



## RAPPORT DE STAGE DE FIN D'ÉTUDES

Structure d'accueil : ..... CIRAD

Thème du stage : Analyse des effets du type et du niveau  
de fertilisation sur l'évolution du rendement fourrager

et de la teneur du sol en Azote et en CEC à la Réunion

Lieu de stage : CIRAD-EMVT Pôle Elevage

Ville : Saint-Pierre, Réunion ... Pays : France



MODELLING DATA, CREATING KNOWLEDGE  
MODELISER LES DONNÉES, CRÉER DU SAVOIR

## Promotion 2013

Maître de stage : Emmanuel TILLARD

Responsable de filière : Myriam VIMOND

Tuteur pédagogique : Myriam VIMOND

Stagiaire: ...Lore LE PABIC.....



## Remerciements

Je tiens avant tout à remercier Emmanuel TILLARD, mon maître de stage pour son encadrement, ses conseils et son implication dans mon travail. J'adresse également mes remerciements à Paulo SALGADO, chercheur agronome, pour ses explications, sa disponibilité et pour l'attention toute particulière qu'il a porté à mon travail.

Je remercie Maëva MIRAILLES-BRUNEAU, chargée d'études agronomiques, et Amandine PIERRE, ingénieur agronome, dont les conseils et encouragements ont été précieux.

Merci à Gisèle MOREL, l'assistante de l'équipe, pour toute l'attention qu'elle a prêtée au bon déroulement de mon stage, l'aide et les informations qu'elle m'a fournies.

Je souhaite par ailleurs remercier l'ensemble de l'équipe Pôle Élevage pour l'accueil qui m'a été réservé.



# Table des matières

<b>Introduction générale</b>	<b>1</b>
<b>1 Dispositif expérimental et données disponibles</b>	<b>3</b>
1.1 Le dispositif expérimental . . . . .	3
1.2 Les données . . . . .	5
1.2.1 Effet de la fertilisation sur le rendement fourrager . . . . .	5
1.2.2 Effet de la fertilisation sur la teneur en Azote des sols . . . . .	6
1.2.3 Effet de la fertilisation sur la CEC . . . . .	6
<b>2 Structure de variance-covariance du rendement fourrager</b>	<b>7</b>
2.1 Problématique et description de l'étude . . . . .	7
2.2 Les structures classiques de covariance . . . . .	8
2.2.1 Absence de structure . . . . .	8
2.2.2 Structure en symétrie composée . . . . .	8
2.2.3 Structure auto-régressive d'ordre 1 . . . . .	8
2.3 Description des données . . . . .	9
2.4 Analyse du rendement fourrager sur le site de SICALAIT . . . . .	10
2.4.1 Analyse préliminaire : ANOVA à deux facteurs . . . . .	10
2.4.2 Détection des anomalies . . . . .	12
2.4.3 Le modèle à effets fixes pour données répétées . . . . .	13
2.4.4 Choix de la structure de variance-covariance . . . . .	14
2.4.5 Résultats . . . . .	15
2.5 Analyse du rendement fourrager sur les sites représentatifs à la Réunion . . . . .	16
2.5.1 Analyse préliminaire : ANOVA à trois facteurs . . . . .	16
2.5.2 Détection des anomalies . . . . .	17
2.5.3 Le modèle général avec l'effet du site pour données répétées . . . . .	17
2.5.4 Choix de la structure de variance-covariance . . . . .	19
2.5.5 Résultats . . . . .	20
2.6 Interprétation agronomique des résultats . . . . .	21
2.6.1 Des différences entre les sites . . . . .	22
2.6.2 Des différences entre les saisons . . . . .	22
2.6.3 Des différences entre les modalités de fertilisation . . . . .	22
2.7 Conclusion . . . . .	23
<b>3 Analyse de l'évolution de la teneur en Azote et de la CEC</b>	<b>24</b>
3.1 Problématique et description de l'étude . . . . .	24
3.2 Analyse exploratoire des données longitudinales . . . . .	24
3.2.1 L'hypothèse de linéarité de l'Azote . . . . .	25
3.2.2 L'hypothèse de linéarité de la CEC . . . . .	26
3.3 Stratégie de la modélisation . . . . .	27
3.3.1 Présentation du modèle additif semi-paramétrique . . . . .	27
3.3.2 Critères et outils de comparaison . . . . .	28
3.4 Le modèle de la teneur en Azote des sols . . . . .	28
3.4.1 Quatre fonctions de lissage et corrélation temporelle . . . . .	28
3.4.2 Hétérogénéité intra-site . . . . .	33
3.4.3 Interprétation agronomique des résultats . . . . .	36
3.5 Le modèle de la CEC des sols . . . . .	38
3.5.1 Un lissage temporel selon le type de site . . . . .	38
3.5.2 Un lissage temporel par fertilisation . . . . .	44

3.5.3	Interprétation agronomique des résultats . . . . .	47
3.6	Conclusion . . . . .	48
	<b>Conclusion générale</b>	<b>49</b>
	<b>Bibliographie</b>	<b>51</b>
	<b>Annexes</b>	<b>57</b>
<b>A</b>	<b>Présentation de la structure d'accueil</b>	<b>58</b>
<b>B</b>	<b>Le cycle de l'Azote</b>	<b>59</b>
<b>C</b>	<b>L'intervalle entre coupes de fourrage</b>	<b>60</b>
<b>D</b>	<b>Analyse exploratoire du rendement fourrager</b>	<b>62</b>
<b>E</b>	<b>Le modèle additif et les splines</b>	<b>66</b>
E.1	Définition du modèle . . . . .	66
E.2	Les fonctions splines . . . . .	66
E.3	Estimation via la modélisation mixte . . . . .	67
E.4	Interactions avec les termes de lissage . . . . .	68
<b>F</b>	<b>Analyse exploratoire de la teneur en Azote et de la CEC des sols</b>	<b>70</b>
F.1	Analyse exploratoire de la teneur en Azote . . . . .	70
F.2	Analyse exploratoire de la CEC . . . . .	71
<b>G</b>	<b>Estimations des modèles</b>	<b>72</b>
G.1	Le rendement fourrager . . . . .	72
G.2	La teneur en Azote . . . . .	73
G.3	La CEC . . . . .	75

# Introduction générale

Dans l'écosystème de l'île de la Réunion, les quelques 10000 hectares de prairies entretenues par les éleveurs contribuent à de multiples services. Elles participent activement à l'emploi agricole, à la qualité du paysage, au contrôle des feux et de l'érosion et à la qualité de l'eau. C'est avant tout un capital écologique qu'il faut à la fois préserver, entretenir et valoriser quant à sa capacité à produire de la biomasse fourragère dont dépendent les élevages. L'émergence de préoccupations quant à la durabilité des systèmes de production, leur aptitude à recycler les effluents et la nécessité de mettre en place des normes et des pratiques de fertilisation adaptées dans les zones occupées par les activités d'élevage appellent à la construction de références particulières, adaptées au contexte local.

Du fait des conditions climatiques contrastées et du gradient altitudinal marqué, les ressources herbagères produites sont de nature diverse et variable dans le temps, sur le plan quantitatif et qualitatif. La difficulté réside dans la gestion de la variabilité de cette ressource en fonction du climat et dans son adaptation permanente aux besoins des animaux.

Les plantes fourragères sont constituées d'eau, de carbone et de nombreux éléments minéraux : les macroéléments (N, P, K, Ca, Mg, S, Si) et les oligoéléments (Fe, Zn, Br, Mn, Cu, etc.). Les macroéléments et les oligoéléments proviennent de l'atmosphère, du sol, et des apports réalisés par l'agriculteur sous la forme d'engrais ou de matières organiques. C'est grâce à ses racines que la plante absorbe les éléments minéraux nutritifs dans le sol. En cas de déséquilibre nutritionnel, la plante subit une carence (manque d'un élément) ou une toxicité (excès d'un élément) qui se répercute sur la croissance et sur la production. Les espèces fourragères n'ont pas toutes les mêmes besoins en éléments nutritifs, d'où l'importance de choisir la culture en fonction des conditions de la parcelle (sol, climat).

L'Azote est l'élément essentiel de la croissance de la plante. La quantité d'Azote absorbée par la plante détermine le niveau d'absorption des autres éléments minéraux. Les besoins en éléments minéraux sont différents selon le stade de développement de la culture : la fertilisation (matières organiques complétée par les engrais minéraux) doit être apportée au bon moment et en quantité suffisante. Des apports trop élevés entraînent des risques d'accidents végétatifs et des pollutions de l'environnement. L'annexe B décrit plus en détail le cycle biogéochimique de l'Azote.

