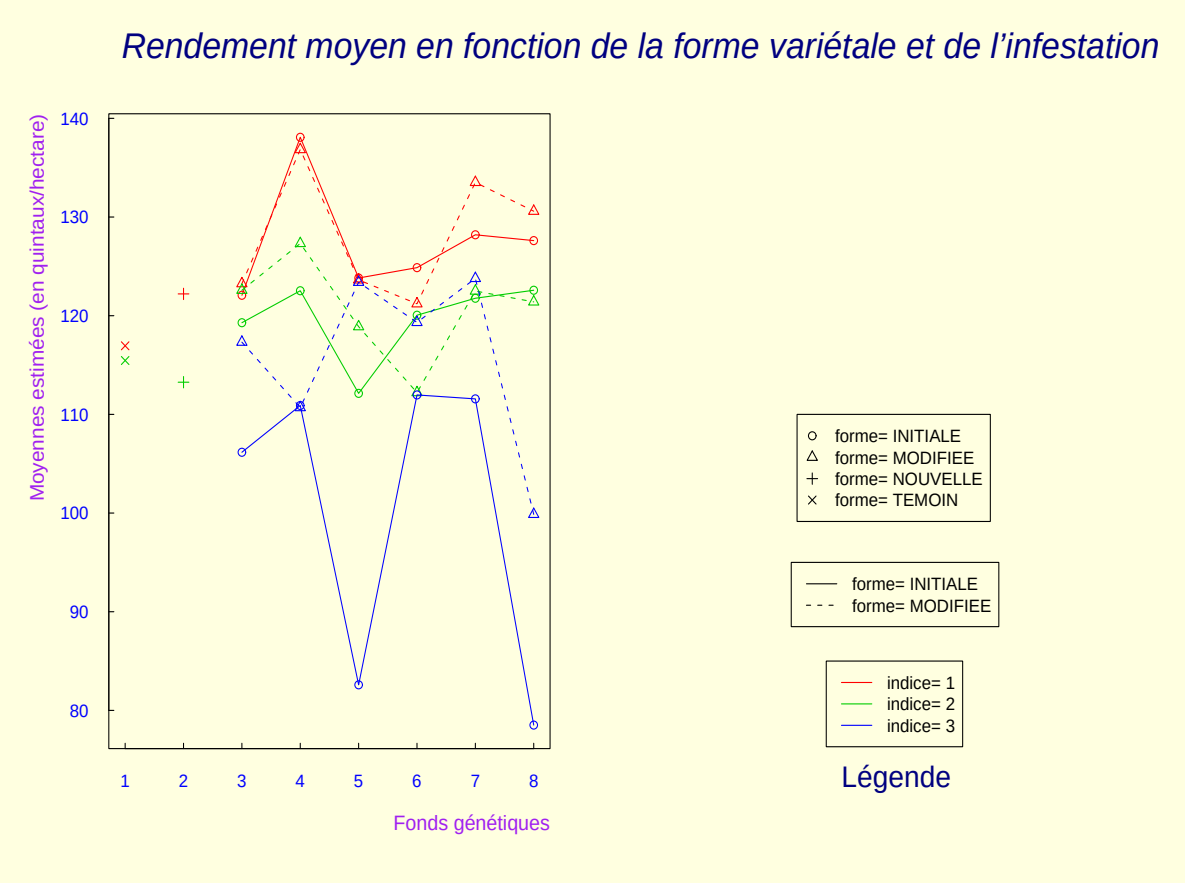

FIG. 3.4 – *Moyennes estimées du rendement, identifiées par couleur, forme et tracé, en fonction du fonds génétique.*

En conclusion :

- les différences de rendement moyen entre formes initiales et modifiées varient selon la valeur de l'indice d'infestation : on observe un écart de rendement s'amplifiant avec le niveau d'attaque des pyrales ;
- la seule différence significative entre les deux formes semble être obtenue pour la variété 5 en présence massive de pyrales.

Le Tableau 3.6 donne les estimations des contrastes pour les différences entre formes initiales et modifiées, et plus généralement entre variétés transgéniques et non-transgéniques, pour chaque valeur de l'indice d'infestation. Comme pour l'estimation des moyennes (Tableau 3.5), nous n'avons pas présenté les résultats concernant les différents fonds génétiques dans le tableau suivant.

Estimates						
Label	**Estimate**	**Standard Error**	**T Value**	**Pr > |T|**	**Lower**	**Upper**
modifiée-initiale indice=1	0.73	1.67	0.44	0.668	-2.80	4.26
modifiée-initiale indice=2	1.09	2.21	0.49	0.629	-3.58	5.76
modifiée-initiale indice=3	15.45	4.43	3.49	0.003	6.11	24.79
OGM-non OGM indice=1	3.00	2.84	1.06	0.305	-2.99	8.99
OGM-non OGM indice=2	-0.55	4.00	-0.14	0.891	-8.97	7.87
OGM-non OGM indice=3	15.45	4.43	3.49	0.003	6.11	24.79

TAB. 3.6 – *Estimation des contrastes.*

Remarque 14 *Nous obtenons les mêmes résultats pour les différences entre variétés initiales et modifiées et celles entre variétés transgéniques et non-transgéniques pour l'indice 3, la variété nouvelle et le témoin n'étant pas étudiés en essais indicés du niveau 3.*

Le Tableau 3.6 nous permet de déduire que les écarts entre formes initiales et modifiées s'accentuent avec le niveau d'infestation :

- l'écart entre formes initiales et modifiées est proche de 0 dans les essais protégés (rappelons que ces essais sont mis en place afin de vérifier le caractère équivalent des deux formes) ;
- les différences sont significatives en essais fortement infestés ($P = 0.003$).

La Figure 3.5 illustre les écarts de rendement moyen pour les six variétés de types initiales/modifiées et pour chaque niveau d'infestation.

La droite horizontale d'ordonnée 0 représente l'écart nul entre formes initiales et modifiées.

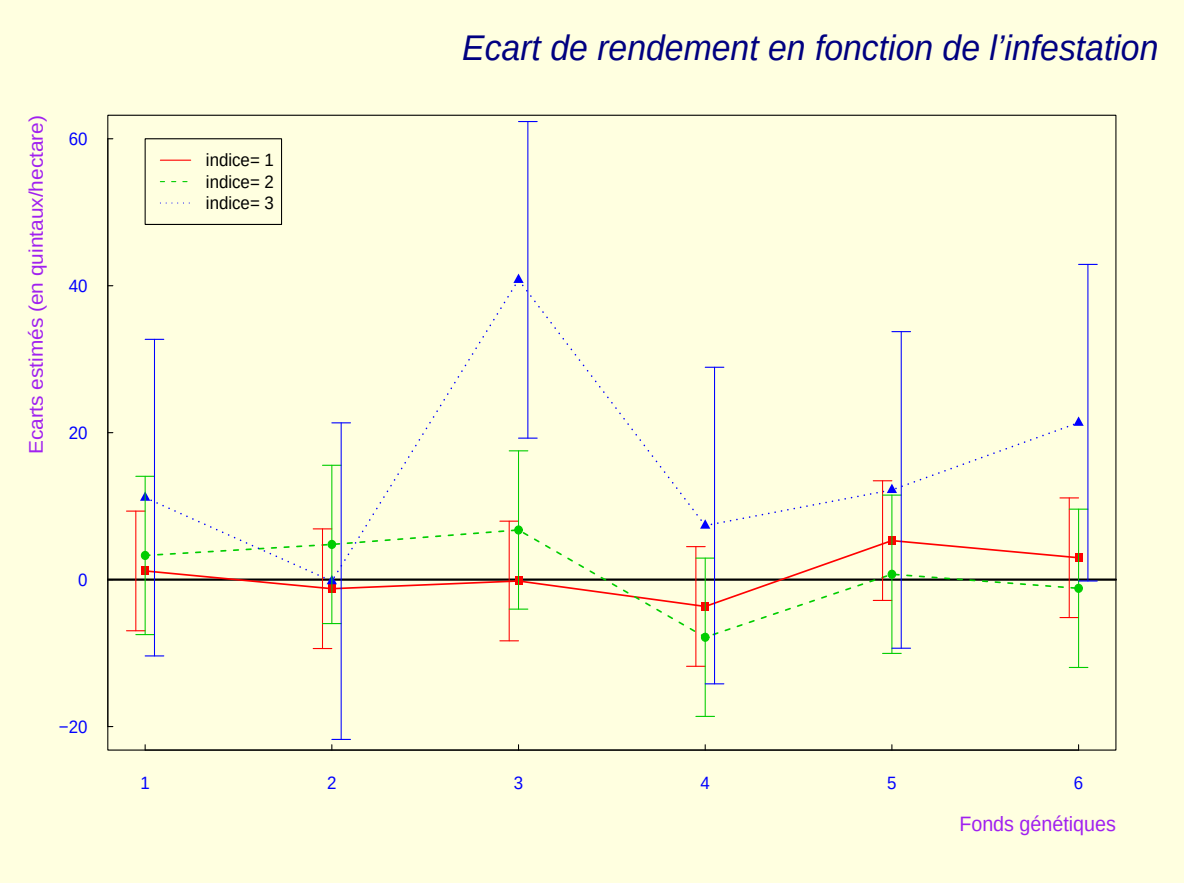

FIG. 3.5 – *Ecarts de rendement entre formes modifiées et initiales, identifiés par couleur et par tracé, en fonction du fonds génétique.*

Nous observons que :

- le tracé des intervalles de confiance nous permet de souligner que la variabilité des moyennes de rendement s'amplifie avec le niveau d'infestation ;
- le caractère significatif de l'écart entre les formes transgénique et non transgénique de la variété que nous avions souligné dans l'analyse des moyennes est confirmé pour l'indice 3 de la variété 3 ($P = 0.0003$) ;
- tous les autres contrastes sont non-significatifs ($P > 0.05$).

A travers la méthode des contrastes, nous avons pu quantifier la présence de l'effet du transgène et compléter l'interprétation du modèle en mettant en évidence l'effet du niveau d'infestation : nous avons observé des écarts entre formes initiales et modifiées en présence de pyrales de l'ordre de 15 quintaux/hectare par rapport à une moyenne générale d'environ 108 quintaux/hectare.

3.2 Exemple de la fourchaison des chênes avec R

3.2.1 L'ajustement du modèle

Considérons le modèle défini en (1.2) et réalisons les analyses avec R en utilisant la méthode ML présentée en (2.1.1). Pour cela, nous utiliserons les notations suivantes : *nf* pour la variable explicative nombre de fourches, *circ* pour la covariable circonférence, et *bloc*, *sujet*, *annee* respectivement pour les facteurs bloc, arbre et âge de l'arbre.

