ACADEMIE DE MONTPELLIER

UNIVERSITE DE MONTPELLIER

MASTER

EPIDEMIOLOGIE, DONNEES DE SANTE ET BIOSTATISTIQUES
DATA ANALYST POUR LES SCIENCES DE LA VIE

MEMOIRE sur le stage de
Malwenn LANGEVIN

Mise en œuvre de modèle linéaire mixte pour l'analyse de données structurées

- effectué du 19/02/2024 au 16/08/2024
- à Montpellier
- sous la direction de Heuclin Benjamin
- par Malwenn LANGEVIN
- soutenu le 10/09/2024
- devant la commission d'examen J.N. BACRO, C. REYNES, R. SABATIER.

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude envers le Cirad pour m'avoir accueillie et de m'avoir offert l'opportunité d'effectuer mon stage au sein de cette entreprise. Mes plus sincères remerciements vont à toute l'équipe d'Aida pour leur accueil chaleureux, leur soutien et leur collaboration tout au long de cette expérience professionnelle enrichissante.

Je tiens également à adresser mes remerciements les plus sincères à mon encadrant, Benjamin Heuclin, pour sa patience, son expertise et ses précieux conseils qui m'ont guidée tout au long de la réalisation de ce stage. Ses retours constructifs ont grandement contribué à l'enrichissement de mes compétences et à la réussite de ce travail. Ainsi, qu'à Albert Flori et Sandrine Le Squin qui m'ont fourni les données nécessaires au stage et à Marie Denis pour ses connaissances approfondies sur les modèles mixtes appliqués à la génétique.

Je souhaite également exprimer ma reconnaissance envers Christelle Reynes de la Faculté de Pharmacie de Montpellier pour son suivi attentif et ses encouragements tout au long de mon parcours académique et professionnel.

Ce mémoire de stage représente le fruit du travail réalisé tout au long de ce stage. Je suis profondément reconnaissante envers toutes les personnes qui ont contribué, de près ou de loin, à la bonne réalisation de ce stage.

Résumé

L'étude de phénomènes complexes, que ce soit en en agronomie, écologie et santé implique généralement le recueil de données structurées. L'analyse de telles données nécessite des méthodes statistiques spécifiques, comme les modèles linéaires mixtes, pour gérer la variabilité intra- et inter-groupes. Cependant, leur application peut se révéler complexe en raison des choix de modèles et des paramètres à définir. Dans un premier temps, nous avons exploré les connaissances existantes sur les modèles mixtes pour élaborer un guide complet. Ce guide aborde divers aspects des modèles, tels que les types de structurations possibles, leur estimation et le développement des librairies utilisés. Ensuite, nous avons testé ces librairies sur plusieurs jeux de données pour évaluer leurs avantages et inconvénients et leur capacité à réaliser des modèles équivalents. Enfin, nous avons recherché des alternatives gratuites au librairie AsReml et avons démontré qu'une telle option est effectivement disponible.

Mots-clés : Modèles Mixtes, Structures de corrélation, Données structurées, Effets aléatoires

Abstract

Structured data have a significant impact on the analysis of complex phenomena in agronomy, ecology, and health. Their processing requires specific statistical methods, such as linear mixed models, to manage intra- and inter-group variability. However, their application can be complex due to model choices and parameters to define. Initially, we explored existing knowledge on mixed models to develop a comprehensive guide. This guide covers various aspects of the models, such as possible structuring types, their estimation, and the development of the librairies used. Subsequently, we tested these librairies on several datasets to assess their advantages and disadvantages and their ability to perform equivalent models. Finally, we searched for free alternatives to the AsReml librairie and demonstrated that such an option is indeed available.

Keywords : Mixed Models, Correlation Structures, Structured Data, Random effects

Sommaire

Table des figures

Liste des tableaux

Introduction

1 Structure d'accueil

Dans cette partie, seront d'abord présentés l'établissement qui a permis la réalisation du stage ainsi que l'entreprise qui l'a financé. Ensuite, le contexte et la problématique du stage seront exposés, en expliquant la pertinence et la nécessité de la création d'un guide sur les modèles mixtes dans ce cadre. Les défis spécifiques rencontrés et les questions auxquelles ce travail a cherché à répondre seront abordés. Une présentation des données utilisées sera également effectuée. Enfin, les objectifs du stage seront définis, tant en termes d'apports théoriques que pratiques.

1.1 Cirad

Le Centre de coopération internationale en recherche agronomique pour le développement (Cirad) est un organisme français spécialisé en recherche agronomique appliquée aux régions tropicales et méditerranéennes. Fondé en 1984 et basé à Montpellier, le Cirad joue un rôle important dans la coopéaration scientifique internationale pour le développement agricole.

Pour le développement durable des systèmes agricoles, le Cirad conduit des recherches innovantes. Il collabore avec des institutions de recherche locales et internationales pour élaborer des solutions adaptées aux défis des différentes régions. Ses domaines de recherche englobent l'agronomie, l'écologie, l'économie, la santé des plantes et des animaux.

En outre, le Cirad promeut la coopération scientifique avec les pays en développement pour renforcer les capacités de recherche et de formation. Il établit des partenariats durables avec des instituts de recherche, des universités afin de partager des connaissances et des compétences.

1.2 PalmElit

PalmElit est une entreprise française spécialisée dans la sélection et la production de semences de palmiers à huile. Fondée en 2008 et basée à Montpellier, PalmElit est une filiale du Cirad, et elle se positionne comme un acteur majeur dans l'amélioration génétique des palmiers à huile.

Elle mène des programmes de recherche avancés pour développer de nouvelles variétés de palmiers à huile. PalmElit se concentre sur l'amélioration génétique des palmiers en utilisant des techniques modernes de sélection, de biotechnologie et de génomique. L'objectif est de créer des variétés plus productives, résistantes aux maladies et adaptées aux conditions climatiques des régions de culture.

En outre, PalmElit garantit la production et la distribution de semences de haute qualité pour les plantations de palmiers à huile. PalmElit supervise l'ensemble du processus de production des semences, depuis la sélection des parents jusqu'à la mise sur le marché des graines certifiées. L'entreprise met en œuvre des pratiques de production rigoureuses pour assurer la qualité, la pureté et la performance des semences, répondant ainsi aux normes internationales et aux besoins des producteurs.

1.3 Contexte

Dans différents domaines de recherche tels que l'agronomie, l'écologie et la santé, l'étude de phénomènes complexes implique généralement l'analyse de données structurées. Elles se présentent sous différentes formes : individus apparentés en génétique, données spatiales en écologie, ou données longitudinales en santé.

Ces données structurées nécessitent des méthodes statistiques spécifiques, comme par exemple les modèles linéaires mixtes. Cependant, leur mise en oeuvre peut être difficile. De nombreux logiciels statistiques tels que *AsReml*, et diverses librairies de R, chacun avec leur spécifités et limitations sont disponibles. Le choix du logiciel/librairie et la mise en oeuvre d'une modèle linéaire mixte peut alors vite devenir complexes.

1.4 Données

Dans ce document, seront présentés deux jeux de données provenant de l'entreprise PalmElit.

1.4.1 Jeux de données n°1 : Regroupement d'essai

Le premier jeu de données est est un regroupement de 28 essais Alpha-Lattice menés sur le palmier à huile (figure 1). Chaque arbre est identifié par un numéro d'essai, un numéro de répétition, un bloc et position dans la parcelle. Les mesures enregistrées incluent le nombre de grappes de fruits sur une période de 3 à 5 ans, le poids des grappes de fruits frais sur une période de 3 à 5 ans (exprimé en kilogrammes), et le poids moyen des fruits sur une période de 3 à 5 ans (exprimé en kilogrammes).

Au total, 452 génotypes différents sont présents et les données comportent 30 409 observations pour 17 variables. Pour analyser ces données, 2 types de modèles sont possibles. Le modèle père/mère qui est formulé en incluant des termes pour les effets fixes des génotypes du père et de la mère, ainsi qu'un terme pour l'erreur aléatoire. Et le modèle animal qui prend en compte les contributions génétiques non seulement des parents mais aussi des ancêtres plus éloignés.

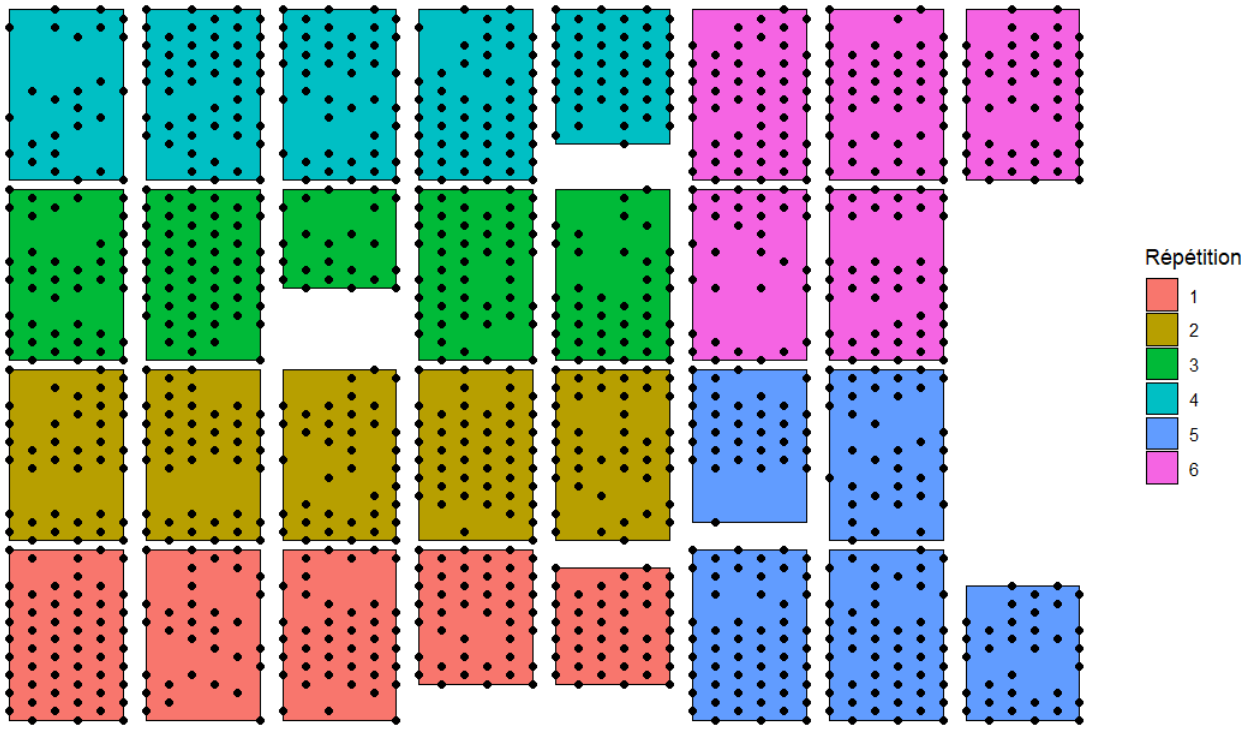

FIGURE 1 – Représentation de la disposition des palmier sur un des essais du regroupement des 28 essais. Les point représentent les palmier, les rectangles colorés représentent les blocs incomplets et les couleurs représentent les répétitions complètes. Les trous sont dus à des palmiers décédés.