La matière organique du sol s'associe aux argiles et aux silicates d'alumine pour former le complexe argilo-humique. Ce complexe donne au sol ses caractéristiques de structure, de porosité, de stockage d'éléments minéraux nutritifs et de stockage de l'eau. La Capacité d'Échange Cationique (CEC) fait référence à la capacité du sol à retenir et échanger les éléments nutritifs facilement disponibles pour les plantes. Il s'agit d'une mesure de l'ensemble des charges négatives qui se trouvent sur l'argile et la matière organique.

Les apports d'éléments minéraux peuvent se faire soit sous forme organique (fumiers, lisiers, composts) qui sont issus des effluents naturels, végétaux ou animaux, soit sous forme minérale (Urée, Potasse, Phosphate, etc.), provenant de gisements ou produits par l'industrie chimique. Dans le cas des engrais organiques, l'Azote qu'ils contiennent, non utilisé directement par la plante, est en partie stocké dans le sol, contribuant ainsi à accroître le "capital de fertilité" de la prairie pour produire un "arrière effet" à plus long terme. Dans le cas de l'apport d'engrais minéral, très soluble, et compte tenu de l'absence de stockage dans le sol, l'Azote qui n'est pas utilisé directement par la plante risque d'être lessivé vers les aquifères (nappes phréatiques) ou perdu dans l'atmosphère (Gaz à Effet de Serre, GES).

---

L'utilisation raisonnée de fertilisants organiques issus du recyclage des effluents d'élevage représente une voie intéressante d'amélioration de la durabilité (économique, sociale et environnementale) des exploitations d'élevage. La valorisation des effluents d'élevage contribue en effet à la diminution de la dépendance du territoire vis-à-vis des engrais chimiques importés et à la réduction des impacts environnementaux négatifs des activités d'élevage. Si les réponses de l'herbe aux apports minéraux sont relativement bien connues et documentées dans la bibliographie, l'effet des pratiques d'épandage de matières organiques sur les prairies demande à être d'avantage documenté. Seul un ouvrage, publié en 2006 (P. Chabalier et al., [2]), résume les connaissances disponibles sur les qualités des matières organiques produites dans le contexte tropical de la Réunion, ainsi que les aspects pratiques et réglementaires de l'épandage.

Ces questionnements portent sur la mise en œuvre de nouvelles pratiques de gestion des effluents, de conduite de la fertilisation organique des prairies et sur les potentialités différenciées des herbages de l'île à recycler les nutriments issus des élevages. Ils ont conduit à la mise en place en 2004 d'un suivi de parcelles expérimentales pour évaluer les effets à long terme de différents types de fertilisation sur la production de biomasse fourragère, le stockage des nutriments dans le sol (Azote, Carbone) et la fertilité des sols, dans quatre sites expérimentaux représentatifs de la diversité agro-écologique des systèmes prairiaux à la Réunion.

L'examen de l'évolution de ces paramètres sur les huit années de l'essai montre la diversité des profils de croissance de l'herbe, l'alternance marquée entre saisons et l'impact des types et des niveaux de fertilisation organique et/ou minérale sur la production annuelle de biomasse. Dans le cadre de cette action de recherche, ma mission au sein de l'équipe Pôle Élevage, consiste alors à élaborer des modèles statistiques permettant de décrire l'évolution annuelle et pluriannuelle du rendement fourrager et de la composition des sols, et de quantifier les effets des facteurs de variation pris en compte dans l'essai (saison, site géographique, type et niveau de fertilisation).

Ce rapport se divise en trois parties. Le chapitre 1 présente le dispositif expérimental dont sont issues nos données et décrit les variables utilisées pour chaque analyse. Le second chapitre se rapporte à l'analyse mise en place pour étudier le rendement fourrager. Le chapitre 3 concerne la composition des sols : d'une part, en terme d'Azote, et d'autre part en terme de Capacité d'Échange Cationique (CEC). Nous terminons par une conclusion générale et par quelques pistes à explorer.



# Chapitre 1

## Dispositif expérimental et données disponibles

### 1.1 Le dispositif expérimental

Un suivi de parcelles fourragères expérimentales a été mis en place depuis 2004 pour évaluer les effets à long terme de différents types de fertilisation (engrais organique et/ou minéraux) et des niveaux d'incorporation (faible ou élevé). Les modalités du protocole ont été raisonnées en étroite collaboration avec les techniciens de l'ARP et des filières de production bovine (SICALAIT, SICAREVIA et SEDAEL).

Les quatre sites expérimentaux retenus sont représentatifs de la diversité agro-écologique des prairies à la Réunion (espèce fourragère dominante, altitude, climat, type de sol). Les situations géographiques des sites sont représentées sur la carte de la figure 1.1, et les caractéristiques associées à chaque site sont listées dans le tableau 1.1.



FIGURE 1.1 – Visualisation de l'emplacement géographique de chaque site sur la carte de l'île de la Réunion.

Lieu	Altitude (m)	Pluies (mm)	Sol	Espèces fourragères dominantes
LYCEE (St Joseph)	10	1000-1500	Sableux	Chloris
SEDAEL (Mont-Vert)	800	1000-1500	Andosol*	Kikuyu
SICALAIT (Plaines des Cafres)	1600	1500-2000	Andosol*	Ray-Gras, dactyle, brome
MARIANNE (Plaine des palmistes)	1200	4000-5000	Andosol*	Ray-Gras, dactyle, brome

TABLE 1.1 – Descriptif de la diversité agro-écologique des sites expérimentaux.

\*Andosol : sol d'origine volcanique riche en matière organique

Les différents types (minérales, organiques et mixtes) et niveaux de fertilisation utilisés dans cette étude sont présentés dans le tableau 1.2.

Code	Nom	Descriptif	Type
F1	Témoin	0 fumure	néant
F2	M70	70 kg engrais minéral	minéral
F3	L40	40 m <sup>3</sup> lisier	organique
F4	L40M30	40 m <sup>3</sup> lisier + 30 kg engrais minéral	mixte
F5	L70	70 m <sup>3</sup> lisier	organique
F6	C7.2	7,2 t compost	organique
F7	C7.2M56	7,2 t compost + 56 kg engrais minéral	mixte
F8	C12	12 t compost	organique
F9	C12M48	12 t compost + 48 kg engrais minéral	mixte
F10	M120	120 kg N engrais minéral	minéral

TABLE 1.2 – Descriptif des fertilisations minérales et/ou organiques utilisées.

L’engrais minéral est composé d’ammonitrate à 26% d’Azote. Le lisier, mélange de déjections d’animaux d’élevage (urines, excrément) et d’eau, est d’origine bovine. Le compost est un compost de fumier de bovin et de paille de canne à sucre. Le lisier et le fumier de bovin utilisés durant l’essai ont été collectés dans deux élevages, les mêmes durant toute la durée de l’essai. Le traitement de fertilisation ”témoin” (aucune fertilisation) est notre traitement de référence.

Dans chacun des quatre sites, les dix types de fertilisation ont été répétés indépendamment sur trois blocs, soient 30 micro-parcelles expérimentales de 4 m<sup>2</sup> chacune, par site. Ce plan expérimental en bloc complet (pour un site) est schématisé sur la figure 1.2. Autrement dit, toutes les combinaisons de toutes les modalités des facteurs sont examinées dans le dispositif.

Bloc	Parcelle	Fertilisation	Bloc	Parcelle	Fertilisation	Bloc	Parcelle	Fertilisation
B1	P1	6	B2	P11	9	B3	P21	3
	P2	2		P12	1		P22	2
	P3	8		P13	10		P23	4
	P4	3		P14	5		P24	7
	P5	1		P15	4		P25	9
	P6	9		P16	8		P26	6
	P7	4		P17	2		P27	8
	P8	10		P18	6		P28	10
	P9	5		P19	7		P29	1
	P10	7		P20	3		P30	5

Légende	
Fertilisation	Composition
1	0 fumure
2	70 kg engrais minéral
3	40 m <sup>3</sup> lisier
4	40 m <sup>3</sup> lisier + 30 kg engrais minéral
5	70 m <sup>3</sup> lisier
6	7,2 t compost
7	7,2 t compost + 56 kg engrais minéral
8	12 t compost
9	12 t compost + 48 kg engrais minéral
10	120 kg N engrais minéral

FIGURE 1.2 – Représentation schématique du plan expérimental en bloc complet pour un site.