La fourchaison des chênes est analysée selon le modèle suivant :

$$nf = annee \times circ + bloc + sujet + sujet \times circ$$

◇ ***Les commandes*** R

Le package base contient également certaines fonctions statistiques, notamment celles qui concernent les modèles linéaires et les analyses de variance.

Il y a trois fonctions statistiques principales dans le package base : aov, glm et lme pour, respectivement, les analyses de variance, les modèles linéaires généralisés, et les modèles linéaires mixtes. Nous utiliserons donc la fonction lme.

⋆ la fonction lme() :

- l'argument principal et obligatoire de lme() est la formule qui précise la réponse à gauche du signe ∼ et le prédicteur à droite,
- l'option data= précise que les variables doivent être prises dans le data.frame *tableR*,
- l'option random=list donne une **liste** des effets aléatoires du modèle,
- l'option method= spécifie la méthode à utiliser dans l'estimation des paramètres de variance-covariance,
- les résultats ne sont pas affichés car ceux-ci sont copiés dans un objet nommé *analyse* ;

⋆ certaines fonctions permettent ensuite d'extraire les résultats désirés :

- la fonction anova() pour afficher les tests de significativité pour les effets fixes :
  - l'option type= "marginal" permet d'obtenir les tests de Fisher marginaux (par défaut, R calcule les tests de Fisher séquentiels)
  - la fonction anova () avec plusieurs arguments permet de comparer des modèles par le calcul des critères AIC,
- la fonction intervals() pour afficher les intervalles de confiance des écarts-types des effets aléatoires,
- la fonction print() pour afficher un bref résumé de l'analyse (essentiellement les paramètres estimés),
- la fonction summary() pour afficher plus de détails dont les tests statistiques.

Les formules sont un élément-clef des analyses statistiques avec R. La notation utilisée est la même pour presque toutes les fonctions : une formule est typiquement de la forme `y ~ model`, où `y` est la réponse analysée et `model` est un ensemble de termes pour lesquels les paramètres sont estimés. Ces termes sont séparés par des symboles arithmétiques mais qui ont une signification particulière.

En voici un résumé :

`a+b`	effets additifs
`a:b`	effet interactif entre `a` et `b`
`a*b`	effets additifs et interactifs (identique à `a+b+a:b`)
`a-b`	supprime l'effet de `b`
`1`	ajuste un modèle sans effets (juste l'`intercept`)

Sous R, les facteurs qualitatifs sont déclarés comme tels dans le tableau de données, à l'aide de la commande `is.factor`, et non dans l'analyse.

⋆ MODÈLE 1 : $F_{bdsa} = \gamma_a^0 + \gamma_a^1 C_{bdsa} + A_b + B_{bds}^0 + B_{bds}^1 C_{bdsa} + \epsilon_{bdsa}$

Considérons d'abord le modèle présenté en 1.2.3.

```
analyse1 <- LME(nf ~ annee*circ, DATA=tableR,
  RANDOM=LIST(bloc=~1, sujet=~circ),METHOD="ML")
ANOVA(analyse1)
INTERVALS(analyse1)
```

Les effets aléatoires se décomposent en un effet du *bloc* sur la moyenne générale (`bloc=~1`), et un effet de l'*arbre* sur la moyenne générale du modèle et sur la relation entre le nombre de fourches et la circonférence de l'arbre (`sujet=~circ`).

⋆ MODÈLE 2 : $F_{bdsa} = \gamma_a^0 + \gamma_a^1 C_{bdsa} + A_b + B_{bds}^1 C_{bdsa} + \epsilon_{bdsa}$

Reprenons à présent le modèle précédent, sans prise en compte de l'effet de l'arbre sur la moyenne générale du modèle.

```
analyse2 <- LME(nf ~ annee*circ, DATA=tableR,
  RANDOM=LIST(bloc=~1, sujet=~circ-1),METHOD="ML")
ANOVA(analyse2)
INTERVALS(analyse2)
```

Les effets aléatoires se décomposent ici uniquement en un effet du *bloc* (`bloc=~1`) et un effet de l'arbre (`sujet=~circ-1`) indépendemment de l'année.

Remarque 15 *L'effet de la densité est pris en compte au travers de la vigueur de l'arbre : nous considérons alors une relation linéaire sans effet de la densité propre, et la circonférence, qui représente la vigueur de l'arbre, est influencée par son âge.*

◇ ***Les résultats des fonctions*** `anova` ***et*** `intervals` ***et leurs interprétations***

Dans toute la suite, nous choisissons de donner les résultats avec une précision de deux chiffres après la virgule, de 3 chiffres pour les p-valeurs et de 4 chiffres pour les écarts-types des effets aléatoires.

⋆ Les tests de Fisher pour les effets fixes

Les tableaux suivants présentent l'analyse de variance pour les deux modèles définis ci-dessus : les tests de Fisher séquentiels sont donnés par les Tableaux 3.7 et 3.9, et les tests de Fisher marginaux par les Tableaux 3.8 et 3.10.

Type 1 Tests of Fixed Effects				
	Num DF	**Den DF**	**F Value**	**p Value**
(intercept)	1	1481	669.87	$< 10^{-4}$
annee	3	1481	353.55	$< 10^{-4}$
circ	1	1481	287.36	$< 10^{-4}$
annee : circ	3	1481	6.18	$< 10^{-4}$

Tab. 3.7 – *Modèle 1 : Tests de Fisher de type* I *pour les effets fixes.*

Type 3 Tests of Fixed Effects				
	Num DF	**Den DF**	**F Value**	**p Value**
(intercept)	1	1481	12.70	$< 10^{-4}$
annee	3	1481	0.06	0.981
circ	1	1481	67.89	$< 10^{-4}$
annee : circ	3	1481	6.18	$< 10^{-4}$

Tab. 3.8 – *Modèle 1 : Tests de Fisher de type* III *pour les effets fixes.*

Type 1 Tests of Fixed Effects				
	Num DF	**Den DF**	**F Value**	**p Value**
(intercept)	1	1481	671.16	$< 10^{-4}$
annee	3	1481	353.56	$< 10^{-4}$
circ	1	1481	287.34	$< 10^{-4}$
annee : circ	3	1481	6.18	$< 10^{-4}$

Tab. 3.9 – *Modèle 2 : Tests de Fisher de type* I *pour les effets fixes.*

Type 3 Tests of Fixed Effects				
	Num DF	Den DF	F Value	p Value
(intercept)	1	1481	12.71	$< 10^{-4}$
annee	3	1481	0.06	0.981
circ	1	1481	67.89	$< 10^{-4}$
annee : circ	3	1481	6.18	$< 10^{-4}$

TAB. 3.10 – *Modèle 2 : Tests de Fisher de type* III *pour les effets fixes.*

Remarque 16 *1. la ligne* ***intercept*** *teste la nullité du paramètre γ^0, c'est-à-dire la significativité du terme constant du modèle.*

2. Les données ont été relevées à 5 âges différents, mais on remarque que le nombre d'observations pour les arbres âgés de 4 ans est faible (car peu d'entre eux ont atteint la hauteur de 1m30). Nous avons alors regroupé les deux premières tranches d'âge et obtenons donc un nombre de degrés de liberté pour le facteur annee égal à 3.

Les Tableaux 3.7 et 3.9 indiquent que l'ajout sccessif de tous les paramètres est significatif : nous rejetons toutes les hypothèses nulles au seuil de 5%, ce qui signifie que tous les paramètres pris en compte ont un effet, et nous observons notamment que l'effet de l'interaction entre l'âge et la cironférence est très significatif.

Les Tableaux 3.8 et 3.10, quant à eux, révèlent une information supplémentaire. Si l'on teste :

$$H_0 : F_{bdsa} = \gamma^0 + \gamma_a^1 C_{bdsa} + A_b + B_{bds}^0 + B_{bds}^1 C_{bdsa} + \epsilon_{bdsa}$$

contre

$$H_1 : F_{bdsa} = \gamma_a^0 + \gamma_a^1 C_{bdsa} + A_b + B_{bds}^0 + B_{bds}^1 C_{bdsa} + \epsilon_{bdsa},$$

nous sommes amenés à accepter H_0 qui supprime l'effet de l'âge sur le terme constant du modèle, mais conserve l'effet de l'interaction *annee*$\times$*circ* : l'effet de l'âge sur la fourchaison n'est significatif qu'au travers de la circonférence.

Finalement, les deux tests de Fisher se complètent. Les tests séquentiels nous révélent que l'ajout d'un écart dû à l'année sur la constante est significatif, alors qu'en testant les paramètres un à un, l'effet de l'année est significatif mais seulement sur le terme linéaire de la circonférence. Dans les tests de type I, la nullité de ce paramètre est masquée par la significativité globale de l'année sur tous les paramètres du modèle.

⋆ Estimation des écarts-types des effets aléatoires

Il apparait que le calcul d'intervalles de confiance pour les écarts-types des effets aléatoires n'est pas possible pour le premier modèle : l'estimation de la matrice de variance-covariance de ces paramètres n'est pas définie positive.

Le calcul des intervalles de confiance pour le deuxième modèle est possible : les intervalles présentés dans le Tableau 3.11 donnent une première idée sur la significativité des effets aléatoires. Nous observons les estimations des écarts-types, la taille des intervalles de confiance calculés à partir de la loi asymptotique des estimateurs, ainsi que leur grandeur relative à l'écart-type résiduel.

Random Effects			
	Lower	**Estimate**	**Upper**
Level : bloc sd(intercept)	0.0266	0.0747	0.2096
Level : sujet sd(circ)	0.0017	0.0019	0.0021
Within-group standard error	0.7581	0.7860	0.8149

Tab. 3.11 – *Modèle 2 : Estimation des écarts-types des effets aléatoires.*

Remarquons d'abord que la résiduelle est estimée à $\sigma = 0.786$ et a un intervalle de confiance assez réduit.

L'examen des valeurs estimées des écarts-types met ensuite en évidence que les effets aléatoires de l'arbre et du bloc sont présents et non négligeables :

- l'effet du bloc est très faible ($\sigma_A = 0.0747$), d'autant plus que l'intervalle de confiance calculé pour l'écart-type est large. Mais la comparaison des critères d'Akaïke présentée dans le tableau 3.12 nous invite à préférer le modèle qui contient l'effet aléatoire du bloc sur la moyenne générale, et de ce fait, nous ne pouvons le négliger ;
- l'effet du sujet sur le terme linéaire de la circonférence ($\sigma_{B^1} = 0.0019$) apparaît faible comparativement à la résiduelle, mais va être multiplié par la circonférence, qui est de l'ordre de 250mm. L'effet sera finalement évalué à $\sigma_{B^1} = 0.475$, ce qui est de l'ordre de la résiduelle et nous amène également à le conserver dans notre modèle.