1.4.2 Jeux de données n°2 : Plan Alpha-Lattice

Le second jeu de données concerne un essai sur le palmier à huile planté au Brésil en 2010 (figure 2). Cet essai vise à étudier l'effet de huit croisements d'hybrides, plantés selon des densités variant de 103 arbres/ha à 143 arbres/ha, sur le diamètre des troncs à 12 ans. Il y a deux parcelles élémentaires de 64 arbres par croisement, dont 49 arbres utiles. Les parcelles sont regroupées en blocs incomplets, eux-mêmes regroupés en répétitions complètes selon un plan Alpha-Lattice. Le champ sur lequel l'essai a été réalisé n'a pas une topographie uniforme donc les données présentent de la corrélation spatiale.

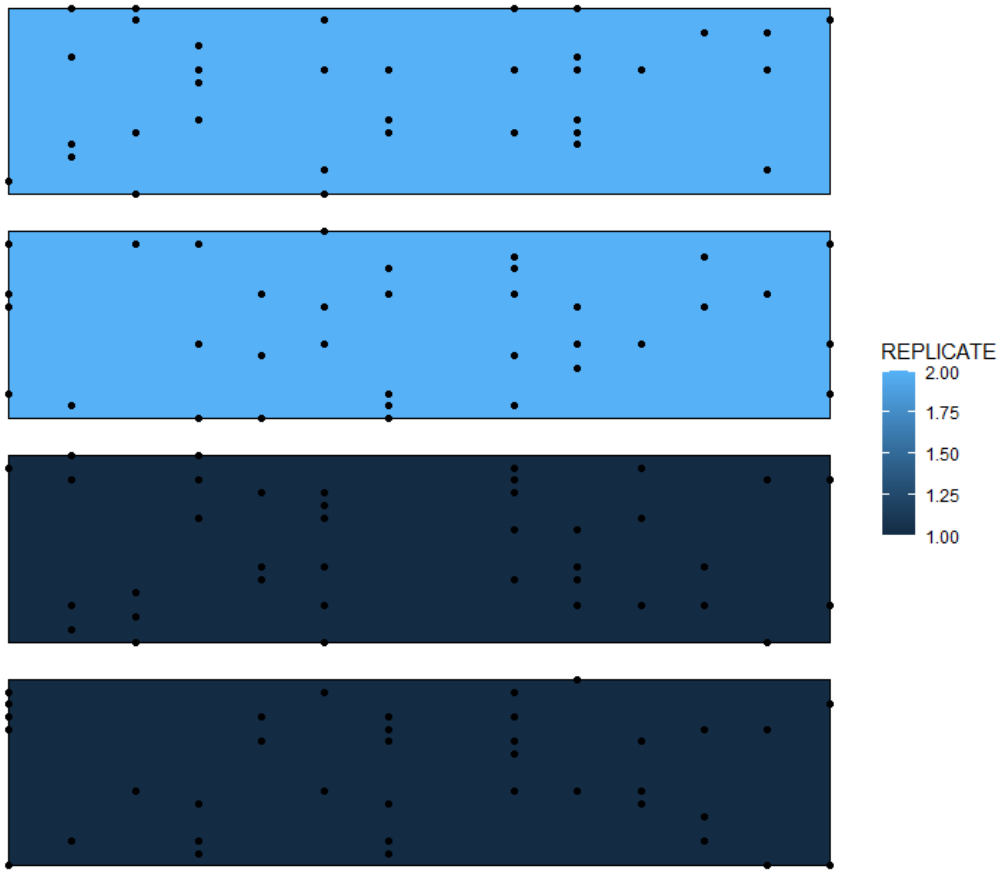

FIGURE 2 – Représentation de la disposition des palmier. Les point représentent les palmier, les rectangles colorés représentent les blocs incomplets et les couleurs représentent les répétitions complètes. Les trous sont dus à des palmiers qui n'ont pas été observés.

1.5 Objectifs

L'un des objectifs principaux du stage étaient de réaliser un guide complet et structuré sur les modèles mixtes, avec pour ambition de centraliser l'ensemble des connaissances nécessaires à leur compréhension et leur application. Ce guide doit servir de référence pour les chercheurs, étudiants, et professionnels intéressés par les modèles mixtes en offrant une compilation des connaissances théoriques, incluant une présentation détaillée des concepts fondamentaux.

En parallèle, le guide devait inclure une revue des outils et des logiciels disponibles pour la mise en oeuvre de ces modèles, présentant les fonctionnalités, avantages, et limites des principales solutions logicielles comme celles proposées par R. Une partie essentielle du travail consistait également à

développer des exemples pratiques et des cas d'étude, illustrant l'application des modèles mixtes à des données réelles, avec des démonstrations de code et des interprétations de résultats pour rendre le guide à la fois théorique et pratique. Tout cela a été réalisé sur R (version 4.3.0 (2023-04-21 ucrt)).

Ainsi, le stage visait à créer une ressource exhaustive qui non seulement présentait les concepts théoriques mais offrait également des outils pratiques et des conseils pour intégrer les modèles mixtes dans divers projets de recherche ou professionnels. L'autre objectif du stage était de rechercher une alternative gratuite à *AsReml*, un librairie de modélisation statistique pour les modèles mixtes qui nécessite l'achat d'une licence (BUTLER et al. 2009), afin de proposer une solution plus en libre accès pour PalmElit. *AsReml* est reconnu pour sa puissance et sa capacité à traiter des modèles mixtes complexes, son coût élevé peut représenter un obstacle pour les entreprises cherchant des solutions moins coûteuses. Le stage visait donc à identifier des librairies libres sur R en examinant leur performance et leur compatibilité avec les besoins de l'entreprise.

Guide

2 Définition d'un modèle mixte

Tout d'abord, il fallait définir le modèle linéaire mixte avant de se plonger dans les différentes composantes de ce même modèle. Une définition claire de ce qu'est un modèle mixte a été établie, en s'appuyant sur les travaux déjà réalisé dans le domaine. Cette étape de définition était importante pour établir les fondations nécessaires à une compréhension des effets fixes et aléatoires. En établissant cette base, il était possible de mieux aborder les aspects spécifiques et les applications pratiques des modèles mixtes dans les sections ultérieures.

Ensuite, décrire les effets fixes et aléatoires est important pour comprendre les modèles linéaires mixtes. Les effets fixes permettent d'évaluer l'impact moyen des variables explicatives sur la variable réponse, posant ainsi les bases pour aborder les effets aléatoires et leurs interactions au sein des modèles mixtes. D'autre part, les effets aléatoires ajoutent une dimension de variabilité aux modèles mixtes en modélisant les différences spécifiques entre les groupes d'observations. Ils introduisent des corrélations intra-groupe et permettent de généraliser les résultats à des niveaux non observés. En détaillant à la fois les effets fixes et aléatoires, on complète la description des composantes du modèle linéaire mixte, garantissant ainsi une compréhension approfondie.

Il est important d'établir des règles de décision claires pour choisir entre effet fixe et effet aléatoire dans les modèles mixtes afin d'assurer la pertinence et la précision des analyses statistiques. Ces règles fournissent un cadre méthodologique permettant de structurer la modélisation en fonction de la nature des données et des objectifs de l'étude. En définissant ces critères, on s'assure de sélectionner le type d'effet approprié pour chaque facteur étudié, ce qui influence directement la façon dont la variabilité et les corrélations entre les observations sont modélisées.

Ainsi, les modèles linéaires mixtes, également connus sous le nom de modèles à effets mixtes, sont une classe de modèle qui englobe les modèles linéaires et les modèles à effets aléatoires (sans effet fixe) (MCCULLOCH et al. 2001, CASELLA et al. 2024). Le principe même des modèles linéaires mixtes se trouve dans leur capacité à intégrer à la fois des composantes fixes représentant les effets moyens des variables explicatives et des composantes aléatoires permettant de capturer les variabilités et les corrélations inter-observations. Cette dualité permet de modéliser de manière plus précise et réaliste des phénomènes complexes, en tenant compte à la fois des tendances générales et des variations spécifiques à chaque unité d'observation.

2.1 Définition d'un effet fixe

Un effet fixe peut-être soit une variable quantitative (numérique), soit une variable qualitative nominale ou ordinale (un facteur) avec un nombre fini de niveaux. Dans un modèle linéaire mixte, l'inclusion d'une variable en tant qu'effet fixe permet d'estimer son impact direct sur la variable réponse. Ce sont généralement des variables qui sont contrôlées ou manipulées par l'expérimentateur. Par exemple, dans une étude agronomique, le type de fertilisant appliqué aux cultures (fertilisant A vs fertilisant B) serait considéré comme un effet fixe.

2.2 Définition d'un effet aléatoire

Un effet aléatoire est toujours une variable qualitative ou un facteur. Il s'agit d'une variable groupante où les niveaux du facteur définissent des groupes d'observations. On suppose qu'il existe un nombre infini de niveaux dans la population entière et que nous n'en observons qu'un sous-échantillon aléatoire (nous observons un nombre fini de niveaux parmi un ensemble infini) (GALECKI et al. 2013). Généralement, ce n'est pas un facteur que l'on contrôle et dont on cherche à estimer l'effet de chacun de ses niveaux sur la variable réponse mais toutefois on suppose qu'il joue un rôle sur la variabilité de la variable réponse (variabilité inter-groupe). L'estimation de cette part de variabilité plutôt que de l'effet de chacun des niveaux présente l'avantage de pouvoir généraliser les résultats quel que soit la valeur prise par ce facteur aléatoire, même pour des valeurs non observées. Il permet également d'introduire dans le modèle des corrélations entre les observations. Nous reviendrons sur ce point dans la section (4).

2.3 Comment choisir entre effet fixe et aléatoire ?

Pour prendre une décision entre utiliser un effet fixe ou aléatoire pour un facteur dans un modèle linéaire mixte, plusieurs règles peuvent guider cette sélection. Ces règles sont basées sur des principes théoriques et des considérations pratiques qui visent à optimiser la pertinence et l'efficacité du modèle.

Les règles suivantes orientent cette décision mais ne sont pas exhaustives. Premièrement, un facteur est généralement choisi comme effet aléatoire lorsque les observations qu'il regroupe présentent une corrélation intra-groupe significative (CHEN et al. 2003). Deuxièmement, l'utilisation d'un effet aléatoire est appropriée lorsque l'on souhaite généraliser les résultats à une population plus large pour laquelle on ne dispose que d'un échantillon aléatoire. Troisièmement, si la variabilité capturée par le facteur est d'une importance primordiale par rapport à l'effet spécifique de ses niveaux individuels sur la variable réponse, un effet aléatoire est souvent préféré (CLARKE et al. 2010). Cela permet de modéliser la variabilité inter-groupe tout en simplifiant l'estimation à un seul paramètre de variance, comparé à plusieurs paramètres d'effet fixe pour chaque niveau du facteur.

Enfin, dans des cas où le facteur comporte un nombre élevé de niveaux et où l'estimation précise de l'effet de chaque niveau est complexe avec un effet fixe, opter pour un effet aléatoire peut être plus économique en termes de degrés de liberté (SEARLE et al. 2009).

3 Formulation et hypothèses

Un modèle linéaire mixte peut être présenté soit sous forme individuelle, soit sous forme matricielle. Il est crucial de développer les deux représentations, car chacune fournit une vision complémentaire. La forme individuelle montre comment chaque observation est traitée en tenant compte des effets fixes et aléatoires propres à chaque cas. Cette approche facilite une compréhension détaillée de l'impact de chaque composante sur les données et rend l'analyse plus accessible.

En revanche, la forme matricielle propose une vue d'ensemble plus structurée et généralisée des relations entre toutes les variables du modèle. Elle est particulièrement utile pour modéliser efficacement les dépendances complexes et les interactions entre les observations et les niveaux des facteurs aléatoires, grâce à l'utilisation de matrices de design.