Trois principes ont été appliqués lors de la planification expérimentale afin d'assurer la validité de l'analyse :

1. *la randomisation* : l'allocation des traitements de fertilisation aux micro-parcelles d'un bloc respecte un tirage aléatoire équiprobable ;
2. *les répétitions* : chaque traitement de fertilisation est affecté à trois micro-parcelles pour un même site afin de pouvoir estimer l'erreur expérimentale résiduelle ;
3. *le contrôle de l'erreur* : la part non contrôlée de l'expérience a été réduite afin de diminuer l'erreur expérimentale.

Les parcelles d'un même site sont fauchées cinq à huit fois par an, selon un rythme de coupe qui dépend des conditions climatiques, qui influent directement sur le rythme de pousse de l'herbe. Les coupes sont ainsi plus espacées au cours de la saison sèche et fraîche (juin à novembre). Les actions menées après chaque coupe sur chacune des micro-parcelles sont les suivantes :

- Récolte et pesée des quantités de fourrage fauchées (Matière Verte MV) ;
- Prélèvement d'un échantillon de fourrage ;
- Fertilisation de la parcelle (type et quantité), selon les modalités précisées (tableau 1.2) ;
- Séchage des échantillons de fourrage et calcul du taux de Matière Sèche (MS) ;
- Analyse de la composition des échantillons de fourrage.

Les prélèvements d'échantillon de sol sont effectués une fois par an, en novembre, sur lesquels deux actions sont menées : séchage/tamissage, et analyse de la composition.

## 1.2 Les données

Une base de données contenant l'ensemble des données collectées sur les micro-parcelles a été créée sous ACCESS : biomasse fourragère (rendement), composition chimique des fourrages et des sols. Trois analyses ont été planifiées dans le cadre de cette étude :

1. l'analyse de l'effet de la fertilisation sur le rendement fourrager ;
2. l'analyse de l'effet de la fertilisation sur la teneur en Azote des sols ;
3. l'analyse de l'effet de la fertilisation sur la Capacité d'Échange Cationique (CEC) des sols.

### 1.2.1 Effet de la fertilisation sur le rendement fourrager

Dans cette première analyse, les variables explicatives sont (1) le traitement de fertilisation, (2) le site et (3) la saison.

Le traitement de fertilisation comporte 10 modalités (tableau 1.2) et le site, 4 modalités (tableau 1.1). Chacune des modalités de la variable *site* représente un contexte pédo-climatique particulier, qui associe une altitude, un climat, un type de sol et une espèce fourragère dominants. L'analyse ne permettra donc pas de distinguer les effets de la fertilisation sur chacun de ces aspects.

L'année biologique a été définie selon l'évolution climatique annuelle observée à la Réunion ; elle s'étale du 01/12 d'une année  $n$  au 30/11 de l'année  $n + 1$ . Les années 2004 et 2005 (11 premiers mois) ont été écartées du champ de l'analyse car elles correspondaient à une période de calage du protocole.

Les cinq à huit coupes par an sont déclenchées en fonction du stade de maturité et de la hauteur d'herbe. Le climat étant variable, la pousse de l'herbe varie selon la saison et l'année, ce qui engendre un rythme de coupe plus ou moins élevé selon les saisons et les années. Avant d'étudier le rendement à l'échelle saisonnière, il a été nécessaire de vérifier si l'intervalle entre coupes est similaire d'un site à un autre, autrement dit si ces délais sont distribués de façon égale pour tous les sites. D'après les résultats de cette analyse préliminaire (annexe C), nous avons déduit une indépendance entre le site et l'intervalle entre coupes selon la saison, et pour toutes les années. Nous pouvons ainsi étudier les rendements par coupe selon le site et la saison, indépendamment de cet intervalle.

Au sein d'une même année biologique, les coupes ont donc été rassemblées en quatre saisons trimestrielles :

- la saison des pluies, notée  $T1$ , qui s'étale du 01/12 au 28/02 de l'année suivante,
- la fin de saison des pluies, notée  $T2$ , qui s'étale du 01/03 au 31/05,
- la saison fraîche, notée  $T3$ , qui s'étale du 01/06 au 31/08,
- la fin de saison fraîche, notée  $T4$ , qui s'étale du 01/09 au 30/11.

La variable à expliquer est la quantité moyenne de biomasse fauchée par coupe, par micro-parcelle, et par saison exprimée en kilogramme de Matière Sèche (MS) par hectare ( $kg\ MS/ha$  /coupe). La quantité de MS par coupe est obtenue en effectuant le produit de la quantité de fourrage vert fauchée ( $g\ MV / 4\ m^2$ ) et du taux de Matière Sèche du fourrage (%), et en ramenant le résultat à l'unité de surface (1 hectare). Nous calculons ensuite, pour chacune des micro-parcelles, la moyenne des quantités de MS par coupe pour les coupes observées au sein de chacune des saisons. Nous notons notre variable à expliquer  $Y$ , qui est donc le rendement moyen des coupes récoltées lors d'une saison et d'une année, pour une micro-parcelle.

### 1.2.2 Effet de la fertilisation sur la teneur en Azote des sols

Concernant la seconde analyse, nous retrouvons (1) le traitement de fertilisation et (2) le site qui seront nos facteurs explicatifs, auxquels nous ajoutons une variable supplémentaire, à savoir (3) l'année.

La variable *annee* a été recodée en variable numérique, allant de 0 à 6, correspondant aux 7 années biologiques comprises entre le 01/12/2005 au 30/11/2012. L'année 0 représentant l'année de référence (2006).

La teneur en Azote du sol est la variable à expliquer de cette seconde étude, notée  $Ntot$ , exprimée en gramme d'Azote par kilogramme d'échantillon de sol sec ( $g\ N/kg\ sol$ ). Les observations de la variable  $Ntot$  sont disponibles pour chaque micro-parcelle de chaque site et pour chaque année, soient 120 séries temporelles, chacune comportant 7 observations. Ainsi, le nombre total d'observations s'élève à 840 pour les quatre sites étudiés. Toutes les mesures ont été effectuées aux mêmes périodes, et le jeu de données est équilibré.

### 1.2.3 Effet de la fertilisation sur la CEC

Dans le cadre de la troisième analyse, nous conservons les variables explicatives de la teneur en Azote, c'est à dire (1) le traitement de fertilisation, (2) le site et (3) l'année, avec en plus (4) la biomasse fourragère annuelle exportée.

La dernière variable *biomasse* est la quantité de biomasse fourragère produite dans l'année biologique (01/12 au 30/11 de l'année suivante). Cette biomasse s'obtient par cumul des rendements obtenus par coupe, par site et par fertilisation sur une année biologique, exprimée en kilogramme de Matière Sèche (MS) par hectare.

La variable à expliquer dans cette analyse est la Capacité d'Échange Cationique (CEC), qui est le second critère de qualité du sol. Cette quantité fait référence à la capacité du sol à retenir et échanger des éléments nutritifs facilement disponibles pour les plantes. Il s'agit d'une mesure de l'ensemble des charges négatives qui se trouvent sur l'argile et la matière organique. La CEC s'exprime en milligrammes d'équivalents par 100 grammes d'échantillon de sol sec ( $meq/100\ g\ sol$ ). Tout comme la teneur en Azote, nous disposons de 120 séries temporelles de 7 observations, correspondants aux prélèvements de sol effectués de 2006 à 2012, soient 840 relevés.

Ainsi, l'étude se décompose en deux grandes parties. La première concerne l'analyse de la quantité de biomasse récoltée à l'échelle de la saison. La seconde se rapporte à la teneur du sol en Azote et de la CEC du sol.

# Chapitre 2

## Structure de variance-covariance du rendement fourrager

### 2.1 Problématique et description de l'étude

*Dans quelle mesure pouvons-nous évaluer l'impact du type et niveau de fertilisation sur la quantité de biomasse produite à l'échelle des saisons ou de l'année ?*

L'objectif principal de cette analyse est de savoir comment le rendement fourrager moyen varie selon la fertilisation utilisée, et si les différences de rendement constatées évoluent au cours des saisons. Plus précisément, nous étudions les liens qui peuvent exister entre la quantité de biomasse produite, les modalités de fertilisation, les caractéristiques pédo-climatiques par saison. Les objectifs restent donc semblables à ceux d'une analyse factorielle classique. Cependant, l'hypothèse d'indépendance des mesures de rendement successives effectuées sur une même micro-parcelle n'est pas valide. Nous sommes alors en présence de répétitions dans les données.