Model	**DF**	**AIC**	**logLik**
Modèle avec effet bloc sur la cste	11	5063.20	-2520.60
Modèle sans effet bloc	10	5064.87	-2522.44

Tab. 3.12 – *Comparaison des modèles avec ou sans effet bloc.*

Nous cherchons ensuite à simplifier le modèle en gardant les effets significatifs. Les modèles seront comparés entre eux à l'aide des critères *AIC* du Tableau 3.13 :

Model	**DF**	**AIC**	**logLik**
Modèle 1 avec effet sujet sur la cste	13	5067.20	-2520.60
Modèle 2 sans effet sujet	11	5063.20	-2520.60

Tab. 3.13 – *Comparaison des modèles avec ou sans effet sujet.*

La comparaison des critères d'Akaike nous conduit donc à retenir le modèle 2 n'incluant pas l'effet de l'arbre sur la moyenne.

Les différents résultats obtenus nous permettent de répondre aux questions des forestiers : les fourches sont plus développées pour des arbres dont la circonférence est importante. Plus exactement, un arbre vieux avec un gros tronc favorisera davantage la fourchaison.

3.2.2 L'analyse des résidus

Après avoir analysé les différents résultats, nous effectuons une étude graphique qui va nous permettre d'une part de représenter l'ajustement, et d'autre part de valider ou non le modèle retenu avec le dessin des résidus.

Nous reprenons les mêmes commandes présentées précédemment dans l'exemple du maïs.

Remarque 17 *Les résidus standardisés définis dans la Remarque 1 sont obtenus par la commande* `residuals(analyse, type='p')`.

◇ ***La comparaison des observés et des ajustés en fonction de la circonférence***

Sur la Figure 3.6, nous avons représenté la relation entre le nombre de fourches et la circonférence des arbres. Les points sont les observations, et les traits sont les ajustés de la modélisation retenue, par modalité du bloc en forme et de l'âge en couleur.

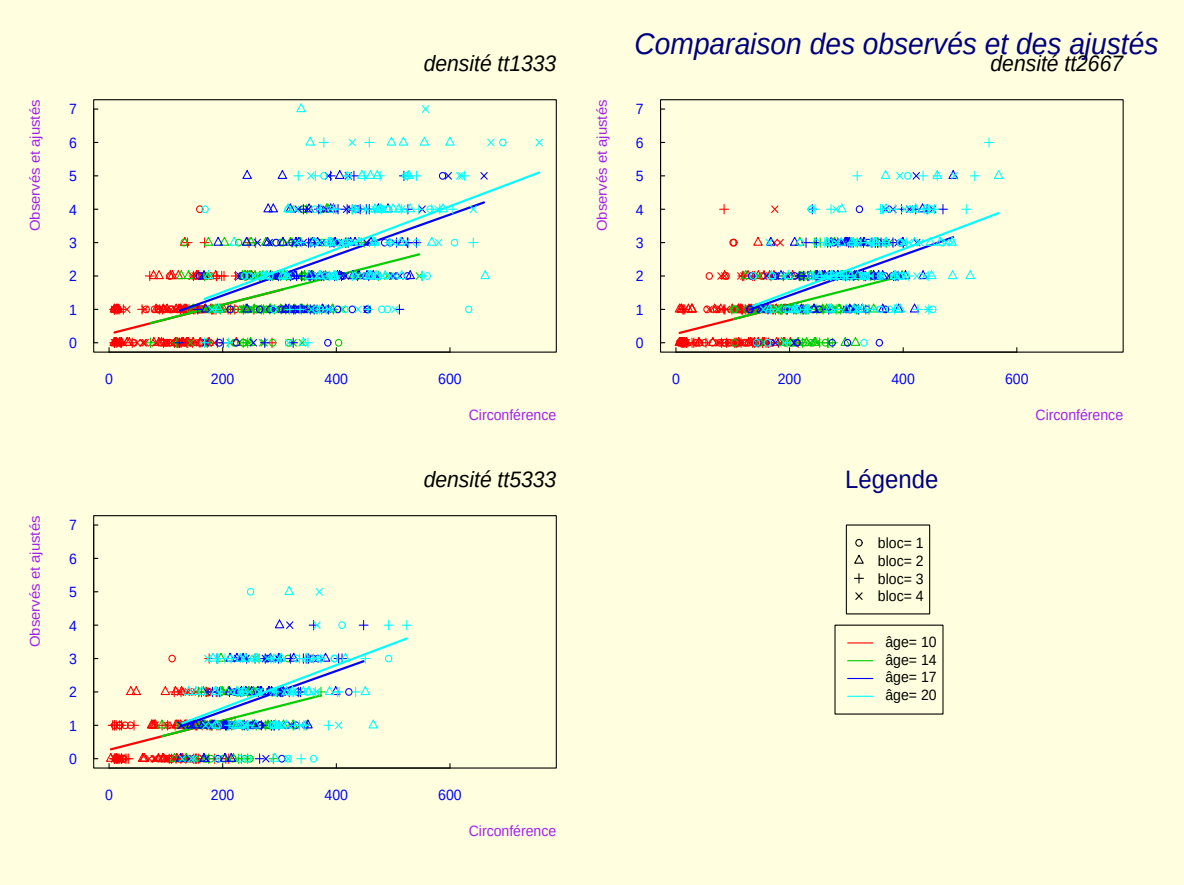

FIG. 3.6 – *Comparaison des nombres de fourches observés et ajustés, identifiés par bloc en forme et par âge en couleur et représentés par densité, en fonction de la circonférence.*

Nous observons :

- un effet de la densité : les arbres de circonférence élevée se trouvent sur les parcelles à densité faible ;

- une corrélation positive entre la fourchaison et la vigueur de l'arbre : le nombre de fourches croît avec la circonférence de l'arbre quelle que soit la densité ;
- un effet de l'âge sur la fourchaison : pour une circonférence donnée, la fourchaison est différente selon l'âge des arbres (par exemple, en prenant une circonférence fixée à 400mm, nous aurons un nombre de fourches égal à 2 pour des arbres âgés de 14 ans, et à 3 pour des arbres âgés de 17 à 20 ans) ;
- une forte significativité de l'interaction entre l'âge et la circonférence indépendamment de la densité, ce qui confirme les résultats numériques obtenus pas les tests de Fisher du Tableau 3.10 (par exemple, en prenant un arbre jeune (10 ans) et gros (300mm), ou un arbre vieux (20 ans) et fin (200mm), nous aurons une fourchaison proche de 0 ; mais pour des arbres à la fois vieux et gros, la fourchaison est favorisée).

◇ ***La comparaison des différents ajustés en fonction de la circonférence***

Sur la Figure 3.7, nous avons tracé les différents ajustés du modèle en fonction de la circonférence, par modalité de la densité en couleur et du bloc en forme. Les traits sont les ajustés des effets fixes, et les points autour sont les ajustés des effets fixes et aléatoires.

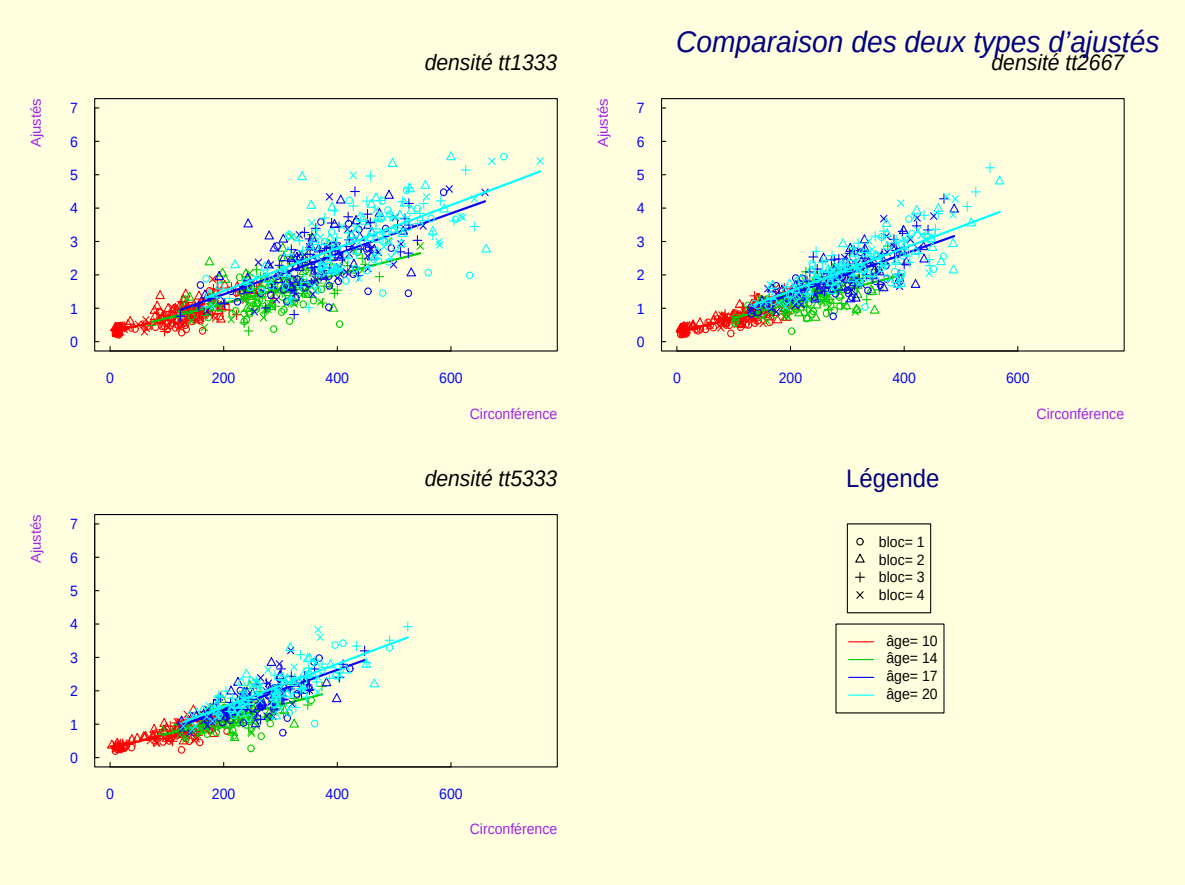

FIG. 3.7 – *Comparaison des nombres de fourches ajustés, identifiés par bloc en forme et par âge en couleur et représentés par densité, en fonction de la circonférence.*

Les traits pleins représentent l'estimation de l'espérance du nombre de fourches pour chaque modalité de l'âge identifiée par couleur, et les points représentent la variabilité, ajoutée sur les différents paramètres, due aux effets aléatoires (effet du bloc sur la constante et effet du sujet sur le terme linéaire de la circonférence).