De plus, détailler la construction de la matrice de design Z et du vecteur d'effets aléatoire U est essentiel pour comprendre comment les effets aléatoires sont modélisés dans ces modèles. Une analyse approfondie de ces matrices permet de saisir les relations complexes entre les facteurs aléatoires et les observations, enrichissant ainsi la compréhension globale du modèle.

Enfin, il est important de détailler l'écriture conditionnelle et marginale pour comprendre comment les effets aléatoires sont intégrés dans les modèles mixtes. L'écriture conditionnelle permet d'examiner les effets des variables explicatives en prenant en compte les effets aléatoires spécifiques, tandis que l'écriture marginale offre une vue d'ensemble des effets moyens à travers l'ensemble des observations. En combinant ces perspectives, on obtient une modélisation plus complète, capturant à la fois les variations spécifiques et les tendances générales.

Tout d'abord, il est important d'introduire quelques notations :

— n : le nombre total d'observations

— p : le nombre de variables à effet fixe

— D : le nombre d'effets aléatoires

— Q_d : le nombre de niveaux observés du d$^{\text{ème}}$ facteur aléatoire, $d = 1, ..., D$

— Q : le nombre total de niveaux observés pour l'ensemble des facteurs aléatoires, $Q = \sum_d^D Q_d$

3.1 Forme individuelle

Comme dit précédemment, un modèle linéaire mixte peut s'écrire sous forme individuelle, c'est-à-dire à l'échelle d'une observation. Néanmoins, cette écriture n'est pas généralisable puisqu'elle dépend des données.

Dans le cadre du regroupement d'essais (exemple n°1), les observations d'un caractère d'intérêt tel que le rendement sont réalisées sur des individus issus du croisement de deux géniteurs (un père et une mère). Certains individus partagent ainsi le même père ou la même mère, créant des données groupées. Cette structure nécessite l'utilisation d'un modèle approprié pour tenir compte de l'ensemble du pédigrée, notamment lorsque les individus sont issus de croisement sur plusieurs générations.

Le modèle s'écrit alors :

$$Y_i = \mu + u_i + Z_T T + Z_R R + Z_B B + \epsilon_i \tag{1}$$

— i est l'indice des individus, $i = 1, ..., n$

— Y_i est l'observation du rendement de l'individu i

— u_i est l'effet aléatoire du ième individu, tel que le vecteur de l'ensemble des effets des groupes $u = (u_1, u_2, \ldots, u_n)^T \sim N_n(0, \sigma_u^2 A)$ avec A étant la matrice de parenté connue calculé à partir du pédigrée de taille $n \times n$

— Z_T est la matrice de design pour les essais où chaque ligne correspond à une observation et chaque colonne correspond à un essai spécifique. T est supposé suivre une distribution normale $T \sim N(0, \sigma_T^2 I)$.

— Z_R est la matrice de design pour les répétitions où chaque ligne correspond à une observation et chaque colonne correspond à une répétition spécifique. R est supposé suivre une distribution

normale $R \sim N(0, \sigma_R^2 I)$.

— Z_B est la matrice de design pour les blocs où chaque ligne correspond à une observation et chaque colonne correspond à un bloc spécifique. B est supposé suivre une distribution normale $B \sim N(0, \sigma_T^2 I)$.

— ϵ_i est le résidu associé à l'observation i, tel que le vecteur de l'ensemble des résidus $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T \sim N_n(0, \sigma_E^2 I_n)$ avec I_n étant la matrice identité de taille $n \times n$

3.2 Forme matricielle

N'importe quel modèle mixte peut s'écrire sous la forme matricielle suivante :

$$Y = X\beta + ZU + \epsilon$$

— Y est la variable réponse de longueur n où n est le nombre total d'observation,

— X est la matrice des effets fixes, pouvant contenir une première colonne de 1 pour modéliser un intercept,

— β est le vecteur des effets fixes associées à X. Il représente les effets moyens des variables explicatives sur la variable réponse,,

— Z est la matrice de design permettant de relier chacune des observations à un ou plusieurs niveaux de l'effet aléatoire U (La construction de ces matrices sera décrite par la suite),

— U est le vecteur d'effet aléatoire,

— ϵ est le vecteur des résidus.

Les hypothèses du modèle linéaire mixte sont : le vecteur d'effet aléatoire U est supposé suivre une distribution normale centrée en zéro, $U \sim N(0, G)$. Le vecteur des résidus est supposé suivre une distribution centrée en zéro, $\epsilon \sim N(0, R)$. Il doit y avoir indépendance entre le vecteur d'effet aléatoire U et les résidus. Les différentes formes des matrices G et R seront présentées plus tard.

Les composantes de la variance désignent l'ensemble des paramètres inconnus dans la partie aléatoire, à savoir les effets aléatoires (G) et les résidus (R), qui doivent être estimés. Ces paramètres sont en fait des paramètres de variance ou de covariance.

3.2.1 Construction de Z et U

La matrice Z est une matrice de design ou d'incidence qui joue un rôle important en établissant la connexion entre les observations et les niveaux des facteurs aléatoires dans le modèle. Chaque ligne de Z correspond à une observation spécifique et indique quels niveaux des facteurs aléatoires sont associés à cette observation. Plus précisément, pour chaque observation, la ligne pertinente de Z est une séquence de 0 et de 1, où une valeur de 1 marque l'association de cette observation avec un niveau particulier d'un facteur aléatoire. Cette structure permet de représenter efficacement comment chaque observation est reliée aux différents effets aléatoires présents dans le modèle. Par exemple, dans un modèle où les effets aléatoires sont attribués à des groupes ou à des individus, la matrice Z organise ces associations de manière à ce qu'on puisse facilement comprendre et quantifier l'impact de ces effets sur les observations.

Quant au vecteur d'effet aléatoire U , il contient les effets spécifiques associés aux niveaux des facteurs aléatoires. Sa longueur D correspond au nombre total de niveaux observés pour tous les facteurs aléatoires présents dans le modèle. Par exemple, dans un modèle génétique où les effets aléatoires incluent ceux des pères et des mères, U pourrait être la concaténation des effets de chaque

parent pour tous les individus observés. En d'autres termes, U regroupe les effets aléatoires en une seule entité, facilitant leur manipulation et leur intégration dans le modèle statistique.

La combinaison de Z et U dans le modèle permet ainsi de relier les effets aléatoires aux observations spécifiques, tout en permettant d'estimer et de contrôler ces effets dans les analyses. En comprenant comment ces deux éléments interagissent, on peut mieux appréhender la structure des effets aléatoires dans les données et tirer des conclusions plus précises sur les sources de variation dans les modèles statistiques.

Reprenons **l'exemple n°1 des Regroupement d'essais** (plan alpha-lattice) qui regroupe plusieurs essais. Concentrons nous sur la construction de la matrice de design de l'effet aéaltoire "*Trial*". Pour simplifier l'exemple, on suppose qu'il y a 4 essais qui contiennent tous 3 palmiers (on oublie les répétions et les blocs). L'effet aléatoire essai T est donné par le vecteur : $T = (t_1, t_2, t_3, t_4)^T$, où t_q est l'effet de l'essai q sur la variable réponse (voir eq 1).

Pour construire la matrice Z_T, il nous faut avoir le tableau des données pour voir comment sont organisées les données et ainsi pouvoir relier chacune des observations à l'essai dans laquelle elle a été recueillie. Imaginons que les données sont organisées par essai (les 3 premières observations proviennent de l'essai 1 et ainsi de suite). La matrice Z_T doit alors être de taille 12 x 4. Sur la première colonne, on retrouvera des 1 si les observations proviennent du premier essai et 0 sinon. Il en est de même pour les colonnes 2, 3 et 4 avec les essais 2, 3 et 4 respectivement.

Cela nous donne la matrice suivante :

$$\left(\begin{array}{cccc} 1&0&0&0\\ 1&0&0&0\\ 1&0&0&0\\ 0&1&0&0\\ \hline 0&1&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&1&0\\ \hline 0&0&1&0\\ 0&0&0&1\\ 0&0&0&1\\ 0&0&0&1 \end{array}\right)$$

3.2.2 Ecriture conditionnelle et marginale

Lorsque l'on travaille en conditionnelle, les résultats dépendent des effets des niveaux de U qui ont été prédits. Cela implique que les observations sont conditionnées par l'ensemble des effets aléatoires U.

$$Y|U \sim N(X\beta + ZU, R)$$
$$U \sim N(0, G)$$

L'écriture marginale consiste à intégrer l'effet aléatoire U. Cette approche permet d'accéder à la distribution de Y sans avoir à connaître les effets des niveaux de U :

$$Y \sim N(X\beta, ZGZ' + R) \tag{2}$$

où ZGZ' est ajouté à R, ce qui augmente la variance. Cette généralisation au niveau de la population permet de faire des prédictions pour un nouvel individu, même lorsque le niveau d'effet aléatoire est inconnu ou non observé dans les données d'ajustement.

Cas particulier du modèle animal :

Dans ce modèle animal présenté à l'équation n°1 (eq 1), l'effet des génotypes (u) nous importe généralement. Pour prédire le rendement d'un génotype, un modèle conditionnel par rapport à u est utilisé, tout en intégrant les effets des facteurs Essai, Répétition et Bloc. Le modèle se formule comme suit :

$$\begin{aligned} Y|u &\sim N(\mu + u, \sigma_T^2 Z_T Z_T' + \sigma_R^2 Z_R Z_R' + \sigma_B^2 Z_B Z_B' + \sigma_e^2 I) \\ u &\sim N(0, \sigma_u^2 A) \end{aligned} \tag{3}$$

Cette approche permet de réaliser des prédictions pour un génotype spécifique tout en les généralisant à n'importe quels essais, répétitions et blocs. En intégrant les effets de ces facteurs, les prédictions tiennent compte des variations introduites par ces éléments, ce qui améliore la robustesse et la fiabilité du modèle. Grâce à cette généralisation, les prédictions restent pertinentes et fiables même lorsqu'elles sont appliquées à différents contextes expérimentaux, optimisant ainsi leur applicabilité.

Pour prédire les génotypes des géniteurs non observés, on utilise une matrice de pédigrée qui modélise les relations génétiques entre les individus. Cette matrice est divisée en quatre parties : la première montre les similarités génétiques entre les géniteurs, la deuxième celles entre les géniteurs et leurs descendants, la troisième reprend cette information dans l'autre sens, et la quatrième montre les similarités entre les descendants. En intégrant ces blocs dans un modèle mixte, les effets aléatoires des géniteurs sont estimés en fonction des données phénotypiques de leur descendance et des marqueurs génétiques. Ce processus permet de modéliser la covariance des effets génétiques, garantissant des prédictions robustes et fiables adaptées à divers contextes expérimentaux complexes.

4 Prise en compte de la structure de corrélation

La structuration introduite pas un effet aléatoire permet d'inclure des corrélations spécifiques entre les observations, ce qui permet de capturer les variations non observées au sein des groupes ou des individus. Cette approche est importante pour modéliser correctement les données longitudinales ou les études avec des mesures répétées, où les observations d'un même sujet sont souvent plus similaires entre elles que celles d'autres sujets. Ignorer ces structures peut conduire à des estimations biaisées des paramètres et à une sous-estimation de l'incertitude associée.