Ce qui distingue l'analyse des mesures répétées de l'ANOVA est la structure de covariance des données observées. Nous pouvons en effet modéliser la corrélation (si elle existe) entre les observations faites sur une même unité au cours du temps. Introduire une "structure de covariance" pour ces observations revient à considérer une matrice de variance-covariance, notée  $R$ , de dimension  $T \times T$ , où  $T$  est le nombre d'instantants d'observation.

Dans le cadre de l'analyse du rendement fourrager, la modélisation des données répétées nécessite l'estimation : des effets fixes, des composantes de la variance et des éléments de la matrice  $R$ . Nous envisageons différentes structures particulières pour la matrice  $R$ , permettant de réduire le nombre de paramètres à estimer. L'estimation se fera alors par l'approche des Moindres Carrés Généralisés (GLS) qui étend la régression linéaire en modélisant l'hétérogénéité et/ou la corrélation des données (J. Pinheiro and D. Bates, [10], A. Zuur et al., [16]). En parallèle, nous testons la significativité des différents effets retenus, pour ainsi chercher à simplifier le plus possible le modèle final. Enfin, nous utilisons les critères d'Akaike et de Schwarz (AIC et BIC) pour comparer les différents modèles obtenus. Rappelons que ces deux derniers sont obtenus de la façon suivante :

$$AIC(M) = -2.L(\hat{\theta}, y) + 2.n_p$$

$$BIC(M) = -2.L(\hat{\theta}, y) + n_p.log(n)$$

où  $n$  est le nombre d'observations dans l'étude,  $n_p$  est le nombre de paramètres à estimer dans le modèle et  $L(\hat{\theta}, y)$  est sa vraisemblance. L'AIC permet de mesurer le compromis entre la précision du modèle et sa complexité. Nous nous appuyons aussi sur le test de rapport de vraisemblance lorsque les modèles sont emboîtés.

Dans un premier temps, nous présentons trois structures usuelles permettant de modéliser les corrélations entre les observations. Une approche descriptive de nos données fait suite à cette partie théorique, conduisant à une analyse du rendement fourrager ciblée pour un site (SICALAIT), que nous élargissons à l'échelle des quatre sites dans un second temps.

## 2.2 Les structures classiques de covariance

Différentes structures sont disponibles sous le logiciel R (version 3.0.0, [12]) pour modéliser la matrice de covariance  $R$ , associée à l'unité expérimentale. Nous comparons ici les plus utilisées, mais il existe un grand nombre de possibilités (M. Davidian et D. Giltinan, [3], P. Diggle et al., [4]). Nous conservons les mêmes effets fixes d'un modèle à un autre, ce qui nous autorise à utiliser les critères d'AIC et de BIC lors du choix cette structure de covariance.

### 2.2.1 Absence de structure

La modélisation de  $R$  la plus générale, mais aussi la plus coûteuse en terme de temps de calcul, est l'absence de structure. Cette forme suppose que chaque variance et chaque covariance est unique. La matrice  $R$  est symétrique et définie positive, et peut s'écrire de la façon suivante :

$$R = \begin{pmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}. \quad (2.1)$$

Tous les éléments de  $R$  doivent être estimés, soit  $\frac{T(T+1)}{2} = 10$  paramètres. Du fait du nombre important de termes, nous cherchons à contourner ce type de modélisation, tout en évitant la structure d'indépendance peu réaliste dans ce contexte.

### 2.2.2 Structure en symétrie composée

La structure en symétrie composée est fréquemment utilisée, afin de réduire le nombre de paramètres à estimer, tout en facilitant l'estimation. Cette forme suppose que la variance de la v.a.r (ici le rendement) associée à toute observation vaut  $\sigma_1^2 + \sigma_2^2$ . La covariance des v.a.r. associées à deux observations réalisées sur la même unité à deux temps différents vaut constamment  $\sigma_2^2$ . L'avantage de cette structure est d'estimer seulement deux paramètres. Dans ce cas, la matrice  $R$  prend la forme :

$$R = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_2^2 & \sigma_2^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \sigma_2^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \sigma_2^2 \\ \sigma_2^2 & \sigma_2^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix}. \quad (2.2)$$

Ainsi, la covariance entre deux observations est telle que :  $cov(Y_t, Y_{t'}) = \sigma_2^2$  si  $t \neq t'$ , et la variance (homogène) d'une observation est :  $\mathbb{V}(Y_t) = \sigma_1^2 + \sigma_2^2$ . Avec ce type de structure, le modèle se ramène à un modèle mixte. La composante de l'erreur, notée  $U$ , peut alors se décomposer en deux parties :

- un effet aléatoire  $A$  rattaché aux unités, dont les composante suivent une loi  $\mathcal{N}(0, \sigma_2^2)$ ,
- l'erreur résiduelle  $U^*$ , dont les termes proviennent d'une loi  $\mathcal{N}(0, \sigma_1^2)$

En notant  $\rho$  la corrélation entre deux mesures, une autre façon de représenter cette matrice est la suivante :

$$R = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}, \quad (2.3)$$

où  $\rho = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$  et  $\sigma^2 = \sigma_1^2 + \sigma_2^2$ .

### 2.2.3 Structure auto-régressive d'ordre 1

L'intérêt de la structure auto-régressive d'ordre 1 ou  $AR(1)$  est aussi de réduire l'estimation à seulement deux composantes, mais cette fois en tenant compte des écarts entre les mesures. En effet, elle est telle que la corrélation des v.a.r. associées à deux mesures réalisées sur la même

micro-parcelle à deux instants différents est inversement proportionnelle à l'écart entre ces deux instants. La structure de la matrice  $R$  s'exprime de la façon suivante :

$$R = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}_{4 \times 4}. \quad (2.4)$$

avec  $|\rho| \leq 1$ . Ainsi, cette structure  $AR(1)$  spécifie une variance homogène  $\mathbb{V}(Y_t) = \sigma^2$ . La covariance entre deux observations faites sur la même unité décroît lorsque le retard ( $|t - t'|$ ) augmente :  $cov(Y_t, Y_{t'}) = \sigma^2 \rho^{|t-t'|}$ . Ainsi, la corrélation entre deux mesures aux temps  $t$  et  $t'$  est donnée par :

$$cor(Y_t, Y_{t'}) = \begin{cases} 1 & \text{si } t = t' \\ \rho^{|t-t'|} & \text{sinon.} \end{cases}$$

Notons que ces différentes structures supposent l'homogénéité de la variance. Ainsi, la variance des observations est supposée être la même quelle que soit la saison. De manière similaire, les variances sont alors égales pour tous les niveaux des facteurs étudiés, puisque  $R$  ne dépend d'aucun indice. Ainsi, la matrice de variance-covariance totale, notée  $\Sigma$  est diagonale par bloc. Tous les blocs sont égaux à  $R$  et chaque bloc correspond à une unité expérimentale. Lorsque l'hypothèse d'homogénéité n'est pas vérifiée, nous pouvons spécifier une variance différente selon les niveaux du facteur. De cette manière, nous obtenons une structure encore plus générale, mais qui implique l'estimation de paramètres supplémentaires.

## 2.3 Description des données

Notre variable à expliquer, notée  $Y$ , représente le rendement (ou la biomasse récoltée) exprimé en kilogramme de matière sèche par hectare et par coupe (kg MS/ha/coupe). L'analyse exploratoire de la variable  $Y$  située en annexe (figures D.3 et D.4) nous conduit à transformer la variable  $Y$  en utilisant son logarithme, afin de la normaliser et de chercher à homogénéiser la variance.

L'effet fixe associé à la variable *annee* n'a pas d'influence très significative sur le rendement pour deux sites : le test de nullité de Student ne rejette pas l'hypothèse d'un coefficient nul pour les sites de SICALAIT et MARIANNE (avec des p-valeurs de 0.579 et de 0.267). De plus, l'évolution moyenne du rendement semble être la même quelle que soit la fertilisation utilisée, comme le montre la figure 2.1, qui représente les interactions entre l'année et la fertilisation.