Nous constatons :

- un effet du sujet, qui est significatif d'après les écarts-types obtenus dans le Tableau 3.11 (en dessinant uniquement l'effet du bloc, qui est relativement faible, les points auraient suivi la tendance des courbes) ;
- un effet de l'âge 4 relativement important.

◇ *Les résidus standardisés en fonction des ajustés et de la circonférence*

Afin de valider ou non le modèle retenu (1.2), nous examinons si les résidus ne présentent pas de structure évidente en traçant les graphiques des résidus réduits et des valeurs ajustées (Figures 3.8 et 3.9). Nous cherchons à éliminer la présence de toute tendance dans les résidus, qui, dans ce cas, nous conduirait à rejeter l'ajustement du modèle retenu.

Les graphiques de la Figure 3.8 représentent les résidus standardisés en fonction des ajustés, par modalité de l'âge en couleur.

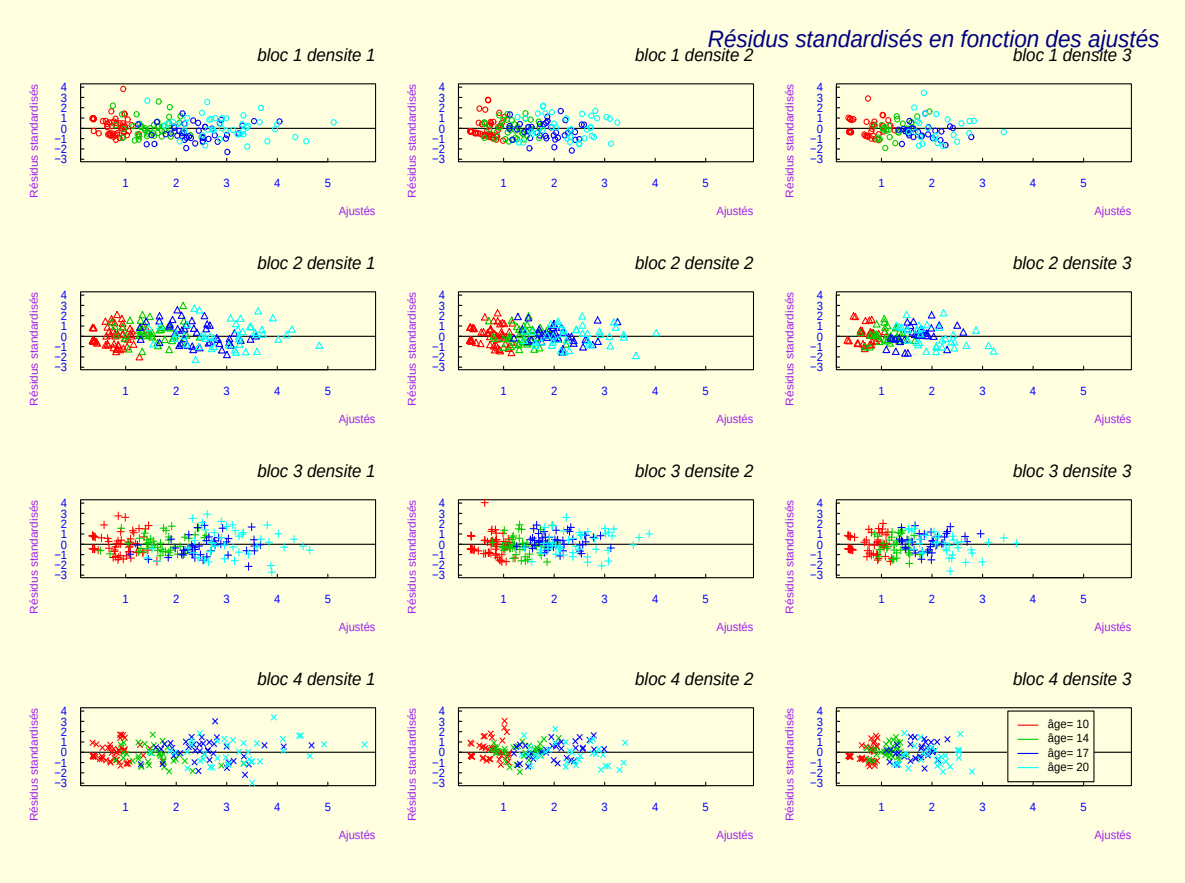

FIG. 3.8 – *Résidus standardisés, identifiés par âge en couleur et représentés par bloc et densité, en fonction des nombres de fourches ajustés.*

Les graphiques de la Figure 3.9 représentent les résidus standardisés en fonction de la cironférence, par modalité de l'âge en couleur.

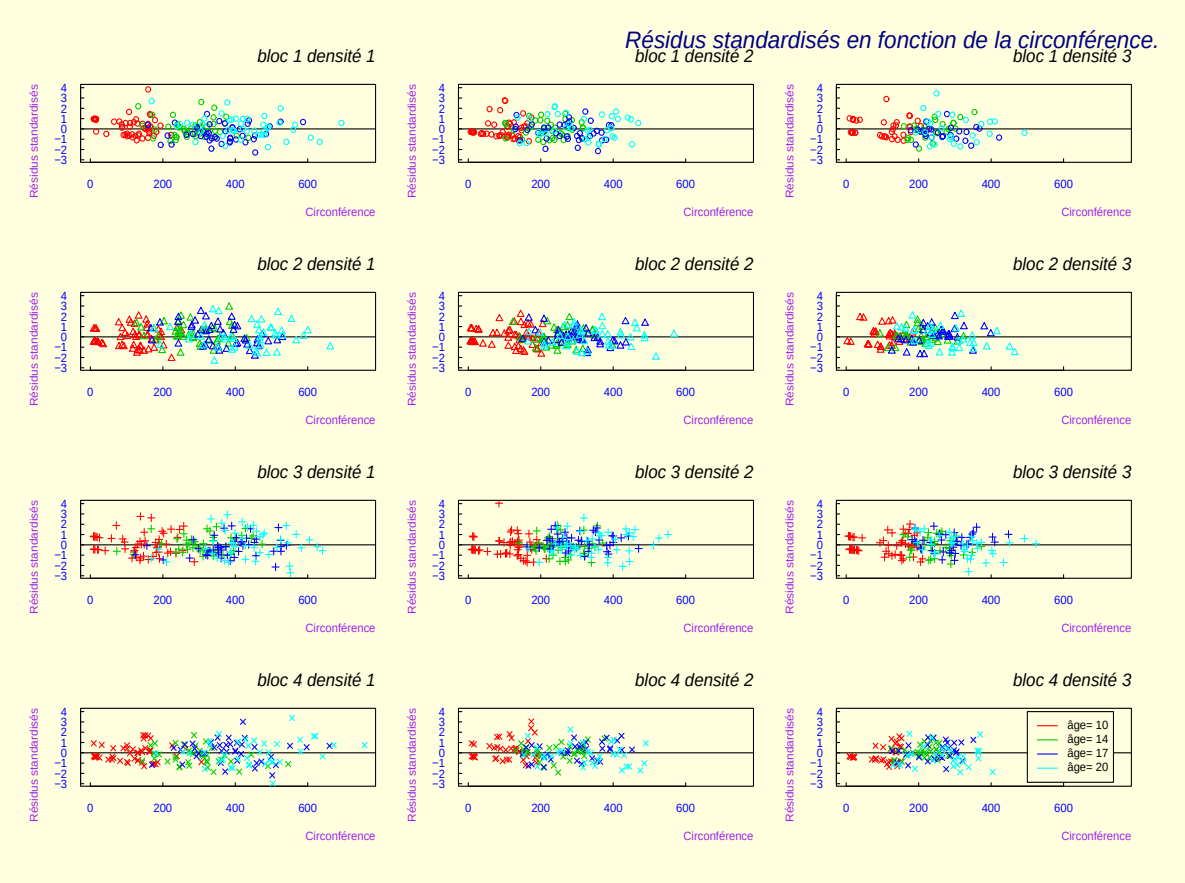

FIG. 3.9 – *Résidus standardisés, identifiés par âge en couleur et représentés par bloc et densité, en fonction de la circonférence.*

Aucune structure n'est repérée : ces résultats nous permettent alors de valider cette modélisation.

Remarquons cependant que les arbres de circonférence élevée se trouvent essentiellement dans des parcelles à faibles densité.

Conclusion

Par ce document, nous avons voulu faciliter une utilisation plus large du *modèle linéaire mixte* dans les contextes d'expérimentations agronomiques. Globalement très similaire au modèle linéaire, le modèle linéaire mixte permet de bien se prêter à un jeu de données présentant plusieurs sources de variabilité. Les tests des paramètres de variance-covariance des effets aléatoires le rendent cependant plus délicat.

Il est souvent intéressant d'effectuer une analyse des données avec plusieurs modélisations, et de disposer d'outils afin de choisir quelle est la plus adaptée aux observations.

Après avoir validé le choix du modèle, les tests d'hypothèse permettent d'apprécier la significativité des effets, et le calcul des contrastes permet à l'expérimentateur de répondre plus précisément aux diverses questions qu'il se pose en réalisant des comparaisons particulières.

Pour l'exemple des variétés de maïs, nous avons constaté l'existence de l'interaction génotype×environnement dans l'ensemble des essais menés en France en 2000 pour des variétés de maïs très précoces à demi-précoces.

Différentes méthodes, notamment l'analyse des contrastes, ont permis de mettre en évidence des écarts de rendement plus ou moins importants entre les formes initiales et modifiées des variétés, ces écarts variant selon le niveau d'attaque des pyrales. Ces résultats ne sont qu'une partie des données disponibles. Ils s'étendent à l'ensemble du *réseau d'essais*, où l'on a souligné un effet du transgène nettement visible dans les essais infestés. (*Schohn*, 2003)

Pour l'exemple des chênes de Normandie, nous pouvons conclure au conditionnement de la circonférence de l'arbre mais aussi de son âge au travers de la circonférence sur le phénomène de fourchaison.