La structuration au travers des résidus permet d'identifier et de modéliser les structures de corrélation non capturées par les effets fixes et aléatoires. En analysant les résidus, on peut détecter des paternes non aléatoires qui révèlent des informations importantes sur la nature des erreurs de modélisation ou sur des aspects non pris en compte dans le modèle initial. Cela aide à améliorer la précision des prédictions et à identifier les modifications potentielles nécessaires dans la spécification du modèle.

En intégrant ces deux aspects dans l'analyse statistique, on renforce la validité et la robustesse des conclusions tirées des modèles mixtes. Une modélisation correcte de la structuration introduite par un effet aléatoire et une analyse appropriée des structures résiduelles garantissent que les relations complexes entre les variables sont correctement capturées et interprétées. Cela conduit à des recommandations et des décisions plus éclairées, basées sur des données plus précises et représentatives des processus sous-jacents étudiés.

4.1 Structuration au travers d'un effet aléatoire

Dans cette partie, l'analyse se concentre sur le cas particulier où le modèle statistique inclut un seul effet aléatoire. Dans ce cadre, les structures de corrélation introduites par cet effet peuvent varier

en termes de complexité et de flexibilité. La structure de corrélation par défaut, souvent utilisée dans les logiciels de statistiques, est la **matrice idendité**. Dans ce cas, la matrice G est un multiple de la matrice idendité, ce qui se traduit par des effets aléatoires $U \sim N(0, \sigma_u^2 I_d)$. Cette structure suppose une corrélation constante entre toutes les observations ayant le même niveau aléatoire.

La **matrice diagonale** est une autre structure couramment utilisée, où chaque niveau de l'effet aléatoire a sa propre variance sans covariance entre les différents niveaux. La matrice de covariance G est alors diagonale, reflétant des variances distinctes pour chaque niveau mais aucune corrélation entre eux.

Une structure de corrélation plus complexe est celle où la **matrice G est connue à une constante près**. Ici, $G = \sigma_u^2 A$, avec A étant une matrice de variance-covariance connue, telle qu'une matrice de parenté en génétique. Les niveaux de l'effet aléatoire ne sont pas indépendants, et les effets aléatoires suivent $U \sim N(0, \sigma_u^2 A)$.

La **structure générale** de corrélation est la plus flexible, permettant à tous les éléments de la matrice G d'être non nuls. Cela signifie que chaque niveau aléatoire peut avoir une variance distincte, et toutes les covariances possibles entre les niveaux aléatoires sont prises en compte. Cette structure, bien que très flexible et capable de modéliser des relations complexes, nécessite l'estimation d'un grand nombre de paramètres.

Enfin, la **structure de bloc** divise la matrice de covariance G en sous-matrices pleines le long de la diagonale. Chaque bloc représente les covariances entre certains niveaux d'effet aléatoire, utile pour modéliser des interactions entre plusieurs effets aléatoires. Par exemple, en génétique, pour étudier l'interaction génotype x environnement dans un essai multi-sites, une matrice de covariance en blocs peut être utilisée. Ici, un effet aléatoire de site (environnement) avec une matrice diagonale (une variance par site) en interaction avec un effet aléatoire de génotype avec une matrice de covariance A donne une nouvelle matrice de covariance structurée par blocs.

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \otimes A = \begin{bmatrix} \sigma_1^2 A & 0 & 0 \\ 0 & \sigma_2^2 A & 0 \\ 0 & 0 & \sigma_3^2 A \end{bmatrix}$$

Cette approche permet de capturer des interactions complexes entre les effets aléatoires, et des logiciels comme sommer, BGlr, et BGGE facilitent la mise en œuvre de ces modèles.

4.2 Structuration au travers des résidus

4.2.1 Différentes structures

Les résidus d'un modèle mixte peuvent être structurés de manière à capturer diverses formes de dépendance ou de variation dans les données qui n'auront pas pu être capturées par des effets aléatoires.

La structuration iid : $R = \sigma_e^2 I$ représente une structure de variance résiduelle où chaque terme d'erreur est indépendant et identiquement distribué, avec une variance constante σ_e^2. Cette hypothèse est par défaut dans tous les logiciels statistiques, car elle simplifie les calculs en supposant que les erreurs sont indépendantes et que leur variance est homogène, ce qui facilite l'ajustement du modèle et l'interprétation des résultats.

La structure diagonale modélise des données où les résidus sont supposés indépendants mais avec une variance hétérogène en fonction d'une variable de regroupement.

TABLE 1 – Matrice de la structure diagonale

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_2^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_3^2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_q^2 \end{bmatrix}$$

Pour des données structurées dans le temps, une **structure autorégressive d'ordre 1 (AR(1))** peut être utilisée pour modéliser la corrélation entre les observations successives sur le même sujet. Ce modèle est défini par un paramètre de corrélation, noté ρ, qui représente la force de la corrélation entre les observations successives. La corrélation entre deux observations diminue exponentiellement à mesure qu'elles s'éloignent dans le temps, ce qui permet de capturer les dépendances temporelles dans les données. La structure AR(1) est particulièrement adaptée pour des séries chronologiques où l'influence d'un événement diminue progressivement au fil du temps. L'avantage des structures autorégressive est que leur matrice de précision est creuse (bande diagonale). Par ailleurs, certains logiciels, tels que *AsReml* ou *INLA*, exploitent la structure creuse de cette matrice pour accélérer les calculs et gérer efficacement de grands pédigrées.

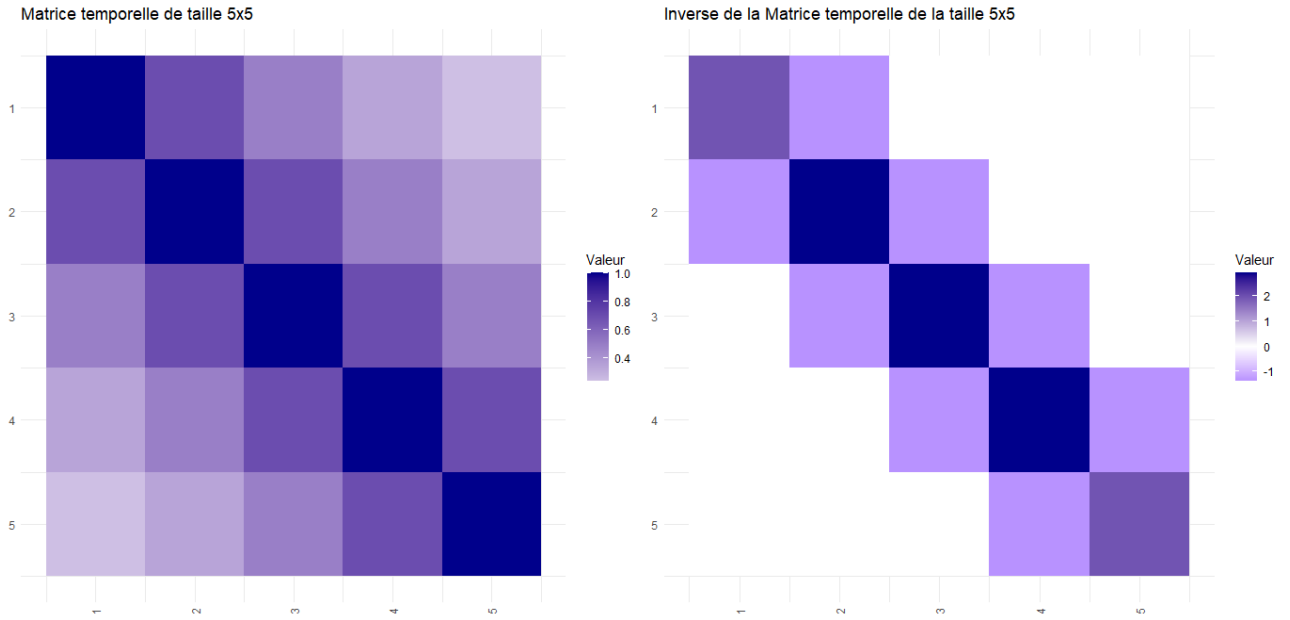

FIGURE 3 – Présentation de la matrice temporelle (AR(1)) et de son inverse

Il est possible d'utiliser diverses structures spatiales pour capturer la dépendance entre les observations en fonction de leur proximité spatiale. Deux classes principales de structures sont couramment utilisées : les structures isotropiques, où la dépendance spatiale est identique dans toutes les directions, et les structures anisotropiques, où la dépendance varie selon la direction. Pour évaluer et choisir parmi ces structures, il est recommandé de calculer le semi-variogramme empirique, qui mesure la variance des écarts entre les observations en fonction de leur distance spatiale.

Le semi-variogramme empirique permet de quantifier comment la variance des différences entre les valeurs mesurées varie en fonction de la distance entre les points de mesure. Pour construire un semi-variogramme empirique, on calcule la variance moyenne des différences entre toutes les paires de valeurs séparées par une distance spécifique, puis on trace ces variances en fonction des distances. L'analyse du graphique résultant fournit des informations cruciales sur la portée de la dépendance spatiale, c'est-à-dire jusqu'à quelle distance les points restent corrélés. Le semi-variogramme aide également à identifier des paramètres tels que le nugget (la variance à très courte distance) et le sill (la variance totale au-delà de la portée). En fournissant une estimation de la structure spatiale, le semi-variogramme empirique guide le choix du modèle de variogramme théorique, essentiel pour l'interpolation spatiale et la modélisation géostatistique, facilitant ainsi une meilleure compréhension et une meilleure gestion des données spatiales. Ce semi-variogramme est souvent calculé en utilisant des distances comme la norme euclidienne (L2) ou la distance de Manhattan (L1).

$$\gamma(distance_h) = 0.5 * Var(\varepsilon_x, \varepsilon_y), \text{ pour tout } x, y \text{ tel que } distance(x, y) \leq h$$

Le variogramme empirique consiste à faire des classes de sitances (ex : [0 ;5[, [5 ;10[, ...) et à calculer la variance entre toutes les observations ayant une distance appartenante à la classe. On peut alors représenter ça graphiquement (1.4.2).

La norme euclidienne est calculée comme la racine carrée de la somme des carrés des différences entre les coordonnées des points. Cette mesure reflète la distance "à vol d'oiseau" dans l'espace, fournissant une mesure directe de la séparation géométrique entre les points. En revanche, la distance de Manhattan est obtenue en additionnant les valeurs absolues des différences entre les coordonnées des points. Cette distance est souvent décrite comme la distance parcourue en suivant un chemin en angle droit, comme les rues d'une grille urbaine.

Chacune de ces distances présente des avantages selon le contexte de l'analyse : la distance Euclidienne est utile pour capturer des différences globales dans un espace continu, tandis que la distance de Manhattan est souvent préférée dans des contextes où les déplacements se font le long d'axes prédéfinis ou lorsque les données sont naturellement alignées sur des grilles.

Les **structures spatiales isotropiques** sont définis par $\gamma(d, a) = 1 - h(d, a)$, où d est la distance spatiale et a est un paramètre de portée contrôlant la décroissance de la corrélation, où h est la fonction de corrélation. Différentes fonctions ont été proposées telles que l'exponentielle, le gaussien, le linéaire, le quadratique rationnel et le sphérique (Jose PINHEIRO et al. 2006). Le tableau (2) détaille les principales fonctions de corrélation et la figure (4) donne une représentation graphique des semi-variogrames de ces fonctions.

Fonctions de Corrélation	**Formule**
Exponentielle	$h(d, a) = exp(-d/a)$
Gaussienne	$h(d, a) = exp[-(d/a)^2]$
Linéaire	$h(d, a) = 1 - (1 - d/a)I(d < a)$
Quadratique Rationnelle	$h(d, a) = (d/a)^2/[1 + (d/a)^2]$
Sphérique	$h(d, a) = 1 - [1 - 1.5(d/a) + 0.5(d/a)^3]I(d < a)$

TABLE 2 – Exemples de fonctions de corrélation et leurs formules correspondantes.