D'autre part le graphique des corrélations entre les différentes années de l'essai (annexe D.2) indique une faible dépendance des observations du rendement. Du point de vue agronomique, ce phénomène se justifie par le fait que la production de biomasse dépend beaucoup du climat, qui est très variable d'une année à l'autre (année sèche, année pluvieuse, etc.). C'est pourquoi nous considérons dans cette analyse que les rendements dépendent pour une large part du climat et sont alors indépendants d'une année à l'autre. La variation étudiée ici est intra-annuelle.

L'unité expérimentale est ici la micro-parcelle observée sur une année, référencée comme "sujet". Ainsi, nous obtenons au total 840 unités, et 210 par site (30 micro-parcelles  $\times$  7 années). Pour chaque unité, nous avons quatre observations de la variable dépendante (le rendement moyen par coupe), associées aux récoltes saisonnières. Le facteur "fertilisation", noté  $F$ , est désigné comme un facteur inter-sujet, tandis que la saison, notée  $T$ , est considérée comme un facteur intra-sujet. Les temps d'observation sont ici les mêmes pour toutes les unités expérimentales : les quatre saisons.

Un tel dispositif est appelé "plan factoriel complet" ou "plan en bloc complet", car il permet de croiser tous les niveaux des facteurs, et il est "équilibré" puisque chaque fertilisation a été reçue par le même nombre d'unités.

Dans un premier temps, nous nous intéressons au rendement saisonnier à l'échelle d'un site, celui de SICALAIT.

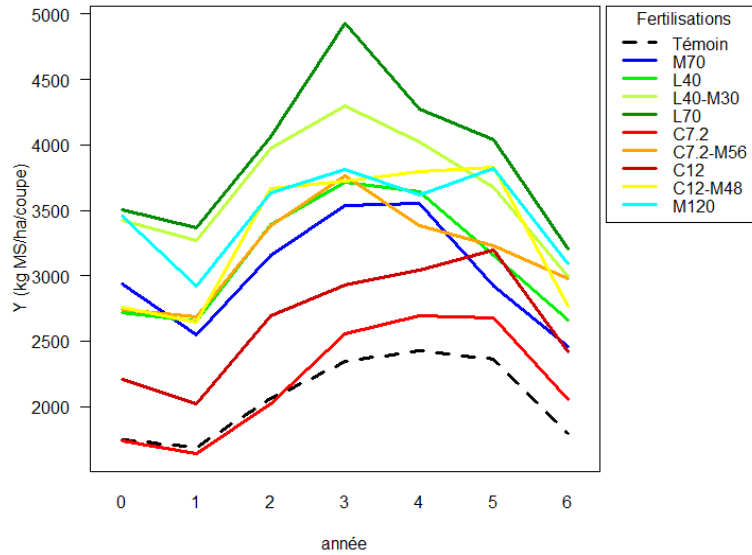


FIGURE 2.1 – Graphique des interactions entre l’année et la fertilisation sur le rendement  $Y$  (en kg MS/ha/coupe).

## 2.4 Analyse du rendement fourrager sur le site de SICALAIT

Dans cette partie, nous considérons un seul facteur à effets fixes autre que la saison : la fertilisation.

### 2.4.1 Analyse préliminaire : ANOVA à deux facteurs

Tout d’abord, nous effectuons une analyse ANOVA à deux facteurs croisés. Cette première approche nous permet d’apporter un complément à la modélisation des données répétées, et d’avoir une idée de l’influence de la fertilisation  $F$  sur le rendement  $Y$ . Cette approche est bien entendu biaisée, puisqu’elle suppose que les observations réalisées sur une même unité au cours du temps sont indépendantes.

Les trois facteurs testés sont : la fertilisation  $F$  (à 10 niveaux, indicés par  $j$ ), la saison  $T$  (à quatre niveaux, indicés par  $t$ ) et l’interaction  $F \times T$  (indiquée par  $jt$ ). Le nombre d’unités étudiées ici est réduit à  $n = 210$ . Pour chaque couple  $(j, t)$  de niveaux, nous avons  $n_{jt} = 21$  observations de la variable dépendante  $Y$ . De ce fait, le modèle de l’analyse de la variance considéré est le suivant :

$$\log(Y_{ijt}) = \mu + \alpha_j + \gamma_t + (\alpha\gamma)_{jt} + \epsilon_{ijt}, \quad (2.5)$$

où  $\mu$  est l’effet général,  $\alpha_j$  est l’effet principal du niveau  $j$  du facteur fertilisation  $F$ ,  $\gamma_t$  est l’effet principal du niveau  $t$  du facteur saison  $T$ , et  $(\alpha\gamma)_{jt}$  est l’effet de l’interaction entre le niveau  $j$  et la saison  $t$ .

Ce modèle suppose que  $\epsilon_{ijt} \sim \mathcal{N}(0, \sigma^2)$ . De ce fait, la variance d’une observation est constamment  $\sigma^2$  et la covariance entre deux observations est nulle :  $\text{cov}(\log(Y_{ijt}), \log(Y_{ijt'})) = 0$  si  $t \neq t'$  et  $\text{V}(\log(Y_{ijt})) = \sigma^2$ . La matrice de variance-covariance résiduelle est diagonale, et possède alors une “structure d’indépendance” telle que :

$$\hat{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2 \end{pmatrix}_{210 \times 210}. \quad (2.6)$$



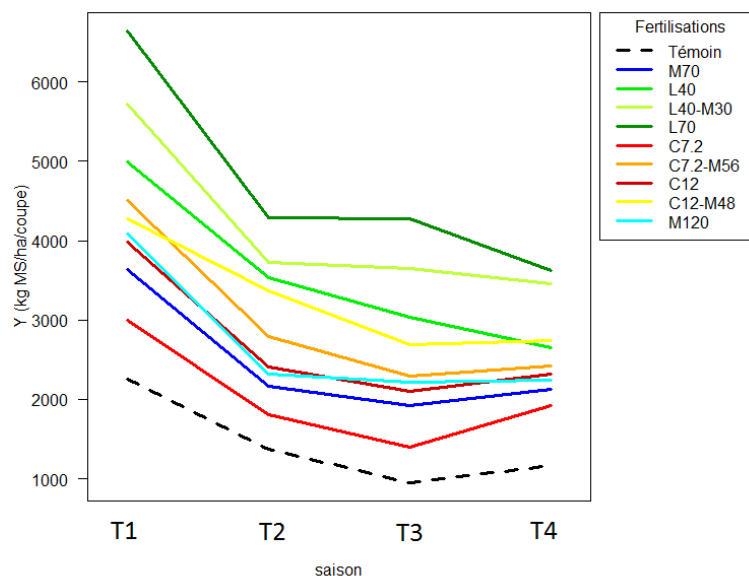
Ainsi, le modèle factoriel ne tient pas compte de la corrélation entre les différentes mesures faites pour une même unité (soient quatre répétitions par unité). En effet, avec ce type de modèle, les quatre mesures trimestrielles sur une même unité sont considérées comme indépendantes. Les facteurs étudiés sont alors testés à partir des moyennes des mesures réalisées pour chaque niveau. Nous obtenons les résultats suivants avec l'ANOVA (Table 2.1) :

	DF	SSE	MSE	F	p-valeur
fertilisation $F$	9	103.3	11.5	73.3	$< 2.2 \times 10^{-16}$
saison $T$	3	51.8	17.3	110.3	$< 2.2 \times 10^{-16}$
interaction $F \times T$	27	4.1	0.2	0.9	0.52
Résidus	800	125.3	0.2		

TABLE 2.1 – Résultats des tests de l'ANOVA pour le modèle 2.5.

Note : la p-valeur est associée à la statistique  $F$  connaissant le nombre de paramètres  $DF$  à estimer.  $SSE$  et  $MSE$  correspondent à la somme et à la moyenne quadratique résiduelle.

D'après le test, il existe au moins une fertilisation et une saison qui influencent le rendement fourrager. L'interaction entre la fertilisation et la saison n'est clairement pas significative (au seuil 5%). Le graphique 2.2 permet de visualiser l'effet de cette interaction entre la saison et la fertilisation sur le rendement.

FIGURE 2.2 – Graphique des interactions entre la fertilisation et la saison sur le rendement  $Y$  (en kg MS/ha/coupe).

Le résultat du test de l'ANOVA est donc confirmé, puisque les droites ne se coupent quasiment pas. Dans la suite, nous n'incluons pas cet effet croisé lors de la modélisation.