Le modèle linéaire mixte permet d'introduire, dans les *essais répétés*, une corrélation entre les différents niveaux d'un facteur considéré aléatoire, c'est le cas notamment pour les arbres. Nous avons vu que différentes modélisations étaient envisageables, le travail du statisticien étant de choisir le meilleur ajustement selon la connaissance des données et son expérience. (*Castelli*, 2003)

Table des figures

Liste des tableaux

Bibliographie

[1] Azaïs, J.M. & Besse, P. & Croquette, A. (2001). *SAS sous UNIX, Logiciel hermétique pour système ouvert.* Publications du Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse III : Toulouse. `http://www.lsp.ups-tlse.fr/Besse`.

[2] Castelli, C. (2003). *Modélisation de la fourchaison des chênes.* Mémoire de fin d'études. Unité de biométrie, INRA : Jouy-en-Josas.

[3] Dervin, C. & Durier, C. (1993). *Analyse de Variance et Régression avec SAS : Proc GLM et Proc REG.* INRA Département et Biométrie : Versailles.

[4] Fox, P.N. & Kempton, R.A. (1997). *Statistical Methods for Plant Variety Evaluation.* Chapman & Hall : Londres.

[5] Gouet, J.P. & Philippeau, G. (1997). *Comment interpréter les résultats d'une analyse de variance ?* ITCF : Boigneville.

[6] Paradis, E. (2002). *R pour débutants.* Publications de l'Institut des Sciences de l'Evolution, Université de Montpellier II : Montpellier. `http://cran.r-project.org/doc/contrib/Rdebuts.pdf`.

[7] Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer : New York.

[8] Schohn, J. (2003). *Analyse statistique de l'expression d'un transgène dans différents contextes génétiques de maïs.* Mémoire de fin d'études. Unité de biométrie, INRA : Jouy-en-Josas.

[9] Schohn, J. (2003). *Analyse statistique de l'expression d'un transgène dans différents contextes génétiques de maïs : zones de précocité ABC1C2.* Complément au mémoire de fin d'études. Unité de biométrie, INRA : Jouy-en-Josas.

[10] Venables, W.N. & Smith, D.M. & l'équipe R Development Core Team (1990-2002). *Introduction to R.* Network Theory Limited : United Kingdom. `http://www.network-theory.co.uk/R/manual`.

[11] Vilain, M. (1999). *Méthodes expérimentales en agronomie : Pratique et analyse.* Tec & Doc : Paris.

[12] Sas Institute Inc. (1997). *SAS/STAT Software : Changes and Enhancements through Release 6.12.* Sas Institute Inc. Cary : North Carolina.