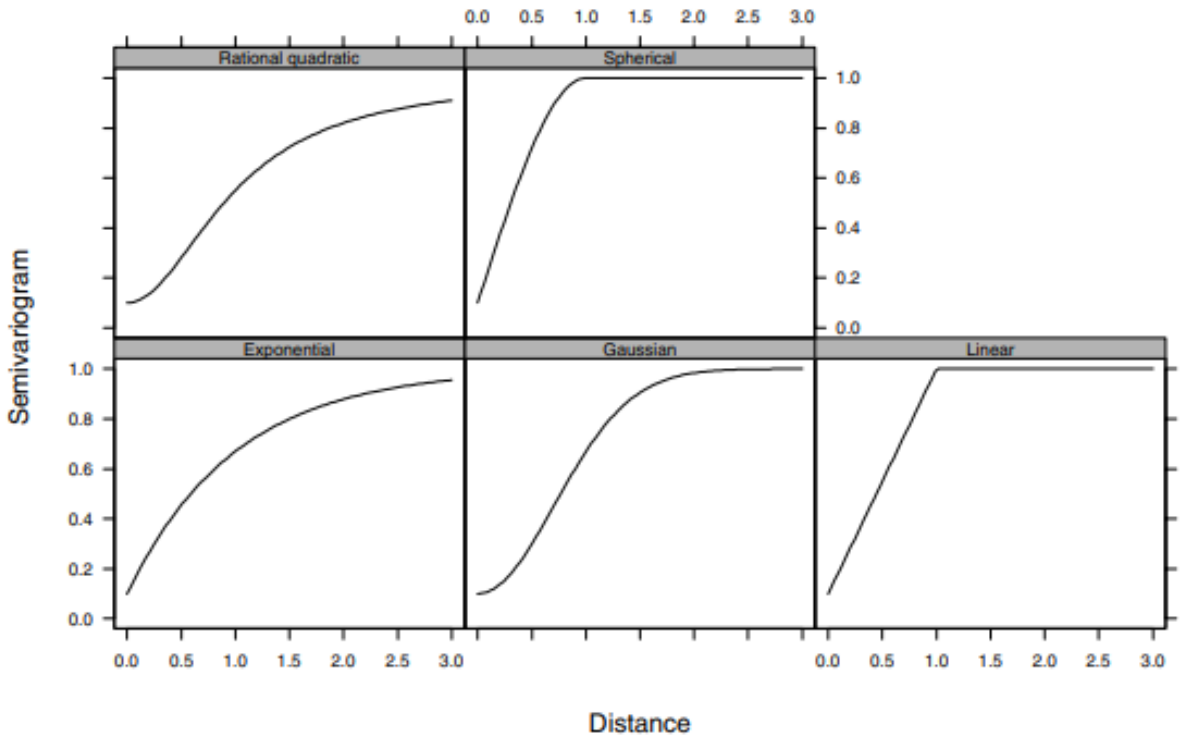

FIGURE 4 – Présentation des variogrammes de chaque structure Jose PINHEIRO et al. 2006

Les **structures spatiales anisotropiques** supposent que la dépendance spatiale varie selon les directions. On peut appliquer des modèles de variogramme différents pour chaque dimension spatiale, comme pour une structuration autorégressive (exponentielle à une dimension), linéaire ou gaussien (tableau 2) sur les lignes et les colonnes. La matrice de variance-covariance est alors définie par un produit de kroneker entre la structure supposée sur les lignes et structure supposée sur les colonnes. Le produit de kronecker présente de bonnes propriétés en algèbre linéaire (l'inverse d'un produit de kronecker est le produit de kronecker des inverses, conservation des formats creux). Des librairies telle que sommer tirent partie de ces propriétés. Il est également possible de modéliser la dépendance spatiale anisotropique au travers de P-splines à deux dimensions. La librairies (mettre le librairie) le propose.

En résumé, le choix de la structure spatiale appropriée dépend de la nature des données et des hypothèses sur la dépendance spatiale. L'analyse du semi-variogramme empirique peut guider ce choix en fournissant des indications sur la portée et la nature de la corrélation spatiale dans les données observées.

4.3 Covariance introduite par un effet aléatoire

Dans les modèles linéaires mixtes, les effets aléatoires permettent d'introduire de la corrélation entre les observations. On peut facilement s'en rendre compte en regardant le modèle marginale (voir eq. 2). La matrice définie par le produit ZGZ' n'est alors pas une matrice diagonale. Cette corrélation dépend évidemment de la matrice de covariance des effets aléatoires G et la matrice de design Z.

4.3.1 Cas d'un effet aléatoire iid

Plaçons nous dans le cas particulier et le plus courant d'un effet aléatoire iid (les niveaux sont indépendants et identiquement distribués). Nous reprenons le modèle *Animal* de l'exemple n°1 (voir eq 1, effets aléatoires "Essai", "Répétition", et "Bloc").

Ce type d'effet aléatoire est conçu pour capturer les variations spécifiques à chaque groupe définie par chacun des niveaux. Cette approche est particulièrement importante lorsque les données présentent des dépendances internes non expliquées par les variables fixes seules.

Si on regarde le modèle conditionnel à l'effet génétique u mais intégré par rapport au effet "Essai", "Répétition", et "Bloc" (modèle semi-marginale) données par l'équation 3, la matrice de covariance est alors définie par la somme de trois matrices blocs diagonales (une par effet aléatoire intégré) additionnées à la matrice diagonale de covariance des résidus :

$$\sigma^2_T Z_T Z'_T + \sigma^2_R Z_R Z'_R + \sigma^2_B Z_B Z'_B + \sigma^2_e I \tag{4}$$

Chacune des ces trois matrices blocs diagonales permettent alors d'introduire de la corrélation entre les observations à différentes échelles. On peut facilement visualiser cela en représentant graphiquement ces matrices. La figure 5-A représente un sous ensemble de la matrice de covariance $\sigma^2_T Z_T Z'_T$ introduite par l'effet aléatoire "Essai". Chaque bloc de couleur représente la corrélation entre les observations issues d'un même essai. (la variance σ^2_T a arbitrairement été fixée à 1). Cette corrélation reflète la structuration présente entre les observations d'un même essai qui ont subis le même environnement. La figure 5-B représente un sous ensemble de la matrice de covariance du modèle semi-marginale (voir eq. 4). On peut alors visualiser la corrélation introduite par les trois effets aléatoires : les grands carrés reflètent la structuration des essais, les moyens carrés reflètent la structuration des répétitions complètes et les petits carrés reflètent la structuration par bloc.

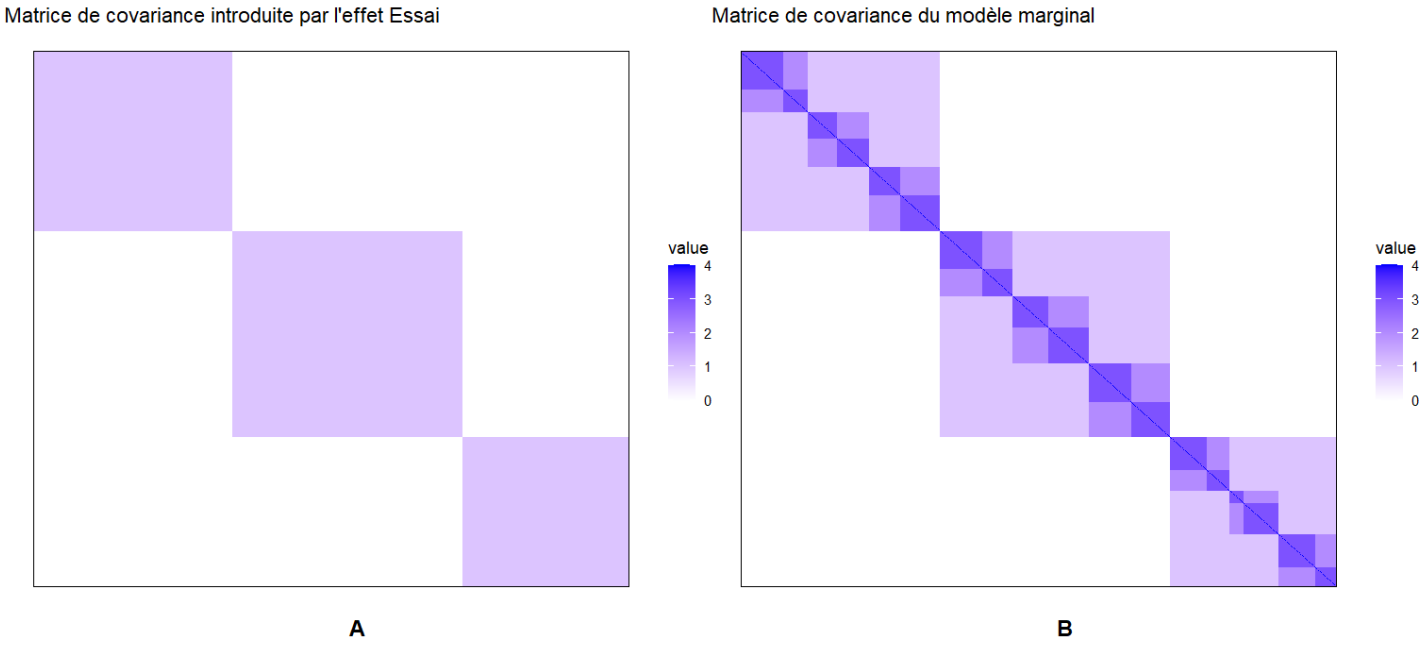

FIGURE 5 – (A) illustre la covariance introduite par l'effet aléatoire "Essai" du modèle *Animal* de l'exemple n°1 (voir eq 4). (B) illustre le sous-ensemble de la matrice de covariance du modèle Animal intégré par rapport aux effets aléatoire "Essai", 'Répétition", "Bloc" (voir eq. 3. Toutes les variances σ^2_T, σ^2_R et σ^2_B ont été arbitrairement fixées à 1.

Comparer ces deux graphiques permet d'évaluer l'impact relatif de chaque effet aléatoire sur la variance totale. Alors que le premier graphique se concentre sur un effet aléatoire spécifique, le second offre une vue consolidée des effets combinés. Initialement, les diverses sources de variation sont identifiées, puis les effets aléatoires sont construits en conséquence.

4.3.2 Cas d'un effet aléatoire de parenté dans un modèle *Animal*

Dans un modèle *Animal* en génétique, l'information de corrélation entre les individus induite par leurs liens de parenté est pris en compte au travers d'un effet aléatoire individuel (autant de niveaux que d'individus) ayant pour matrice de covariance $G = \sigma_u^2 A$ où A est une matrice de corrélation connue calculée à partir de l'information de pédigrée (voir eq 1).

Pour calculer la matrice A, il est nécessaire d'utiliser les informations disponibles dans le pédigrée des individus. La matrice A est construite en déterminant les coefficients de parenté entre chaque paire d'individus (CNAAN et al. 1997). Les éléments diagonaux A_{ii} sont calculés comme $1 + F_i$, où F_i est le coefficient de consanguinité de l'individu i. Les éléments non diagonaux A_{ij} représentent la proportion de gènes partagés entre les individus i et j en fonction de leur ascendance commune.