Malgré l'éventuelle présence de corrélations dans les données, les mesures de la qualité de la régression par le  $R^2$  et par le  $R_{aj}^2$  restent justes. Rappelons que ces outils statistiques se calculent de la façon suivante :

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2},$$

$$R_{aj}^2 = 1 - \frac{1}{(n-p)} \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2},$$

où  $n$  est le nombre d'observations,  $p$  est le nombre de paramètres à estimer. Nous obtenons ici  $R_{aj}^2 = 0.538$  et l'erreur résiduelle estimée par  $\hat{\sigma} = 0.396$ .

## 2.4.2 Détection des anomalies

### Hétérogénéité saisonnière

Jusqu'à présent, nous avons supposé que la variance intra-saisonnière était homogène, c'est à dire que la variabilité des mesures est la même pour chaque saison :

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2.$$

Malgré la transformation de la variable dépendante, cette variance intra-saisonnière n'est pas stabilisée. Le graphique de dispersion (boxplot) D.7 de l'analyse exploratoire située en annexe montre bien une dispersion différente selon la saison. Le test de Levene, utilisant la modification de Brown-Forsythe, rejette l'hypothèse d'égalité des variances saisonnières, avec une p-valeur de  $3.63 \times 10^{-4}$  (elle était de  $4.38 \times 10^{-8}$  sur les données brutes). À un seuil de significativité de 5%, nous avons donc des preuves que les variances intra-saisonnières ne sont pas égales. Le test d'homogénéité de Bartlett confirme aussi cette différence de variabilité, avec une p-valeur de  $2.29 \times 10^{-4}$ , initialement de  $9.93 \times 10^{-14}$  sans la transformation logarithmique.

Ainsi, l'utilisation d'une analyse de la variance classique est à proscrire, puisque l'hypothèse d'homogénéité de la variance n'est pas vérifiée.

### Corrélation résiduelle

En présence d'autocorrélation dans les résidus, les estimateurs des moindres carrés ordinaires ne sont plus de variance minimale. Les tests d'hypothèses usuels ne sont alors plus valides.

Dans cette expérience, les mesures sont effectuées à des intervalles de temps égaux. Afin de détecter une éventuelle corrélation intra-sujet, il est possible d'utiliser la fonction d'autocorrélation. La figure 2.3 représente la fonction d'autocorrélation empirique associée aux résidus standardisés issus du modèle linéaire, sans aucune structure de corrélation sur les répétitions.

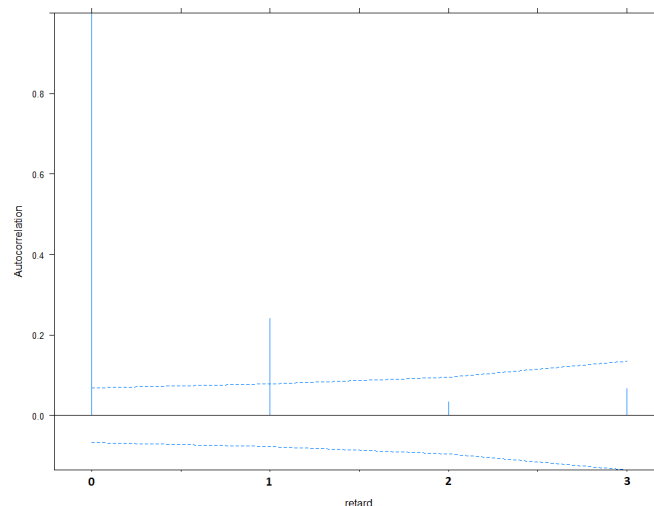


FIGURE 2.3 – Fonction d'autocorrélation empirique des résidus standardisés du modèle (2.5).

D'après l'examen visuel de la figure 2.3, nous constatons une présence d'autocorrélation significative parmi les résidus empiriques. Sur le graphique 2.4, nous représentons la matrice des corrélations estimées sur les saisons, ainsi que les nuages de points associés.

Les mesures du rendement lors de la saison 2 et de la saison 3 sont très corrélées, tout comme pour la saison 3 et la saison 4. Cette corrélation décroît lorsque l'on compare des saisons plus

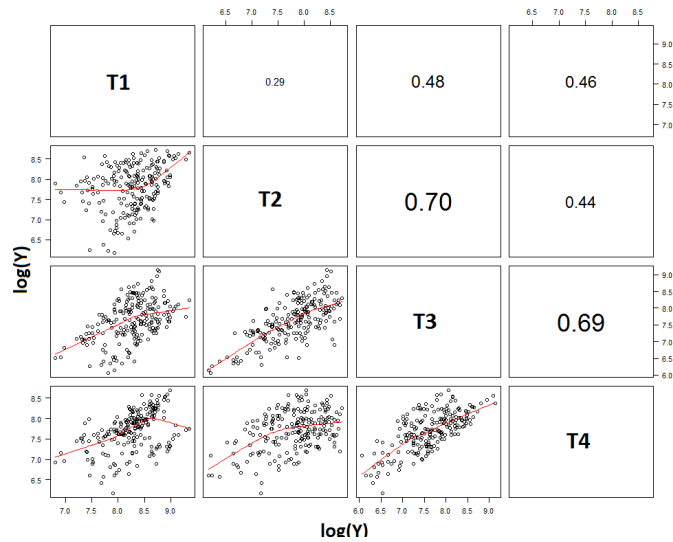


FIGURE 2.4 – Représentation de la matrice des corrélations estimées sur le logarithme du rendement  $Y$  pour chaque couple de saisons.

éloignées. La première saison semble tout de même avoir un comportement particulier. Sur les graphiques de dispersion du rendement pour chaque paire de saisons, nous pouvons aussi constater une diminution de la corrélation lorsque l'intervalle de temps entre les mesures augmente. En effet, les points sont plus regroupés pour les saisons 2 et 3, ainsi que pour les saisons 3 et 4, ce qui n'est pas le cas pour des observations plus distantes (saisons 2 et 4).

Par conséquent, l'analyse de la variance standard n'est probablement pas adaptée à ces données. Ainsi, un autre type d'analyse doit être utilisé. Dans la suite, nous cherchons à tenir compte de la structure de corrélation entre les mesures successives sur une même unité expérimentale.

### 2.4.3 Le modèle à effets fixes pour données répétées

L'objectif ici est d'utiliser directement une structure de corrélation pour modéliser l'association entre les erreurs intra-parcelles. Les quatre répétitions effectuées sur une même micro-parcelle seront alors dépendantes. Les effets fixes du modèle sont ceux de la saison et de la fertilisation. Les effets d'interactions entre la fertilisation et la saison n'étant pas significatifs, nous ne les incluons pas dans le modèle.

Le principe général est le suivant : pour un niveau  $j$  du facteur fertilisation  $F$  ( $j = 1, \dots, 10$ ), et pour une observation  $i$  prise en compte à ce niveau là ( $i = 1, \dots, n_j = 1, \dots, 21$ ), on considère le vecteur aléatoire :  $\log(Y_{ij}) = (\log(Y_{ij1}), \dots, \log(Y_{ijT}))'$  de  $\mathbb{R}^T$ , avec ici  $T = 4$ , et on pose :

$$\log(Y_{ij}) \sim \mathcal{N}_T(\mu_j, R),$$

où  $\mu_j = (\mu_{j1}, \dots, \mu_{jT})'$ . Par ailleurs, on définit :

$$\mu_{jt} = \mu + \alpha_j + \gamma_t, \quad (2.7)$$

où  $\mu$  est l'effet général,  $\alpha_j$  est l'effet principal du niveau  $j$  du facteur fertilisation  $F$  et  $\gamma_t$  est l'effet principal de la saison  $t$ .

On considère la matrice  $R$  qui est  $T \times T$ , symétrique et strictement définie positive ; elle représente la structure de covariance des observations répétées sur une même unité relative à la fertilisation  $j$ .

Le modèle considéré ici s'écrit donc :

$$\log(Y_{ijt}) = \mu_{jt} + U_{ijt}, \quad (2.8)$$

où l'erreur  $U_{ijt}$  est la  $t$ -ième composante du vecteur aléatoire gaussien centré  $U_{ij}$  de  $\mathbb{R}^T$ , de matrice de variance-covariance  $R$ . Les composantes de ce vecteur aléatoire ne sont donc pas indépendantes ( $R$  n'est en effet pas diagonale).