Annexe A

Programmation exemple 1

```
PROGRAMMATION SAS
*********************************************************************

LIBNAME bdd2 '/home/schohn/donnees';

/*CREATION DE LA TABLE ANAABCMOY*/

PROC IMPORT OUT=bdd2.fin_rdt
            DATAFILE= "/home/schohn/donnees/fin_rdt.csv"
            DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

/*zone AB=1911 observations*/
DATA bdd2.fin_AB;
SET bdd2.fin_rdt;
IF serie=118 OR serie=119 OR serie=177 OR serie=178 OR serie=290 OR serie=291;
RUN;

/*zone C1C2=1570 observations*/
DATA bdd2.fin_C1C2;
SET bdd2.fin_rdt;
IF serie=128 OR serie=129;
RUN;

/*zone ABC1C2=3481 observations*/
DATA bdd2.fin_ABC;
SET bdd2.fin_AB bdd2.fin_C1C2;
RUN;

PROC SORT DATA=bdd2.fin_ABC;
BY annee serie N_essai denom forme;
RUN;
```

```
PROC MEANS DATA=bdd2.fin_ABC NOPRINT;
VAR rdt;
BY annee serie N_essai denom forme;
OUTPUT OUT=bdd2.anaABCmoy mean=rdt;
RUN;

PROC SORT DATA=bdd2.fin_ABC (DROP=rdt);
BY N_essai ;
RUN;

PROC SORT DATA=bdd2.anaABCmoy (DROP=_TYPE_ _FREQ_);
BY N_essai ;
RUN;

DATA bdd2.tri (KEEP=N_essai indice indice2 ville);
SET bdd2.fin_ABC;
BY N_essai;
IF FIRST.N_essai=1;
RUN;

DATA bdd2.anaABCmoy (DROP=N_essai);
MERGE bdd2.anaABCmoy (IN=A) bdd2.tri;
BY  N_essai ;
IF A;
essai=N_essai;
RUN;


/*ANALYSE DE VARIANCE ET DES CONTRASTES SUR LA ZONE DE PRECOCITE ABC POUR L'ANNEE 2000*/

TITLE2 'MODELE 2, paramétrisation 3 (anova)' ;
PROC GLM DATA=bdd2.anaABCmoy;
    CLASS forme denom indice2  essai;
MODEL rdt=
   forme forme*denom
   indice2 indice2*essai
   indice2*forme indice2*forme*denom
   indice2*essai*forme ;
WHERE annee=2000;
RUN;


/*nous supprimons dans un premier temps les variétés de types
initiales et modifiées qui ne sont pas associées aux trois niveaux
d'infestation présents en 2000*/

DATA bdd2.anaABCmoy00;
```

```
SET bdd2.anaABCmoy;
IF annee=2000;
IF denom = 'RIVALDO' OR denom = 'SANDRINA'  THEN DELETE;
RUN;

TITLE2 'MODELE 2, paramétrisation 3 (modèle mixte)' ;
PROC MIXED DATA=bdd2.anaABCmoy00 CL COVTEST;
CLASS forme denom indice2 essai;
MODEL rdt=
   forme forme*denom
   indice2
   indice2*forme indice2*forme*denom /OUTP=bdd2.av2000
 ;
RANDOM indice2*essai indice2*essai*forme ;
LSMEANS forme*indice2 /CL ;
LSMEANS forme*denom /CL;
LSMEANS forme*denom*indice2 /CL ;
ESTIMATE 'MI'
    forme -1 1 0 0/CL;
ESTIMATE 'MI 1'
    forme -1 1 0 0
    forme*indice2 -1 0 0  1 0 0 /CL ;
ESTIMATE 'MI 2'
    forme -1 1 0 0
    forme*indice2 0 -1 0  0 1 0 /CL ;
 ESTIMATE 'MI 3'
    forme -1 1 0 0
    forme*indice2 0 0 -1  0 0 1 /CL ;
ESTIMATE 'MI 1 V1'
    forme -1 1 0 0
    forme*indice2 -1 0 0  1 0 0
    forme*denom -1 0 0 0 0 0  1 0 0 0 0 0
    forme*denom*indice2
     -1 0 0  0 0 0  0 0 0  0 0 0  0 0 0  0 0 0
      1 0 0  0 0 0  0 0 0  0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 1 V2'
    forme -1 1 0 0
    forme*indice2 -1 0 0  1 0 0
    forme*denom 0 -1 0 0 0 0  0 1 0 0 0 0
    forme*denom*indice2
     0 0 0  -1 0 0  0 0 0   0 0 0  0 0 0  0 0 0
     0 0 0   1 0 0  0 0 0   0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 1 V3'
    forme -1 1 0 0
    forme*indice2 -1 0 0  1 0 0
    forme*denom 0 0 -1 0 0 0  0 0 1 0 0 0
```

```
    forme*denom*indice2
     0 0 0  0 0 0  -1 0 0  0 0 0  0 0 0  0 0 0
     0 0 0  0 0 0   1 0 0  0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 1 V4'
    forme -1 1 0 0
    forme*indice2 -1 0 0  1 0 0
    forme*denom 0 0 0 -1 0 0  0 0 0 1 0 0
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  -1 0 0  0 0 0  0 0 0
     0 0 0  0 0 0  0 0 0   1 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 1 V5'
    forme -1 1 0 0
    forme*indice2 -1 0 0  1 0 0
    forme*denom 0 0 0 0 -1 0  0 0 0 0 1 0
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  0 0 0  -1 0 0  0 0 0
     0 0 0  0 0 0  0 0 0  0 0 0   1 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 1 V6'
    forme -1 1 0 0
    forme*indice2 -1 0 0  1 0 0
    forme*denom 0 0 0 0 0 -1  0 0 0 0 0 1
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  0 0 0  0 0 0   -1 0 0
     0 0 0  0 0 0  0 0 0  0 0 0  0 0 0  1 0 0 /CL
    ;
ESTIMATE 'MI 2 V1'
    forme -1 1 0 0
    forme*indice2 0 -1 0  0 1 0
    forme*denom -1 0 0 0 0 0  1 0 0 0 0 0
    forme*denom*indice2
     0 -1 0  0 0 0  0 0 0  0 0 0  0 0 0  0 0 0
     0 1 0   0 0 0  0 0 0  0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 2 V2'
    forme -1 1 0 0
    forme*indice2 0 -1 0  0 1 0
    forme*denom 0 -1 0 0 0 0  0 1 0 0 0 0
    forme*denom*indice2
     0 0 0  0 -1 0  0 0 0  0 0 0  0 0 0  0 0 0
     0 0 0  0 1  0  0 0 0  0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 2 V3'
    forme -1 1 0 0
    forme*indice2 0 -1 0  0 1 0
    forme*denom 0 0 -1 0 0 0  0 0 1 0 0 0
```

```
    forme*denom*indice2
     0 0 0  0 0 0  0 -1 0  0 0 0  0 0 0  0 0 0
     0 0 0  0 0 0  0 1  0  0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 2 V4'
    forme -1 1 0 0
    forme*indice2 0 -1 0  0 1 0
    forme*denom 0 0 0 -1 0 0  0 0 0 1 0 0
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  0 -1 0  0 0 0  0 0 0
     0 0 0  0 0 0  0 0 0  0 1  0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 2 V5'
    forme -1 1 0 0
    forme*indice2 0 -1 0  0 1 0
    forme*denom 0 0 0 0 -1 0  0 0 0 0 1 0
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  0 0 0  0 -1 0  0 0 0
     0 0 0  0 0 0  0 0 0  0 0 0  0 1  0  0 0 0 /CL
    ;
ESTIMATE 'MI 2 V6'
    forme -1 1 0 0
    forme*indice2  0 -1 0  0 1 0
    forme*denom 0 0 0 0 0 -1  0 0 0 0 0 1
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  0 0 0  0 0 0  0 -1 0
     0 0 0  0 0 0  0 0 0  0 0 0  0 0 0  0 1  0 /CL
    ;
ESTIMATE 'MI 3 V1'
    forme -1 1 0 0
    forme*indice2 0 0 -1   0 0 1
    forme*denom -1 0 0 0 0 0  1 0 0 0 0 0
    forme*denom*indice2
     0 0 -1   0 0 0  0 0 0  0 0 0  0 0 0  0 0 0
     0 0 1    0 0 0  0 0 0  0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 3 V2'
    forme -1 1 0 0
    forme*indice2 0 0 -1   0 0 1
    forme*denom 0 -1 0 0 0 0  0 1 0 0 0 0
    forme*denom*indice2
     0 0 0  0 0 -1   0 0 0  0 0 0  0 0 0  0 0 0
     0 0 0  0 0 1    0 0 0  0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 3 V3'
    forme -1 1 0 0
    forme*indice2 0 0 -1   0 0 1
    forme*denom 0 0 -1 0 0 0  0 0 1 0 0 0
```

```
    forme*denom*indice2
     0 0 0  0 0 0  0 0 -1   0 0 0  0 0 0  0 0 0
     0 0 0  0 0 0  0 0  1   0 0 0  0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 3 V4'
    forme -1 1 0 0
    forme*indice2 0 0 -1   0 0 1
    forme*denom 0 0 0 -1 0 0  0 0 0 1 0 0
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  0 0 -1   0 0 0  0 0 0
     0 0 0  0 0 0  0 0 0  0 0  1   0 0 0  0 0 0 /CL
    ;
ESTIMATE 'MI 3 V5'
    forme -1 1 0 0
    forme*indice2 0 0 -1   0 0 1
    forme*denom 0 0 0 0 -1 0  0 0 0 0 0 1 0
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  0 0 0  0 0 -1   0 0 0
     0 0 0  0 0 0  0 0 0  0 0 0  0 0  1   0 0 0 /CL
    ;
ESTIMATE 'MI 3 V6'
    forme -1 1 0 0
    forme*indice2  0 0 -1   0 0 1
    forme*denom 0 0 0 0 0 -1  0 0 0 0 0 0 1
    forme*denom*indice2
     0 0 0  0 0 0  0 0 0  0 0 0  0 0 0  0 0 -1
     0 0 0  0 0 0  0 0 0  0 0 0  0 0 0  0 0 1 /CL
    ;
ESTIMATE 'ON'
        forme -1 1 0 0/CL;
ESTIMATE 'ON 1'
        forme -1 1 1 -1
        forme*indice2 -1 0 0  1 0 0  1 0  -1 0  /divisor=2 CL;
ESTIMATE 'ON 2'
        forme -1 1 1 -1
        forme*indice2 0 -1 0  0 1 0  0 1  0 -1  /divisor=2 CL;
ESTIMATE 'ON 3'
        forme -1 1 0 0
        forme*indice2 0 0 -1  0 0 1 /CL;
ODS OUTPUT LSMEANS=bdd2.moy2000 ESTIMATES=bdd2.cont2000;
RUN;


/*EXPORTATION DES TABLES SAS AV2000, MOY2000 et CONT2000*/

PROC EXPORT DATA=bdd2.av2000
       OUTFILE='R/av2000.txt'
       DBMS=DLM REPLACE;
```

```
DELIMITER='00'x;
RUN;

PROC EXPORT DATA=bdd2.moy2000
       OUTFILE='R/moy2000.txt'
       DBMS=DLM REPLACE;
DELIMITER='00'x;
RUN;

PROC EXPORT DATA=bdd2.cont2000
       OUTFILE='R/cont2000.txt'
       DBMS=DLM REPLACE;
DELIMITER='00'x;
RUN;


PROGRAMMATION R
************************************************************************

#MISE EN PLACE DES DATA.FRAME

#Analyse des résidus
#av2000 est la table contenant les ajustés et les résidus corrigés

av2000 <- read.table("/home2/schohn/R/reseau/av2000.txt",header=T)
dim(av2000)
names(av2000)

#av2000b est la table contenant les ajustés et les résidus pour
les effets fixes seulement

av2000b <- read.table("/home2/schohn/R/reseau/av2000b.txt",header=T)
dim(av2000b)
names(av2000b)

#on construit les résidus standardisés
av2000$Standard <- av2000$Resid/av2000$StdErrPred

#on supprime les colonnes inutiles et on renomme l'indice
av <- av2000[,-c(1,2,6,7,9,11,12,13)]
av$indice <- av$indice2
av <- av[,-c(4)]

#on convertit les variables indice2 et forme en facteurs et on
#renomme le facteur indice2 en indice
av$indice <- as.factor(av$indice)
av$forme <- as.factor(av$forme)
```

```
#on introduit la colonne num pour les graphes des résidus
av$num <- 1:156


#Analyse des moyennes
#moy2000 est la table contenant les estimations des moyennes

moy2000 <- read.table("/home2/schohn/R/reseau/moy2000.txt",header=T)

#on conserve les colonnes utiles
moy <- moy2000[,c(2:5,11,12)]

#on convertit le facteur denom en variable numérique fonds(1 à 8)
moy$fonds <- as.numeric(moy$denom)-1

#on convertit les variables indice2 et forme en facteurs et
#on renomme le facteur indice2 en indice
moy$indice <- as.factor(moy$indice2)
moy <- moy[,-3]
moy$forme <- as.factor(moy$forme)

#on supprime les estimations ne tenant pas compte de l'indice
#puis des fonds génétiques dans "moyennes"
moyi <- moy[moy$indice !='_',]
moyennes <- moyi[-(1:10),]

#"moyglobal" regroupe les lsmeans globaux par indice
moyglobal <- moy[moy$fonds=='0',]
moyglobal <- moyglobal[,-c(2,6)]
moyglobal$indice <- as.numeric(moyglobal$indice)-1


#Analyse des contrastes
#cont2000 est la table contenant les tests sur les contrastes

cont2000 <- read.table("/home2/schohn/R/reseau/cont2000.txt",header=T)

#on conserve les colonnes utiles
cont <- cont2000[,-(4:7)]

#on introduit les labels(forme indice (facteurs) et dénomination
#(variable numérique))
cont$forme <- as.factor(rep(c('MI','OGM'),c(22,4)))
cont$indice <- as.factor(c('nan',1:3,rep(1:3,rep(6,3)),'nan',1:3))
cont$denom <- as.numeric(c(rep('nan',4),rep(1:6,3),rep('nan',4)))
cont <- cont[,-1]

#on supprime les estimations ne tenant pas compte de l'indice et
```

```
#des fonds génétiques dans "contrastes"
contrastes <- cont[cont$indice!='nan'& cont$denom!='NaN',]

#"contglobal" regroupe les contrastes globaux par indice
contglobal <- cont[cont$denom=='NaN' & cont$indice!='nan',]
contglobal <- contglobal[,-7]
contglobal$indice <- as.numeric(contglobal$indice)

#on renomme les différences de rendement
contglobal$forme <- as.factor(c(rep('modifiée-initiale',3),
 rep('ogm-non ogm',3)))


/*ETUDE GRAPHIQUE*/

#résidus standardisés en fonction des ajustés en distinguant les
  formes variétales
#initiales et modifiées
postscript("residu.ps",horizontal=T)