L'inverse de la matrice A, ou matrice de précision, est typiquement une matrice creuse, ce qui signifie qu'elle contient de nombreux éléments nuls (exemple figure 6). Cette structure creuse est exploitée par certains logiciels tels que AsReml et Inla pour accélérer l'inférence du modèle. Des méthodes algébriques, telles que la décomposition de Cholesky, sont couramment utilisées pour obtenir A^{-1}. La décomposition de Cholesky consiste à factoriser A en une matrice triangulaire inférieure L, telle que $A = LL^T$, sans avoir à calculer directement la matrice de pédigrée A. La librairie ***pedigreemm*** (VAZQUEZ et al. 2010) implémente cette approche dans la fonction ***getAInv***.

Cette méthode est particulièrement avantageuse pour les matrices symétriques définies positives, comme celles rencontrées en génétique quantitative, car elle est numériquement stable et nécessite moins d'opérations que d'autres méthodes de factorisation, telles que la décomposition LU. De plus, la matrice L obtenue de cette manière permet une interprétation des contributions généalogiques, offrant des indications intéressants dans l'analyse de la parenté.

Dans le cadre des modèles mixtes, la décomposition de Cholesky est importante pour le calcul des BLUPs, où elle permet une estimation efficace des valeurs génétiques en tenant compte des effets aléatoires modélisés par la matrice de pedigree.

En outre, la densité de la loi normale multivariée fait intervenir la matrice de précision, ce qui justifie l'intérêt de travailler directement avec cette matrice, comme le font AsReml et Inla. Pour les matrices de grande taille ou lorsque les méthodes directes sont impraticables, les techniques d'approximation basées sur les chaînes de Markov, telles que les algorithmes de Monte Carlo par chaînes de Markov (MCMC), sont souvent employées. Ces algorithmes exploitent la structure creuse de la matrice de précision pour fournir des estimations approximatives de A^{-1}, permettant ainsi de gérer efficacement la complexité computationnelle. Elles sont particulièrement utiles pour traiter de grandes matrices, où les méthodes algébriques directes seraient trop coûteuses en termes de calcul.

5 L'estimation d'un modèle linéaire mixte

Il est important d'expliquer l'estimation d'un modèle linéaire mixte en raison de la diversité des méthodes disponibles et de leurs implications sur l'analyse des données. Les modèles linéaires mixtes, largement utilisés pour analyser des données avec des structures hiérarchiques ou répétées, peuvent être estimés à l'aide de divers librairies et techniques, chacun offrant des avantages et des limitations distincts.

Les librairies comme *NLME* (*Nonlinear Mixed-Effects Models*), *LME4* (*Linear Mixed-Effects Models*), *AsReml* (*Average Information Restricted Maximum Likelihood*), et *Sommer* (*Solving Mixed Model Equations in R*) se basent sur des techniques d'estimation telles que le Maximum de Vraisemblance (ML), le Maximum de Vraisemblance Restreint (REML) ou encore l'*Average Information Algorithm*. Une des méthodes d'inférence phare en fréquentiste est le maximum de vraisemblance

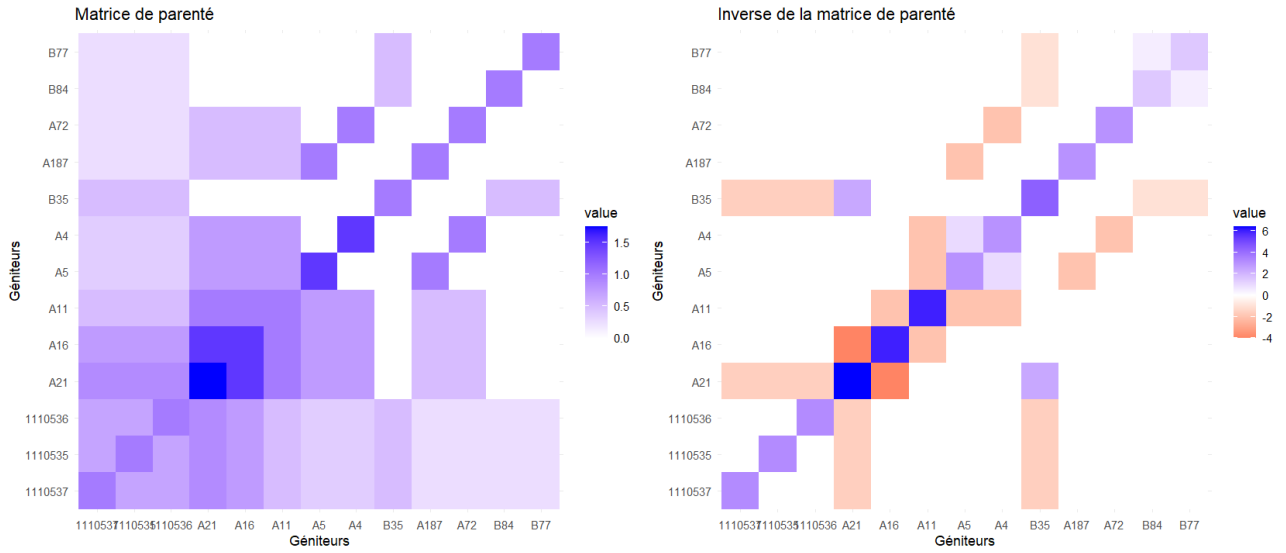

FIGURE 6 – Exemple de matrice de parenté et de son inverse

(ML pour *maximum likelihood*). Cependant, l'estimation par ML produit une estimation biaisée des composantes de la variance parce qu'elle tend à sous-estimer la variance des effets aléatoires. Cela se produit parce que la méthode ML maximise la vraisemblance des données observées sans ajuster pour la perte de degrés de liberté due à l'estimation des paramètres du modèle. En conséquence, les variances estimées des effets aléatoires sont souvent plus petites que les variances réelles.

Par ailleurs, des librairies bayésiennes comme *BGLR* (*Bayesian Generalized Linear Regression*), *BGGE (Bayesian Genomic Generalized Estimation*), et *INLA* (*Integrated Nested Laplace Approximations*) utilisent des algorithmes de Monte Carlo par Chaînes de Markov (MCMC) ou des approximations intégrées de Laplace pour la modélisation bayésienne. Ces techniques bayésiennes sont particulièrement puissantes pour incorporer des informations a priori et pour estimer la distribution a posteriori des paramètres, offrant ainsi une flexibilité et une profondeur d'analyse accrues dans la modélisation statistique.

6 Librairies R pour les modèles mixtes

Examiner les différents librairies ainsi que réaliser des simulations, dans le cadre de ce travail. Chaque librairie offre des fonctionnalités distinctes et des approches variées pour l'analyse des modèles mixtes et la génétique quantitative, ce qui facilite le choix de l'outil le mieux adapté aux besoins spécifiques de l'étude. Par exemple, *INLA* se distingue par sa capacité à gérer efficacement les matrices creuses, offrant ainsi une alternative économique à *AsReml*, qui est payant. Les simulations permettent de tester les performances des différents outils dans des scénarios contrôlés, fournissant des preuves empiriques sur leur efficacité, précision et capacité à gérer des données complexes. En comparant les résultats issus de ces simulations, il est possible d'évaluer les avantages et les limitations de chaque méthode, ce qui aide à faire un choix éclairé et à optimiser les analyses en fonction des exigences du projet.

6.1 Librairies R

Les librairies *NLME, LME4, LME4GS, SOMMER, BGLR, BGGE, BRMS, INLA* et *AsReml* sont des outils clés pour l'analyse des modèles mixtes, chacun ayant ses propres spécificités et avantages en fonction du contexte d'application.

NLME (José PINHEIRO et al. 2017) et *LME4* (**librairielme**) sont toutes deux des librairies R largement utilisées pour ajuster des modèles linéaires et non linéaires à effets mixtes, mais elles diffèrent dans leur approche. *NLME* est particulièrement adapté aux modèles non linéaires, offrant des fonctionnalités robustes pour les données longitudinales et les effets aléatoires, ce qui le rend idéal pour les situations nécessitant une grande flexibilité dans la modélisation. *LME4*, quant à lui, est réputé pour son efficacité dans les modèles linéaires à effets mixtes grâce à des algorithmes d'optimisation avancés, et est souvent préféré pour sa capacité à modéliser des structures de covariance complexes dans divers domaines, de la psychologie à la biologie.

En comparaison, *LME4GS* (CAAMAL-PAT et al. 2021) est une extension de *LME4* conçue spécifiquement pour la sélection génomique. Cette librairie est particulièrement utile pour les analyses génomiques, permettant d'intégrer des effets aléatoires génétiques et facilitant l'estimation des valeurs de reproduction et des effets génétiques additifs dans les populations animales et végétales. *SOMMER*(COVARRUBIAS-PAZARAN 2016) se positionne également dans le domaine de la génétique quantitative, mais avec une orientation plus marquée vers les modèles mixtes complexes, incluant des structures de covariance spécifiques aux relations génétiques et des modèles multi-traits.

Pour les approches bayésiennes, *BGLR* (PÉREZ et al. 2014) et *BGGE* (GRANATO et al. 2018) offrent des outils puissants. *BGLR* se distingue par sa flexibilité dans la modélisation des effets génétiques complexes et est largement utilisé pour les études de sélection génomique, grâce à ses capacités d'estimation bayésienne. *BGGE* étend cette flexibilité en intégrant des interactions génotype-environnement, ce qui le rend particulièrement pertinent pour les études sur l'adaptation génétique et la performance dans des environnements variés.

BRMS (BÜRKNER 2017) se distingue par sa capacité à construire des modèles de régression bayésiens en utilisant Stan. Il offre une grande flexibilité pour définir des modèles hiérarchiques et mixtes, en supportant une large gamme de distributions et de structures de covariance. Son utilisation des algorithmes MCMC dans Stan permet des inférences précises pour des modèles complexes, le plaçant comme un choix privilégié pour des analyses bayésiennes sophistiquées.

INLA (GÓMEZ-RUBIO 2020) propose une alternative efficace aux méthodes MCMC pour les modèles statistiques complexes, en particulier pour les modèles spatiaux et temporels. *INLA* permet une estimation rapide et précise des paramètres et des distributions a posteriori, ce qui est très avantageux pour les applications nécessitant une gestion rapide des modèles spatiaux, comme en épidémiologie et en écologie.

Enfin, *AsReml* (BUTLER et al. 2009) est un logiciel spécialisé dans l'estimation des modèles mixtes linéaires, particulièrement prisé en génétique animale et végétale. Il est reconnu pour sa capacité à traiter de grands ensembles de données et des structures de covariance complexes, fournissant des estimations précises des composantes de variance et des effets aléatoires. *AsReml* est souvent choisi pour les analyses de performance animale et les essais variétaux en sélection génétique, en raison de son efficacité et de sa précision dans ces contextes.

6.2 Simulations

Des simulations de données ont été réalisées à partir d'un pédigrée issu d'un des essais de l'exemple n°1 (voir section 1.4.2), avec pour objectif la prédiction des valeurs génétiques (BLUPs) des géniteurs à partir des descendants. Pour l'ensemble des descendants et géniteurs, une variable réponse a été simulée suivant le modèle :

$$
\begin{aligned}
Y &= \mu + u + e \\
u &\sim N(0, \sigma_u^2 A) \\
e &\sim N(0, \sigma_e^2 I)
\end{aligned}
$$

avec $\mu = 10$, $\sigma_u^2 = 2$, A la matrice de pédigrée issu de données réelles et $\sigma_e^2 = 1$. Le jeu de données simulées a ensuite été séparé en deux avec les descendants d'un côté pour l'entraînement des modèles et les géniteurs de l'autre pour tester les modèles. Ce jeu de données comportes 1227 descendants et 57 géniteurs pour un total de 1284 individus.