Dans ce modèle multidimensionnel complet, nous distinguons :

- Un effet général  $\mu$  ;
- 10 effets principaux du facteur fertilisation  $F$  (les  $\alpha_j$ ) avec une contrainte de centrage ;
- 4 effets principaux du facteur saison  $T$  (les  $\gamma_t$ ) aussi avec une contrainte de centrage.

Les deux premiers types d'effets sont des effets inter-sujets (indépendants du temps) et le dernier est un effet intra-sujets (dépendant du temps). En comptant les éléments de la matrice  $R$ , cela fait au total  $(J + T - 1) + \frac{T(T+1)}{2} = 23$  paramètres indépendants à estimer pour le modèle homoscédastique, avec  $n * T = 210 * 4 = 840$  observations ( $n = \sum_{j=1}^J n_j$ ). Il est recommandé de disposer d'un nombre total de sujets  $n$  vérifiant :  $n \geq J + \frac{T+1}{2}$ . Dans cet essai, nous disposons bien de  $n = 210 \geq 12.5$ .

#### 2.4.4 Choix de la structure de variance-covariance

Les différents modèles construits ont été estimés à l'aide de la fonction  $gls()$  du package *nlme* (J. Pinheiro, D. Bates, S. DebRoy et D. Sarkar, [11]) du logiciel R (version 3.0.0, [12]). Le tableau 2.2 résume les résultats obtenus avec les trois structures de variance-covariance mises en oeuvre sur les données.

	Symétrie composée			Autoregressive 1			Non-structurée		
	AIC	BIC	log-V	AIC	BIC	log-V	AIC	BIC	log-V
Homogénéité de la variance intra-saison	880.6	951.4	-425.3	869.9	940.7	-419.9	815.2	909.5	-387.6
Hétérogénéité de la variance intra-saison	879.1	964.0	-421.5	<b>867.6</b>	952.5	-415.8	809.5	918.0	-381.8

TABLE 2.2 – Comparaison des différentes structures de variance-covariance.

Note : la log-V correspond à la log-vraisemblance du modèle.

D'après le critère de l'AIC, une structure de corrélation autorégressive d'ordre 1 (AR(1)) semble la plus appropriée aux données (Table 2.2). Cependant, la p-valeur du test de Levene pour l'égalité des variances résiduelles par saison est de  $6.15 \times 10^{-4}$  ( $3.48 \times 10^{-4}$  pour le test de Bartlett). Au seuil 5%, nous rejetons donc l'hypothèse d'égalité des variances des résidus selon les saisons. Il est alors nécessaire de tenir compte de cette variabilité saisonnières. Pour cela, nous spécifions une variance différente selon la saison, c'est à dire :

$$\sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2 \neq \sigma_4^2.$$

Ainsi, trois paramètres supplémentaires sont à estimer. La structure de la matrice  $R$  s'écrit alors :

$$R = \begin{pmatrix} \sigma_1^2 & \sigma_2\sigma_1\rho & \sigma_3\sigma_1\rho^2 & \sigma_4\sigma_1\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_3\sigma_2\rho & \sigma_4\sigma_2\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_4\sigma_3\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}_{4 \times 4}. \quad (2.9)$$

Afin de vérifier si la structure en AR(1) modélise correctement la dépendance intra-sujet, nous représentons le fonction d'autocorrélation empirique pour les résidus normalisés du modèle. Le graphique 2.5 nous indique que la corrélation intra-sujet semble totalement expliquée par le modèle (valeurs significatives), et donc que la structure en AR(1) représente de manière adéquat la dépendance entre les mesures.

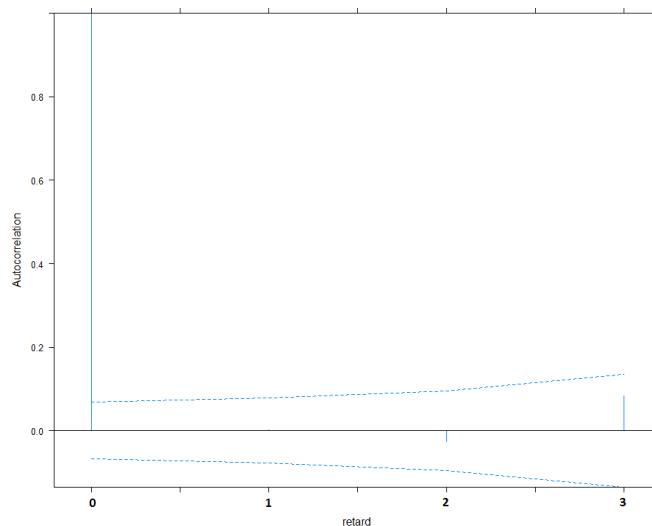


FIGURE 2.5 – Fonction d'autocorrélation empirique des résidus normalisés du modèle 2.8 comportant la structure de variance-covariance 2.9

Nous effectuons ensuite un test de Levene sur les résidus selon les saisons afin de confirmer la prise en compte de l'hétérogénéité. Cette fois, l'hypothèse nulle n'est pas rejetée au seuil 5% (la p-valeur est de 0.1302). De même, le test de Bartlett ne rejette pas le postulat d'homogénéité des résidus selon la saison.

Comme le confirme le test du rapport de vraisemblance (Table 2.3), ce modèle est significativement meilleur (au seuil 5%) que le modèle supposant seulement l'homogénéité des variances. Il est aussi préféré à celui ne tenant pas compte de la structure de corrélation.

Modèle	df	AIC	BIC	logLik
1 : hétérogénéité	17	895.69	975.90	-430.84
2 : AR(1)	15	869.88	940.65	-419.94
3 : hétérogénéité + AR(1)	18	867.60	952.52	-415.80

Test	L.Ratio	p-valeur
1 vs 3	30.09	< .0001
2 vs 3	8.28	0.04

TABLE 2.3 – Comparaison des modèles retenus.

### 2.4.5 Résultats

Le modèle ainsi choisi pour le rendement fourrager sur le site de SICALAIT comporte les effets principaux des saisons et des fertilisations. L'interaction entre ces deux effets fixes n'est pas significative (à 5%) : l'effet de la fertilisation sur la biomasse ne varie donc pas selon la saison. La structure de variance-covariance retenue pour le modèle à effets fixes est de type autorégressive d'ordre 1 et tient compte de l'hétérogénéité saisonnière. L'intervalle de confiance obtenu pour l'estimateur du paramètre de corrélation  $\rho$  est :  $[0.145; 0.301]$ , et  $\hat{\rho} = 0.225$ . Le modèle implique aussi l'estimation des quatre composantes de la variance associées aux saisons. Ces variabilités sont estimées en considérant la première saison comme étant la saison de référence. Plus précisément, les variances des trois autres saisons sont calculées en multipliant la variance obtenue lors de la saison  $T1$  par un facteur de variation (ou pondération) estimé à partir des données. Les estimations de ces dispersions résiduelles sont résumées dans le tableau 2.4.

Nous constatons que la saison  $T2$  présente un ratio légèrement plus élevé, indiquant une variabilité plus importante lors de cette période de l'année.

<i>saison<sub>j</sub></i>	saison 1 (ref.)	saison 2	saison 3	saison 4
$\hat{\sigma}_j$	0.377	0.432	0.411	0.361
$\hat{\sigma}_j/\hat{\sigma}_1$	1	1.147	1.09	0.957

TABLE 2.4 – Estimations des écart-types et des facteurs multiplicatifs associés à chaque saison.

Les trois autres sites ont été étudiés en suivant la même approche et la même structure de variance-covariance a été retenue.

## 2.5 Analyse du rendement fourrager sur les sites représentatifs à la Réunion

Dans cette partie, nous élargissons l'étude à l'échelle des quatre sites. Pour cela nous utilisons l'ensemble des données disponibles en ajoutant au modèle un facteur "site". Comme précédemment, nous commençons par une analyse de la variance classique, mais cette fois avec un facteur supplémentaire.

### 2.5.1 Analyse préliminaire : ANOVA à trois facteurs

Au préalable, nous étudions les effets de trois facteurs principaux (la fertilisation  $F$ , le site  $S$  et la saison  $T$ ), et les effets de deux interactions (fertilisation/site et saison/site) sur le rendement  $Y$ . L'interaction entre la fertilisation et la saison n'étant significative sur aucun des sites, nous ne l'incluons pas dans le modèle.