#la fonction postscript() permet d'ouvrir un dispositif graphique
layout(matrix(c(1,2),1,2))
#la fonction layout() partitionne le graphique actif en plusieurs
 parties sur lesquelles sont affichés les graphes successivement
par(bg="lightyellow", col.axis="blue", col.lab="purple",
col.main="navy", font.main=3, cex.main=2, cex.lab=1.5, tcl=0.25,
 adj=1, las=1)
#la fonction par() permet d'enregistrer les changements des paramètres
  graphiques
#de façon permanente
ajustes <- av$Pred
for (f in 1:2){
plot(range(ajustes),range(av$Standard),type='n',xlab='Ajustés',
ylab='Résidus standardisés')
#la fonction plot() créée un nouveau graphe bivarié
  indf <- av$forme==levels(av$forme)[f]
    for (i in 1:3) {
      indi <- av$indice==i
      indic<- (1:length(ajustes))[indi&indf]
      obs <- ajustes[indic]
      res <- av$Standard[indic]
      points(obs,res,pch=f,col=i+1)
      #la fonction points() ajoute des points au graphe existant
      ind <- (1:length(ajustes))[(abs(av$Standard)>3)&indf]
      ytext <- av$Standard[ind]
      xtext <- ajustes[ind]
      label <- as.factor(av$denom[ind])
      text(xtext-0.8, ytext-0.2, label)
      #la fonction text() ajoute le texte spécifié par label
```

```
                  }
      title(main=paste("forme",levels(av$forme)[f]),col.main="black",
 cex.main=1.5)
      indici <- (1:length(ajustes))[indf]
                  }
legend(80,-4,paste("indice=",levels(av$indice)),lty=rep(1,3),col=2:4)
#la fonction legend() ajoute la légende avec les symboles précisés
title(main=paste("Résidus standardisés en fonction des ajustés"),
outer=TRUE,line=-1.5)
#la fonction legend() ajoute un titre
dev.off()
#la fonction dev.off() permet de fermer un dispositif graphique


#Un graphe global en distinguant les formes par des caractères et des couleurs,
#et avec deux droites représentant les formes initiales/modifiées.
postscript("lsmean1.ps")
par(mfrow=c(1,1), bg="lightyellow", col.axis="blue", col.lab="purple",
 col.main="navy", font.main=3, cex.main=2, cex.lab=1.2, tcl=0.25, adj=1, las=1)
plot(range(moyglobal$indice), c(70,150), type='n', xaxt='n',
 xlab="Niveaux d'infestation", ylab='Moyennes estimées (en quintaux/hectare)')
axis(side=1, c(1,2,3), tcl=0.25)
for (f in 1:4){
  indf <- moyglobal$forme==levels(moyglobal$forme)[f]
  for (i in 1:3) {
    indi <- moyglobal$indice==i
    indic <- (1:length(moyglobal$forme))[indi&indf]
    points(moyglobal$indice[indic], moyglobal$Estimate[indic],
     pch=f+14, col=f+1)
                }
  indicf <- (1:length(moyglobal$forme))[indf]
  lines(moyglobal$indice[indicf][order(moyglobal$indice[indicf])],
   moyglobal$Estimate[indicf][order(moyglobal$indice[indicf])],
   col=f+1, lwd=1.5, lty=f)
             }
indi1 <- moyglobal$forme==levels(moyglobal$forme)[1]
indic1 <- (1:length(moyglobal$forme))[indi1]
for (k in 1:length(indic1)) {
  lines(rep(k-0.02,2), c(moyglobal$Lower[indic1[k]],
   moyglobal$Upper[indic1[k]]), col=2)
  lines(c(k-0.04,k), rep(moyglobal$Lower[indic1[k]],2), col=2)
  lines(c(k-0.04,k), rep(moyglobal$Upper[indic1[k]],2), col=2)
                            }
indi2 <- moyglobal$forme==levels(moyglobal$forme)[2]
indic2 <- (1:length(moyglobal$forme))[indi2]
for (k in 1:length(indic2)) {
  lines(rep(k,2), c(moyglobal$Lower[indic2[k]],
   moyglobal$Upper[indic2[k]]), col=3)
```

```
  lines(c(k-0.02,k+0.02), rep(moyglobal$Lower[indic2[k]],2), col=3)
  lines(c(k-0.02,k+0.02), rep(moyglobal$Upper[indic2[k]],2), col=3)
                              }
indi3 <- moyglobal$forme==levels(moyglobal$forme)[3]
indic3 <- (1:length(moyglobal$forme))[indi3]
for (k in 1:length(indic3)) {
  lines(rep(k+0.02,2), c(moyglobal$Lower[indic3[k]],
   moyglobal$Upper[indic3[k]]),col=4)
  lines(c(k,k+0.04), rep(moyglobal$Lower[indic3[k]],2), col=4)
  lines(c(k,k+0.04), rep(moyglobal$Upper[indic3[k]],2), col=4)
                              }
indi4 <- moyglobal$forme==levels(moyglobal$forme)[4]
indic4 <- (1:length(moyglobal$forme))[indi4]
for (k in 1:length(indic4)) {
  lines(rep(k+0.04,2),c(moyglobal$Lower[indic4[k]],
   moyglobal$Upper[indic4[k]]), col=5)
  lines(c(k+0.02,k+0.06),rep(moyglobal$Lower[indic4[k]],2), col=5)
  lines(c(k+0.02,k+0.06),rep(moyglobal$Upper[indic4[k]],2), col=5)
                              }
legend(1, 90, paste("forme=",levels(moyglobal$forme)), pch=15:18, col=2:5)
title(main=paste("Rendement moyen en fonction de la forme variétale et
 de l'infestation"), cex=0.5, outer=TRUE, line=-1.5)
dev.off()


#Un graphe précisant les fonds génétiques pour chaque niveau d'infestation
#en distinguant les formes par des couleurs, et avec deux droites représentant
#les formes initiales et modifiées pour chaque indice.
postscript("lsmean2.ps")
layout(matrix(c(1,2,3,4),2,2))
par(bg="lightyellow", col.axis="blue", col.lab="purple",
 col.main="navy", font.main=3, cex.main=2, cex.lab=1.2, tcl=0.25, adj=1, las=1)
for (i in 1:3){
  plot(range(moyennes$fonds), c(53,155), type='n', xlab='Fonds
  génétiques', ylab='Moyennes estimées (en quintaux/hectare)')
  indi <- moyennes$indice==i
  for (f in 2:4) {
    indf <- moyennes$forme==levels(moyennes$forme)[f]
    indic <- (1:length(moyennes$denom))[indi&indf]
    points(moyennes$fonds[indic], moyennes$Estimate[indic],
     pch=14+i, col=f+2)
    lines(moyennes$fonds[indic][order(moyennes$fonds[indic])],
     moyennes$Estimate[indic][order(moyennes$fonds[indic])], col=f+2, lwd=1)
    for (j in 1:length(indic)) {
      lines(c(moyennes$fonds[indic[j]], moyennes$fonds[indic[j]]),
       c(moyennes$Lower[indic[j]], moyennes$Upper[indic[j]]), col=f+2)
      lines(c(moyennes$fonds[indic[j]]-0.05,
       moyennes$fonds[indic[j]]+0.05), rep(moyennes$Lower[indic[j]],2), col=f+2)
```

```
      lines(c(moyennes$fonds[indic[j]]-0.05,
       moyennes$fonds[indic[j]]+0.05), rep(moyennes$Upper[indic[j]],2), col=f+2)
                                }
                  }
indf1 <- moyennes$forme==levels(moyennes$forme)[1]
indic1 <- (1:length(moyennes$denom))[indi&indf1]
points(moyennes$fonds[indic1], moyennes$Estimate[indic1],
 pch=14+i, col=3)
lines(moyennes$fonds[indic1][order(moyennes$fonds[indic1])],
 moyennes$Estimate[indic1][order(moyennes$fonds[indic1])], col=3, lwd=1)
for (k in 1:length(indic1)) {
  lines(c(moyennes$fonds[indic1[k]]-0.04,
   moyennes$fonds[indic1[k]]-0.04),
   c(moyennes$Lower[indic1[k]],moyennes$Upper[indic1[k]]), col=3)
  lines(c(moyennes$fonds[indic1[k]]-0.09,
   moyennes$fonds[indic1[k]]+0.01), rep(moyennes$Lower[indic1[k]],2), col=3)
  lines(c(moyennes$fonds[indic1[k]]-0.09,
   moyennes$fonds[indic1[k]]+0.01), rep(moyennes$Upper[indic1[k]],2), col=3)
                            }
title(main=paste("infestation", levels(moyennes$indice)[i+1]),
 col.main="black" , cex.main=1.5)
                  }
plot(range(moyennes$fonds),range(moyennes$Estimate), axes=FALSE,
 main='Légende', font.main=1, cex.main=1.7, xlab='', ylab='', adj=0.5,
 type='n ')
legend(3, 140, paste("forme=", levels(moyennes$forme)), lty=rep(1,4), col=3:6)
title(main=paste("Rendement moyen en fonction de la forme variétale et
 du niveau d'infestation"), outer=TRUE, line=-1.5)
dev.off()


#Un graphe "résumé" précisant les fonds génétiques en distinguant les formes
#par des caractères et les niveaux d'infestation par des couleurs, et avec
#six droites représentant les formes initiales/modifiées pour chaque niveau
#d'infestation.
postscript("lsmean3.ps")
layout(matrix(c(1,2),1,2))
par(bg="lightyellow", col.axis="blue", col.lab="purple",
 col.main="navy", font .main=3, cex.main=2, cex.lab=1.2, tcl=0.25, adj=1, las=1)
plot(range(moyennes$fonds),range(moyennes$Estimate), type='n',
 xlab='Fonds génétiques', ylab='Moyennes estimées (en quintaux/hectare)')
for (i in 1:3){
  indi <- moyennes$indice==i
  for (f in 1:4) {
    indf <- moyennes$forme==levels(moyennes$forme)[f]
    indic <- (1:length(moyennes$denom))[indi&indf]
    points(moyennes$fonds[indic],moyennes$Estimate[indic],pch=f,col=i+1)
    lines(moyennes$fonds[indic][order(moyennes$fonds[indic])],
```

```
     moyennes$Estimate[indic][order(moyennes$fonds[indic])], col=i+1, lwd=1, lty=f)
                }
             }
plot(range(moyennes$fonds), range(moyennes$Estimate), axes=FALSE,
 font.main=1, cex.main=1.7, xlab='',ylab='', adj=0.5, type='n')
mtext("Légende", col="navy", font=1, cex=1.7, side=1,line=1)
legend(3.3,85, paste("indice=",levels(moyennes$indice)[-1]),
 lty=rep(1,3), col=2:4)
legend(2.8,110, paste("forme=",levels(moyennes$forme)), pch=1:4)
legend(2.7,95, paste("forme=", levels(moyennes$forme)[(1:2)]), lty=1:2)
title(main=paste("Rendement moyen en fonction de la forme variétale et
 de l'infestation"), outer=TRUE, line=-1.5)
dev.off()


#Un graphe représentant les écarts entre les formes modifiées et initiales
#pour chaque fonds génétique en distinguant les niveaux d'infestation par
#des caractères et des couleurs, et avec deux droites représentant ces
#différents niveaux.
postscript("estimate.ps", horizontal=T)
par(mfrow=c(1,1), bg="lightyellow", col.axis="blue", col.lab="purple",
 col.main="navy", font.main=3, cex.main=2, cex.lab=1.2, tcl=0.25, adj=1, las=1)
plot(range(contrastes$denom), c(-20,60), type='n', xlab='Fonds
 génétiques', ylab='Ecarts estimés (en quintaux/hectare)')
abline(h=0, lwd=2, col='black')
#la fonction abline() trace une lilgnbe horizontale sur l'ordonnée précisée
for (i in 1:3){
  indi <- contrastes$indice==i
  indic <- (1:length(contrastes$forme))[indi]
  points(contrastes$denom[indic],contrastes$Estimate[indic], pch=14+i,
   col=i+1)
  lines(contrastes$denom[indic][order(contrastes$denom[indic])],
   contrastes$Estimate[indic][order(contrastes$denom[indic])],
   col=i+1, lwd=1.5, lty=i)
             }
indi1 <- contrastes$indice==1
indic1 <- (1:length(contrastes$denom))[indi1]
for (k in 1:length(indic1)) {
  lines(rep(k-0.05,2), c(contrastes$Lower[indic1[k]],
   contrastes$Upper[indic1[k]]), col=2)
  lines(c(k-0.1,k), rep(contrastes$Lower[indic1[k]],2), col=2)
  lines(c(k-0.1,k), rep(contrastes$Upper[indic1[k]],2), col=2)
                           }
indi2 <- contrastes$indice==2
indic2 <- (1:length(contrastes$denom))[indi2]
for (k in 1:length(indic2)) {
  lines(rep(k,2),
   c(contrastes$Lower[indic2[k]],contrastes$Upper[indic2[k]]), col=3)
```

```
  lines(c(k-0.05,k+0.05), rep(contrastes$Lower[indic2[k]],2),
   col=3)
  lines(c(k-0.05,k+0.05), rep(contrastes$Upper[indic2[k]],2),
   col=3)
                              }
indi3 <- contrastes$indice==3
indic3 <- (1:length(contrastes$denom))[indi3]
for (k in 1:length(indic3)) {
  lines(rep(k+0.05,2),
   c(contrastes$Lower[indic3[k]],contrastes$Upper[indic3[k]]), col=4)
  lines(c(k,k+0.1), rep(contrastes$Lower[indic3[k]],2), col=4)
  lines(c(k,k+0.1), rep(contrastes$Upper[indic3[k]],2), col=4)
                               }
legend(1,60, paste("indice=", levels(contrastes$indice)[-4]), lty=1:3, col=2:4)
title(main=paste("Ecart de rendement en fonction de l'infestation"),
 outer=TRUE, line=-1.5)
dev.off()
```