Pour l'ajustement du modèle, plusieurs librairies R ont été utilisés, à savoir *lme4GS, sommer, BGLR, BGGE, BRMS, INLA* et*AsReml* ont été mise en oeuvre. Chaque librairie a été évalué en termes de précision (R^2) et de robustesse en comparant les variances définies au variances estimées des prédictions. Chacun a permis l'ajustement de modèles mixtes linéaires, intégrant des matrices de covariance génétique pour estimer les BLUPs. Les modèles ont été spécifiés avec des effets aléatoires pour les individus, et les matrices de covariance ont été utilisées pour représenter les relations génétiques.

Les valeurs prédictives des BLUPs pour les descendants ont été obtenues à partir des effets aléatoires estimés par chaque modèle. Ensuite, les prédictions des géniteurs ont été calculées en utilisant les équations mixtes de Henderson, multipliant les BLUPs des descendants par les sous-matrices appropriées de la matrice de relations génétiques : $\hat{u}_g = A_{gd}\hat{u}_d$ où $\hat{u}_g$ représente les BLUPs prédites pour les géniteurs, $\hat{u}_d$ représente les BLUPs estimées pour les descendants, et A_{gd} est la sous-matrice de la matrice de relations génétiques A, qui décrit les relations entre les géniteurs et les descendants. L'inverse de cette matrice de pédigré des génétieurs étant : $\hat{u}_d = A_{gd}^{-1}\hat{u}_g$.

Le tableau 3 présente, pour chaque librairie, les estimations de la variance de l'effet aléatoire génétique ainsi que celle des résidus, le R^2 calculé sur les descendants (données d'entrainement) et celui sur les géniteurs (données de test). Le R^2, mesure la proportion de la variance totale expliquée par le modèle. Un R^2 proche de 1 indique que le modèle explique bien la variabilité des données, tandis qu'un R^2 proche de 0 indique que le modèle n'explique pas bien cette variabilité. L'objectif de cette section est de comparer les performances d'estimation des différentes librairies au travers de simulation. Nous avons simulé des données à partir d'un pédigré issu de l'exemple n°1 (1.4.1).

Librairies	**Variance Aléatoire Estimée**	**Variance Résiduelle Estimée**	**R^2 descendants**	**R^2 géniteurs**
LME4GS	2,72	1,69	0.78	0,36
SOMMER	2,72	1,69	0.78	0,36
BGLR	2,77	1,69	0.77	0,38
BGGE	3,25	0,25	0.82	0,38
BRMS	2,82	1,64	0,82	0,36
INLA	2.84	1.59	0.82	0.36
AsReml	2.72	1.69	0.78	0.36

TABLE 3 – Tableau des Variances estimées et du R^2 pour les différents librairies

Les valeurs estimées des variances aléatoires présentent des différences entre les méthodes utilisées. Par exemple, la variance estimée varie de 2,72 pour les méthodes comme *LME4GS*, *SOMMER*, et *AsReml*, jusqu'à 3,25 pour *BGGE*. La variance résiduelle varie également, avec *BGGE* affichant une valeur particulièrement basse à 0,25. Il est également notable que l'estimation de la variance aléatoire pour *BGGE* soit supérieure à celle des autres librairies. Cela pourrait s'expliquer par les spécificités de la méthode d'estimation utilisée, notamment la décomposition en valeurs singulières (SVD), qui permet une capture plus détaillée de la structure des données. En somme, les différences entre les méthodes sont subtiles, ce qui signifie que les méthodes offrent des performances globalement comparables mais avec des nuances spécifiques.

Coté performance de prédiction, toutes les librairies donnent des R2 du même ordre de grandeur

avec des valeurs allant de 0.77 à 0.82 sur les descendants et de 0.36 à 0.38 sur les géniteurs. On observe ainsi que la librairie *BGGE* donnant des estimations de composantes de la variance différentes des autres librairies performe aussi bien en terme de prédiction.

Ces simulations sont donc rassurantes et montrent que toutes les librairies utilisées sont bien adaptées pour l'inférence de modèle *Animal*. Nous avons cependant observer des différence non négligeables en terme de temps de calcul. Nous reviendrons sur ce point dans la section suivante.

7 Exemples d'application des différents librairies et données

Présenter des exemples d'applications pour les différents librairies est nécessaire pour illustrer la pertinence et l'efficacité de chaque outil dans des contextes spécifiques. En montrant comment chaque librairie est utilisé dans des études réelles, on peut mettre en évidence les capacités uniques et les avantages de chaque méthode pour des types d'analyses particuliers. Par exemple, SOMMER est démontré comme particulièrement utile pour la génétique quantitative en agriculture. Ces exemples aident non seulement à comprendre les applications pratiques des outils, mais aussi à faire des choix informés en fonction des exigences spécifiques du projet. En fournissant des cas concrets, il devient plus facile de comparer les performances et les adéquations des différentes méthodes, facilitant ainsi la sélection de la librairie le mieux adapté pour répondre aux objectifs de recherche.

7.1 Jeux de données n°1 : Regroupement d'essai

Pour évaluer les composantes de variance liées aux effets génétiques, aux essais, aux répétitions et aux blocs dans nos données, les mêmes modèles statistiques ont été ajustés en utilisant différentes librairies R : *LME4GS, SOMMER, BGLR, BGGE, INLA*, et *AsReml*. Il est important de rappeler que le jeu de données se compose de 28 essais, 6 répétitions, 30 blocs et 30 409 génotypes, offrant une base solide pour l'analyse statistique. Le modèle utilisé est représenté par l'équation (1). Les performances de ces modèles ont été comparées sur la base des variances estimées pour chaque effet ainsi que du R^2 sur les données d'entraînement, fournissant ainsi une vue d'ensemble de l'ajustement des modèles. Les temps d'exécution ont également été relevés afin de comparer la rapidité de chaque librairie.

L'un des indicateurs clés utilisés pour évaluer les modèles est l'héritabilité, notée h^2, qui mesure la proportion de la variance phénotypique totale attribuable à la variance génétique additive. L'héritabilité est calculée à partir des composantes de variance à l'aide de la formule suivante :

$$H2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_T^2 + \sigma_R^2 + \sigma_B^2 + \sigma_E^2} \tag{5}$$

où σ_A^2 représente la variance génétique additive et σ_E^2 la variance résiduelle. Une héritabilité élevée indique que la variation observée dans le trait étudié est principalement due à des différences génétiques, ce qui est crucial pour évaluer le potentiel de sélection dans les programmes d'amélioration génétique.

Le tableau (4) présente l'estimation des composantes de la variance, les R^2 ainsi que les temps d'exécution pour chaque librairie.

Le tableau 4 montre des différences en terme de qualité d'estimation et de prédiction.

Nous constatons que les librairies LME4GS, SOMMER et BGLR donnent des estimations similaires avec une variance génétique proche de 10, des variances Essai, Répétition et Bloc inférieures à 1 et une variance résiduelle autour de 5.5. Les héritabilités (voir eq : 5) et les R2 sont également similaires (autour de 0.6 et 0.82) respectivement.

Librairies	σ^2_u	σ^2_T	σ^2_R	σ^2_B	σ^2_e	H2	R^2	Temps d'exécution
LME4GS	9.31	0.58	0.27	0.47	5.78	0.57	0.83	8 min
SOMMER	8.77	0.70	0.30	0.50	6.06	0.54	0.81	1h40
BGLR	9.93	0.31	0.48	0.006	5.49	0.61	0.84	1h20
BGGE	10.89	20.97	5.02	1.76	4.98	0.25	0.87	42 min
INLA	25.38	0.17	0.30	0.45	0.13	0.96	0.99	12s
AsReml	19.38	0.24	0.27	0.46	2.49	0.84	0.97	1s

TABLE 4 – Tableau des Variances et R^2 pour les différentes librairies

La librairie BGGE donne, comme sur les simulations, des estimations qui diffèrent largement des autres librairies avec des variances Essai, Répétition et blocs bien plus élevées ce qui implique une héritabilité bien plus faible (0.25). Toutefois la qualité prédictive ne semble pas impactée et est même supérieure aux trois librairies précédentes (R2 = 0.87).

Enfin, les librairies INLA et AsReml donnent des estimations assez similaires avec un variance génétique 2 fois plus élevée (25 et 19 respectivement) par rapport à celles estimées avec les autres librairies (autour de 10). Les estimations des variances Essai, Répétition et Bloc sont quand à elles inférieures à 1. Les estimations de la variance résiduelle sont plus faible qu'avec les autres librairies (0.13 et 2.49 respectivement). Ces estimations impliquent des héritabilités bien suppérieures (0.96 et 0.84 respectivement). La qualité prédictive de ces deux librairies semble aussi très bonne avec un R2 proche de 1.

Les libraies INLA et AsReml semblent être plus performante en terme de prédiction par rapport aux autres librairies au vu des R2 calculé. Toutefois ces R2 ont été calculés sur les données d'entraînement et peuvent présenter du sur-ajustement.

Pour finir, nous avons observé des différences de temps d'exécution considérables. Rappelons ici que les données analysé sont constituées de 6 essais parmi les 33 disponibles. Cela totalise 7132 individus (génotypes). Nous pouvons constater que les librairies INLA et AsReml ont été peut impacté avec des temps d'exécution de l'ordre des secondes contrairement aux autres librairies. Les librairies INLA et AsReml sont adapté pour les grands jeux de données avec des optimisations de calcul (travail sur l'inverse des matrices de covariance avec format creux par exemple). Ce sont les deux seuls librairies qui ont pu analyser l'ensemble des 33 essais totalisant plus de 33 000 génotype et ceux en des temps raisonnables de l'ordre de la minute pour INLA et des secondes pour AsReml (ces analyses ne sont pas présenté dans ce rapport).

7.2 Jeux de données n°2 : Plan Alpha-Lattice

Plusieurs modèles statistiques ont été inférés à l'aide de différentes librairies R pour prédire les descendants en tenant compte des effets aléatoires et des structures de covariance spécifiques (spatiales) (5). Les librairies utilisées incluent *NLME, SOMMER, BRMS,* et *BGLR*. La performance de ces modèles a été comparée à travers plusieurs critères, notamment l'AIC (*Akaike Information Criterion*), le coefficient de détermination R², et les résultats du test de Moran. On rappelle que le jeu de données se compose de 2 répétitions, 4 blocs et 145 génotypes, offrant une base solide pour l'analyse statistique.

Le test de Moran est un indicateur utilisé pour détecter la présence d'autocorrélation spatiale dans les résidus d'un modèle. Autrement dit, il permet de vérifier si les erreurs de prédiction sont spatialement corrélées, ce qui indiquerait une mauvaise prise en compte de la structure spatiale dans le modèle. Le test de Moran repose sur deux hypothèses principales. L'hypothèse nulle (H0) stipule qu'il

n'y a pas d'autocorrélation spatiale dans les résidus, ce qui signifie que les erreurs sont distribuées de manière aléatoire dans l'espace. En d'autres termes, les résidus ne présentent aucune tendance géographique particulière. À l'inverse, l'hypothèse alternative (H1) postule l'existence d'une autocorrélation spatiale dans les résidus. Cela implique que des valeurs géographiquement proches ont tendance à présenter des erreurs similaires, suggérant une influence spatiale non capturée par le modèle.