Annexe B

Programmation exemple 2

```
#On remarque que le nombre des observations pour les arbres ages de 4 ans
#est faible (en effet peu d'entre eux ont atteints la hauteur 1,30m).
#Aussi,on regroupe les deux premieres tranches d'age en une seule.

ann4 <- as.numeric(dataCirc$annee)
ann4[ann4==1]<-2
dataCirc$ann4 <- factor(ann4,labels=levels(dataCirc$annee)[-1])


#MODELISATION D'UNE FONCTION LINEAIRE DE LA CIRCONFERENCE

#On modelise la relation entre le nombre de fourches et la circonference par
#une fonction lineaire dont les parametres dependent de la densite et de l'age.
#On introduit une effet aleatoire du bloc et de l'arbre sur la moyenne
#generale et un effet aleatoire de l'arbre sur la pente de la regression
#en fonction de la circonference.
#Les modeles seront comparés entre eux a l'aide
#du critere d'Akaike.

res1.lin.lme <- lme(Nf ~ ann4*circ*dens,data=dataCirc,random=list(bloc=~1,
sujet=~circ),method="ML")

anova(res1.lin.lme)
#               numDF denDF  F-value p-value
#(Intercept)        1  1467 673.8004  <.0001
#ann4               3  1467 352.3548  <.0001
#circ               1  1467 286.4901  <.0001
#dens               2   445   0.9053  0.4051
#ann4:circ          3  1467   6.2090  0.0003
#ann4:dens          6  1467   1.4108  0.2068
#circ:dens          2  1467   0.4968  0.6085
#ann4:circ:dens     6  1467   1.7797  0.0997

intervals(res1.lin.lme)
```

```
#Error in intervals.lme(res1.lin.lme) : Cannot get confidence
 intervals on var-cov components: Non-positive definite approximate
 variance-covariance.

#Il apparait que le calcul d'intervalles de confiance sur les
#ecarts-types des effets aleatoires n'est pas possible,
#car l'estimation de la matrice de variance de ces parametres n'est pas
#definie positive.

summary(res1.lin.lme)
#Random effects:
# Formula: ~1 | bloc
#        (Intercept)
#StdDev:  0.07387879
#
# Formula: ~circ | sujet %in% bloc
# Structure: General positive-definite, Log-Cholesky parametrization
#             StdDev      Corr
#(Intercept) 0.011389633 (Intr)
#circ        0.001923732 -0.425
#Residual    0.780931764

#Un examen des valeurs estimees des écarts-types met
#en evidence que la variance de l'effet aleatoire de l'arbre sur la moyenne
#generale est faible (\sigma_{B^0} = 0.011) comparativement l'écart-type
#de l'effet aleatoire de l'arbre sur la pente de la regression
#(\sigma_{B^1} * circ a pour ordre de grandeur 0.002*250). Nous allons
#donc supprimer l'effet aleatoire de l'arbre  sur la moyenne  generale :

res2.lin.lme <-
  update(res1.lin.lme,data=dataCirc,random=list(bloc=~1,
 sujet=~circ-1),method="ML")

intervals(res2.lin.lme)
# Random Effects:
#  Level: bloc
#                     lower       est.     upper
#sd((Intercept)) 0.02605752 0.07376845 0.2088373
#  Level: sujet
#                lower        est.       upper
#sd(circ) 0.001711996 0.001909981 0.002130863
#
# Within-group standard error:
#    lower      est.     upper
#0.7530787 0.7807417 0.8094209

#Le calcul des intervalles de confiance pour
#les parametres est maintenant possible.
```

```
#On compare les modèles :
anova(res2.lin.lme,res1.lin.lme)
#              Model df      AIC      BIC    logLik
#res2.lin.lme     1 27 5073.095 5223.483 -2509.547
#res1.lin.lme     2 29 5077.185 5238.713 -2509.592

#On preferera donc le deuxieme modele qui a un critere d'Akaike plue petit.

#Puis on cherche a "simplifier le modele", c'est a dire a garder les
#effets significatifs.
#On suppose une relation lineaire sans effet de la densite :

res3.lin.lme <- update(res2.lin.lme,fixed=Nf ~ ann4*circ)

anova(res3.lin.lme)
#            numDF denDF  F-value p-value
#(Intercept)     1  1481 671.1611  <.0001
#ann4            3  1481 353.5624  <.0001
#circ            1  1481 287.3395  <.0001
#ann4:circ       3  1481   6.1776   4e-04

#Enfin on compare ces modeles :
anova(res2.lin.lme,res3.lin.lme)
#              Model df      AIC      BIC    logLik
#res2.lin.lme     1 27 5073.095 5223.483 -2509.547
#res3.lin.lme     3 11 5063.199 5124.468 -2520.599


#La comparaison des critères d'Akaike conduit à choisir le modèle 3.
#Pour le valider, nous examinons les graphiques
#des residus reduits et les valeurs ajustees.


#ETUDE GRAPHIQUE

postscript("arbre1.ps")
par(mfrow=c(2,2), bg="lightyellow", col.axis="blue", col.lab="purple",
 col.main="navy", font.main=3, cex.main=2, cex.lab=1, tcl=0.25, adj=1, las=1)
ajustes <- fitted(res3.lin.lme,0)
for (d in 1:3){
  plot(range(dataCirc$circ), range(dataCirc$Nf), type='n',
   xlab='Circonférence', ylab='Observés et ajustés')
  indd <- dataCirc$dens==levels(dataCirc$dens)[d]
  for (a in 1:5) {
    inda <- dataCirc$ann4==levels(dataCirc$ann4)[a]
    for (b in 1:4) {
      indb <- dataCirc$bloc==b
```

```
      indic<- (1:1939)[indd&indb&inda]
      circ <- dataCirc$circ[indic]
      points(circ,dataCirc$Nf[indic],pch=b,col=a+1)
                    }
    aj <- ajustes[(1:1939)[indd&inda]]
    circ <- dataCirc$circ[(1:1939)[indd&inda]]
    lines(circ[order(circ)],aj[order(circ)],col=a+1,lty=1,lwd=2)
                  }
  title(main=paste("densité", levels(dataCirc$dens)[d]), col.main="black", cex.main=1.5)
              }
plot(range(dataCirc$circ),range(dataCirc$Nf), axes=FALSE,
 main='Légende', font.main=1, cex.main=1.7, xlab='', ylab='', adj=0.5, type='n')
legend(300,7,paste("bloc=",levels(dataCirc$bloc)),pch=1:4)
legend(280,4,paste("âge=",levels(dataCirc$ann4)),lty=rep(1,4),col=2:5)
title(main=paste('Comparaison des observés et des ajustés'),
 outer=TRUE, line=-1.5)
dev.off()


postscript("arbre2.ps")
par(mfrow=c(2,2), bg="lightyellow", col.axis="blue", col.lab="purple",
 col.main="navy", font.main=3, cex.main=2, cex.lab=1, tcl=0.25, adj=1, las=1)
ajR <- fitted(res3.lin.lme,2)
for (d in 1:3){
  plot(range(dataCirc$circ), range(dataCirc$Nf), type='n',
   xlab='Circonférence', ylab='Ajustés')
  indd <- dataCirc$dens==levels(dataCirc$dens)[d]
  for (a in 1:5) {
    inda <- dataCirc$ann4==levels(dataCirc$ann4)[a]
    for (b in 1:4) {
      indb <- dataCirc$bloc==b
      indic<- (1:1939)[indd&indb&inda]
      circ <- dataCirc$circ[indic]
      points(circ,ajR[indic],pch=b,col=a+1)
                    }
    aj <- ajustes[(1:1939)[indd&inda]]
    circ <- dataCirc$circ[(1:1939)[indd&inda]]
    lines(circ[order(circ)],aj[order(circ)],col=a+1,lty=1,lwd=2)
                  }
  title(main=paste("densité", levels(dataCirc$dens)[d]), col.main="black", cex.main=1.5)
              }
plot(range(dataCirc$circ), range(dataCirc$Nf), axes=FALSE,
 main='Légende', font.main=1, cex.main=1.7, xlab='', ylab='', adj=0.5, type='n')
legend(300,7,paste("bloc=",levels(dataCirc$bloc)),pch=1:4)
legend(280,4,paste("âge=",levels(dataCirc$ann4)),lty=rep(1,4),col=2:5)
title(main=paste("Comparaison des deux types d'ajustés"), outer=TRUE, line=-1.5)
dev.off()
```

```
postscript("arbre3.ps")
par(mfrow=c(4,3), bg="lightyellow", col.axis="blue", col.lab="purple",
 col.main="navy", font.main=3, cex.main=2, cex.lab=1, tcl=0.25, adj=1, las=1)
residus <- residuals(res3.lin.lme,type='p')
for (b in 1:4) {
  indb <- dataCirc$bloc==b
  for (d in 1:3) {
    indd <- dataCirc$dens==levels(dataCirc$dens)[d]
    plot(range(ajustes), range(residus), type="n", xlab='Ajustés',
     ylab='Résidus standardisés')
    title(main=paste('bloc',b,'densite',d), col.main="black", cex.main=1.5)
     abline(0,0)
    for (a in 1:5) {
      inda <- dataCirc$ann4==levels(dataCirc$ann4)[a]
      indic<- (1:1939)[indd&indb&inda]
      points(ajustes[indic],residus[indic],pch=b,col=a+1)
                  }
                }
              }
legend(3.5,4,paste("âge=",levels(dataCirc$ann4)),lty=rep(1,4),col=2:5)
title(main='Résidus standardisés en fonction des ajustés', outer=TRUE,
 line=-1.5)
dev.off()


postscript("arbre4.ps")
par(mfrow=c(4,3), bg="lightyellow", col.axis="blue", col.lab="purple",
 col.main="navy", font.main=3, cex.main=2, cex.lab=1, tcl=0.25, adj=1, las=1)
for (b in 1:4) {
  indb <- dataCirc$bloc==b
  for (d in 1:3) {
    indd <- dataCirc$dens==levels(dataCirc$dens)[d]
    plot(range(dataCirc$circ), range(residus), type="n",
     xlab='Circonférence', ylab='Résidus standardisés')
    title(main=paste('bloc',b,'densité',d),col.main="black", cex.main=1.5)
     abline(0,0)
    for (a in 1:5) {
      inda <- dataCirc$ann4==levels(dataCirc$ann4)[a]
      indic<- (1:1939)[indd&indb&inda]
      points(dataCirc$circ[indic],residus[indic],pch=b,col=a+1)
                  }
                }
              }
legend(500,4,paste("âge=",levels(dataCirc$ann4)),lty=rep(1,4),col=2:5)
title(main='Résidus standardisés en fonction de la circonférence.',
 outer=TRUE, line=-1.5)
dev.off()
```