Chaque modélisation a été réalisé à modèle équivalent, mais avec des approches différentes pour modéliser la corrélation spatiale. La librairie NLME a appliqué une structure de corrélation spatiale sphérique prenant en compte la dépendance entre les observations en fonction de leurs coordonnées spatiales (x et y) (tableau : 2). En revanche, le librairie SOMMER a utilisé des splines bidimensionnelles pour intégrer la corrélation spatiale, permettant ainsi de capturer des variations spatiales complexes de manière lisse. La librairie BRMS, quant à lui, implémente un structure SAR (Spatial Autoregressive) en utilisant une matrice de voisinage pour modéliser la dépendance spatiale explicitement intégrant ainsi une matrice de poids spatiaux (M). Pour la librairie BGLR, la corrélation spatiale a été modélisée à travers un modèle SAR (Spatial Autoregressive) basé sur les distances euclidiennes entre les observations avec un hyperparamètre rho de 0,95 (équivalent à une structure isotropique exponentielle avec une porté de 20, voir tableau : 2).

Le tableau 5 présente les critères AIC ainsi que le R² pour chaque librairie. L'AIC (*Akaike Information Criterion*) est un critère utilisé pour comparer la qualité relative de plusieurs modèles statistiques en fonction de leur ajustement aux données et de leur complexité. Il favorise les modèles qui offrent un bon ajustement tout en pénalisant ceux qui sont trop complexes. L'AIC est donné par la formule suivante :

$$\text{AIC} = 2k - 2\ln(L)$$

où k représente le nombre de paramètres estimés dans le modèle, et L est la valeur du maximum de vraisemblance du modèle. Un AIC plus faible indique un modèle plus performant en termes d'équilibre entre précision et complexité.

Ce tableau présente également la p-value du test de Moran, qui permet de tester la présence d'autocorrélation spatiale dans les résidus. Une p-value faible indiquerait une autocorrélation significative, suggérant que des structures spatiales n'ont pas été correctement capturées par le modèle.

Librairies	**AIC**	**R^2**	**Test de Moran (p-value)**
NLME	-117.18	0.32	< 0,001
SOMMER	-118.29	0.32	< 0,001
BRMS	-226.46	0.30	< 0,001
BGLR	NA	0.70	0.53

TABLE 5 – Tableau des AIC, Variances et R^2 pour les différents librairies

L'AIC permet de comparer la qualité des modèles : une valeur plus faible, voire négative, indique un meilleur ajustement. Notamment, le modèle ajusté avec *BRMS* a présenté le plus faible AIC (-226.46), suggérant un meilleur ajustement par rapport aux autres modèles. Cependant, *BGLR* est un modèle bayésien, il n'est donc pas possible de calculer l'AIC puisque c'est un critère fréquentiste. Il était possible de calculer le DIC (déviance information bayésien) mais cela n'est pas comparable aux autres modèles donc il n'a pas été calculé. L'AIC a pu être calculé pour *BRMS* malgré le fait que ce soit une librairie bayésienne. Les modèles bayésiens, comme celui ajusté avec *BGLR*, utilisent des critères différents pour l'évaluation de la qualité de l'ajustement. Le R² de *BGLR* (0.70) est également plus élevé que celui des autres modèles, indiquant une proportion plus élevée de variance expliquée.

Le test de Moran a été utilisé pour évaluer la corrélation spatiale résiduelle après ajustement du modèle. Une p-value inférieure à 0,05 indique qu'il reste de la corrélation spatiale non expliquée, ce qui suggère que les modèles n'ont pas complètement capturé les effets spatiaux. Tous les modèles, sauf celui ajusté avec *BGLR*, ont montré des p-values inférieures à 0,001, indiquant une corrélation spatiale résiduelle significative. Cependant, *BGLR* a présenté une p-value plus élevée (0,53), ce qui pourrait indiquer que ce modèle a mieux géré les effets spatiaux.

En conclusion, bien que certains modèles présentent de meilleures performances en termes d'AIC et de R^2, la persistance d'une corrélation spatiale résiduelle suggère que des améliorations sont nécessaires pour modéliser correctement la structure spatiale dans les données.

7.3 Alternatives à *AsReml*

Une alternative notable à *AsReml* pour l'estimation des modèles mixtes linéaires est *INLA*. *INLA* propose une approche innovante pour l'inférence bayésienne dans les modèles statistiques complexes, notamment ceux impliquant des effets spatiaux et temporels. Contrairement aux méthodes classiques de maximum de vraisemblance restreinte utilisées par *AsReml*, *INLA* utilise des approximations de Laplace imbriquées pour obtenir des estimations précises et rapides des distributions a posteriori.

L'une des principales forces de *INLA* est de pouvoir gérer des effets aléatoires de grande taille délors que leurs inverse de matrice de covariance sont creuses. Les matrices creuses, qui contiennent un grand nombre d'éléments nuls, sont courantes dans les grands modèles statistiques, en particulier dans les contextes de génétique et d'écologie où les matrices de précision sont souvent de grande taille. *INLA* utilise des techniques telles que la décomposition de Cholesky de la matrice de parenté, permettant une gestion mémoire et un temps de calcul considérablement réduits par rapport aux matrices denses. Cela rend l'analyse de grands ensembles de données et de modèles complexes plus rapide et plus économique en termes de ressources computationnelles.

En comparaison, *AsReml* est un logiciel payant, ce qui peut représenter un obstacle pour certains chercheurs et praticiens. Bien que *AsReml* soit puissant et capable de gérer des modèles mixtes complexes, le coût associé à son utilisation peut être un désavantage significatif.

En contraste, *INLA* offre une solution open-source et est optimisé pour les matrices creuses, ce qui améliore l'efficacité de l'inférence bayésienne et réduit les coûts computationnels. Cette capacité à gérer les matrices creuses, combinée à son accessibilité sans frais, fait de *INLA* une alternative particulièrement compétitive à *AsReml*, surtout pour les applications nécessitant une analyse rapide et précise de modèles mixtes tout en minimisant les coûts.

Il a également été réalisé l'analyse sur les 33 essais de l'exemple n°1 (1.4.1), portant sur le palmier à huile, qui comprennent plus de 30 000 génotypes. Cette vaste quantité de données a nécessité une évaluation rigoureuse des outils d'analyse statistique disponibles. Il en ressort que, parmi les différentes librairies testées, seules INLA et ASReml ont été capables de traiter ces données de manière efficace et en des temps raisonnables. Plus précisément, INLA a terminé l'analyse en 1 minute 30 secondes, tandis qu'ASReml a démontré une rapidité impressionnante avec un temps de traitement de seulement 10 secondes. Ces résultats mettent en évidence l'efficacité et la robustesse de ces deux outils pour gérer des ensembles de données complexes à grande échelle.

8 Conclusion

En conclusion, ce guide relatif aux modèles mixtes a été réalisé, complété, et mis à disposition sur GitLab pour consultation. Ce document constitue une ressource essentielle pour quiconque souhaite acquérir une compréhension approfondie des méthodes et pratiques associées aux modèles mixtes,

tout en fournissant des instructions détaillées pour leur mise en œuvre. En outre, il offre un cadre théorique solide accompagné d'exemples pratiques, facilitant ainsi l'application des modèles mixtes à des problèmes réels.

L'analyse approfondie menée au cours de ce projet a permis d'identifier *INLA* comme l'alternative la plus pertinente à *AsReml* pour notre étude spécifique. Cette méthode a démontré une efficacité remarquable, en particulier dans le traitement de données complexes où la précision des résultats est cruciale. Comparée à *AsReml*, *INLA* offre des performances de calcul comparables, ce qui en fait un choix judicieux pour les futurs travaux nécessitant une modélisation rapide et précise.

Pour les étapes futures, plusieurs perspectives prometteuses s'ouvrent. L'une des priorités sera l'implémentation de la validation croisée pour certaines bibliothèques, telles que *INLA*, afin de renforcer la robustesse des modèles développés. De plus, l'exploration de l'utilisation de *BGLR* pour la modélisation de la corrélation spatiale dans le cadre des modèles mixtes, comme illustré dans l'exemple 2, représente une piste intéressante à suivre. Ces démarches devraient non seulement améliorer la précision des analyses, mais aussi élargir les possibilités d'application des modèles mixtes dans des contextes de plus en plus variés. Il serait également pertinent de poursuivre l'étude des performances de ces méthodes à travers un jeu de données de validation distinct, non utilisé lors de l'ajustement initial. Cette étape permettra d'évaluer la généralisation des modèles et d'identifier les éventuelles limitations à surmonter.

D'un point de vue personnel, cette expérience a été particulièrement enrichissante. Elle m'a offert l'opportunité d'approfondir mes connaissances en matière de modèles mixtes et de renforcer mes compétences en analyse statistique avancée. La découverte et l'application de nouveaux outils ont non seulement élargi mes horizons, mais ont également ouvert de nouvelles perspectives pour la recherche future dans ce domaine. Je me sens désormais mieux équipé pour aborder des problématiques complexes et contribuer de manière significative à l'avancement des méthodes statistiques en recherche scientifique.

Bibliographie

I. **Articles**

Cnaan, Avital, Laird, Nan M. & Slasor, Peter (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16(20), 2349–2380.

Foulley, Jean-Louis (2003). Le modele lineaire mixte. *Cours accessible a l'adresse* $http : //pbil.univ-lyon1.fr/members/fpicard/franckpicard_fichiers/pdf/cours.foulley.pdf$.

Baragatti, Meıli (s. d.). Modele lineaire, puissance statistique, modeles lineaires mixtes et modeles lineaires generalises : entre theorie et pratique.

Clarke, Paul, Crawford, Claire, Steele, Fiona & Vignoles, Anna F. (2010). The choice between fixed and random effects models : some considerations for educational research. *IZA Discussion Paper*.

Lee, H.S. & Lim, J.H. (2013). Statistical librairie for the social sciences. *JypHyunJae Publication*.

Chen, Zhen & Dunson, David B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4), 762–769.

Pinheiro, José et al. (2017). librairie 'nlme'. *Linear and nonlinear mixed effects models, version*, 3(1), 274.

Bates, Douglas et al. (2015). librairie 'lme4'. *convergence*, 12(1), 2.

Caamal-Pat, Diana et al. (2021). lme4GS : an R-librairie for genomic selection. *Frontiers in Genetics*, 12, 680569.

Covarrubias-Pazaran, Giovanny (2016). Genome-assisted prediction of quantitative traits using the R librairie sommer. *PloS one*, 11(6), e0156744.

Pérez, Paulino & de los Campos, Gustavo (2014). BGLR : a statistical librairie for whole genome regression and prediction. *Genetics*, 198(2), 483–495.

Granato, Italo et al. (2018). librairie 'BGGE'.

Bürkner, Paul-Christian (2017). brms : An R librairie for Bayesian multilevel models using Stan. *Journal of statistical software*, 80, 1–28.

II. **Livres**

Pinheiro, Jose & Bates, Douglas (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.

McCulloch, Charles E., Searle, Shayle R. & Neuhaus, John M. (2001). *Generalized, linear, and mixed models*. Wiley Online Library.

Searle, Shayle R., Casella, George & McCulloch, Charles E. (2009). *Variance components*. John Wiley & Sons.

Galecki, Andrzej & Burzykowski, Tomasz (2013). *Linear mixed-effects model*. Springer.

Gómez-Rubio, Virgilio (2020). *Bayesian inference with INLA*. Chapman and Hall/CRC.

III. **Rapports Techniques**

Butler, D. G., Cullis, Brian R., Gilmour, A. R. & Gogel, B. J. (2009). *ASReml-R reference manual*. The State of Queensland, Department of Primary Industries and Fisheries, Brisbane.