Afin de tester ces différents effets, nous considérons le modèle suivant :

$$\log(Y_{ijkl}) = \mu + \alpha_j + \gamma_t + \theta_k + (\alpha\theta)_{jk} + (\gamma\theta)_{tk} + U_{ijkl} \quad (2.10)$$

où nous distinguons par :

- $\mu$  l'effet général ;
- $\alpha_j$  l'effet principal du niveau  $j$  du facteur fertilisation  $F$  ;
- $\gamma_t$  l'effet principal du niveau  $t$  du facteur saison  $T$  ;
- $(\alpha\theta)_{jk}$  l'effet de l'interaction entre la fertilisation  $j$  et le site  $k$  ;
- $(\gamma\theta)_{tk}$  l'effet de l'interaction entre la saison  $t$  et le site  $k$ .

De cette manière, nous testons les effets fixes du modèle, sans hypothèse sur la structure de covariance. Les résultats obtenus avec l'ANOVA sont présentés dans la table 2.5.

	DF	SSE	MSE	F	p-valeur
fertilisation $F$	9	167.2	18.6	92.8	$< 2.2 \times 10^{-16}$
saison $T$	3	216.8	72.3	361.1	$< 2.2 \times 10^{-16}$
site $S$	3	101.6	33.9	169.3	$< 2.2 \times 10^{-16}$
interaction $T \times S$	9	70.4	7.8	39.1	$< 2.2 \times 10^{-16}$
interaction $F \times S$	27	44.35	1.6	8.2	$< 2.2 \times 10^{-16}$
Résidus	3308	662.0	0.2		

TABLE 2.5 – Résultats des tests de l'ANOVA pour le modèle 2.10.

Note : la p-valeur est associée à la statistique  $F$  connaissant le nombre de paramètre  $DF$  à estimer.  $SSE$  et  $MSE$  correspondent à la somme et à la moyenne quadratique résiduelle.

Ainsi, d'après les tests, tous les effets sont significatifs au seuil 5%. Afin de vérifier la validité des effets des deux facteur croisés, nous effectuons les graphiques des interactions (Figure 2.6).

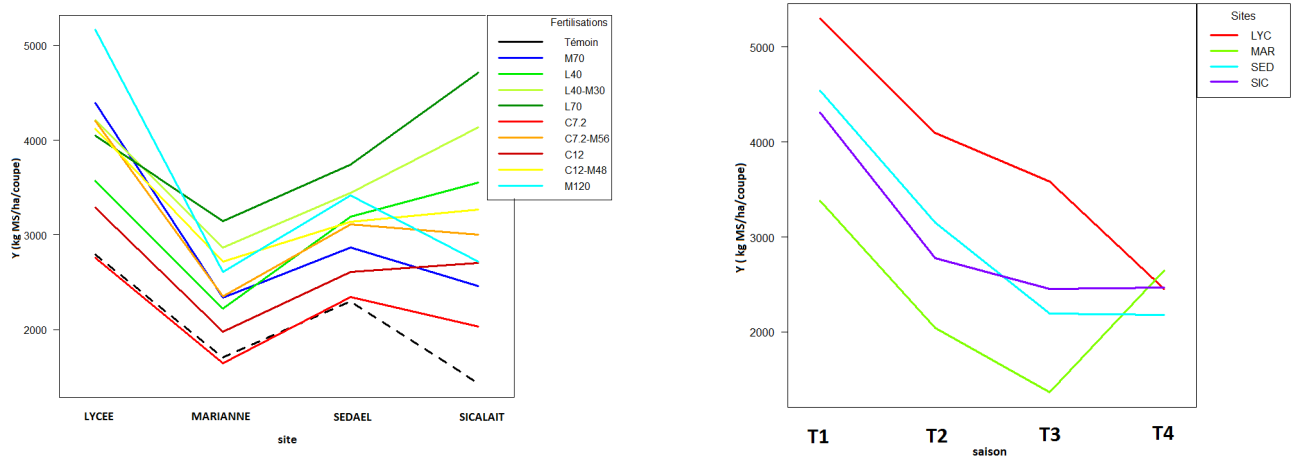


FIGURE 2.6 – Graphique des interactions a) entre la fertilisation et la site, b) entre la saison et le site, sur le rendement (en  $kg\ MS/ha/coupe$ ).

Sur les deux graphiques, nous constatons que les droites se croisent dans plusieurs cas. En conséquence, les modèles élaborés dans la suite comportent ces termes d'interaction.

## 2.5.2 Détection des anomalies

### Hétérogénéité saisonnière

Jusqu'à présent, nous n'avons pas tenu compte d'une éventuelle hétérogénéité de la variance. Cela signifie que la dispersion des mesures est supposée être la même pour tous les niveaux des facteurs. Dans la section 2.4, nous avons mis en évidence la différence de variabilité saisonnière à l'échelle d'un site. Ici, nous constatons que ces différences ne sont pas les mêmes d'un site à un autre. Le graphique 2.7 permet de visualiser la variabilité intra-site de la saison, ainsi que la variabilité inter-site. Malgré la transformation logarithmique du rendement, la variance ne semble toujours pas stabilisée. En posant  $\sigma_{k,t}$  l'écart-type associé aux mesures du site  $k$  lors de la saison  $t$ , nous supposons que :

$$\sigma_{k,t} \neq \sigma, \forall k, t$$

tels que  $k, t \in \{1, \dots, 4\}$ . Le test de Levene rejette l'hypothèse d'égalité de ces variances avec une p-valeur inférieure à  $2.2 \times 10^{-16}$  significative à 5%, tout comme le test de Bartlett (avec une p-valeur de  $1.45 \times 10^{-8}$ ). L'hypothèse d'homogénéité des groupes n'étant pas validée, il semble important de séparer les variances lors de l'estimation du modèle.

### Corrélation résiduelle

Nous cherchons ici à identifier les possibles corrélations entre les unités observées. Pour cela, nous représentons la fonction d'autocorrélation empirique obtenue à partir des résidus standardisés du modèle supposant l'indépendance des observations (Figure 2.8).

D'après la figure 2.8, nous observons des autocorrélations significatives parmi les résidus empiriques. De plus, les corrélations estimées (Figure 2.9) témoignent d'une dépendance entre les saisons. Nous avons donc recours à un modèle plus élaboré, permettant d'évaluer ces corrélations, tout en expliquant les effets fixes identifiés.

## 2.5.3 Le modèle général avec l'effet du site pour données répétées

Dans cette partie, nous souhaitons modéliser la dépendance entre les quatre répétitions effectuées sur une même micro-parcelle. Nous ajoutons l'effet fixe du site au modèle précédent (2.8), ainsi que les effets des interactions avec la saison d'une part et avec la fertilisation d'autre part. Les effets d'interactions entre la fertilisation et la saison ne sont toujours pas significatifs,

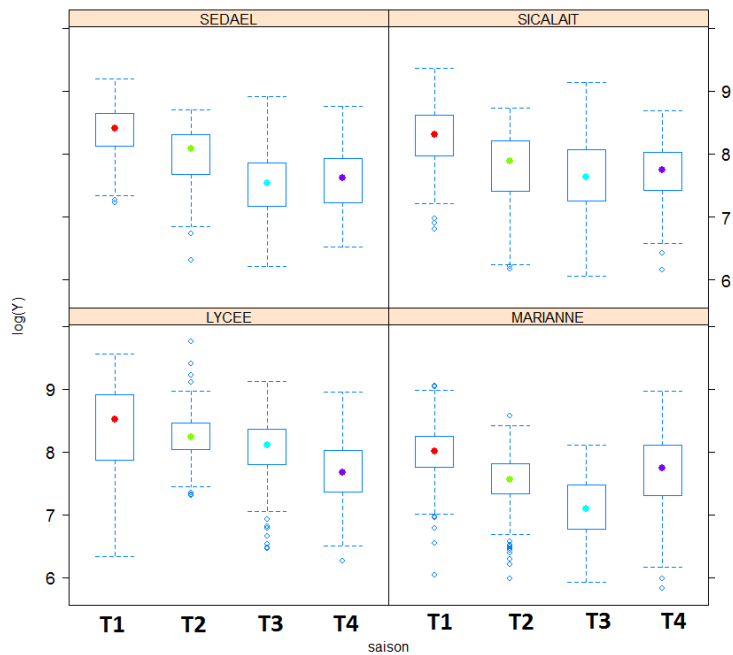


FIGURE 2.7 – Représentation en boîte à moustaches de la dispersion des observations du logarithme du rendement  $Y$  par site et par saison.

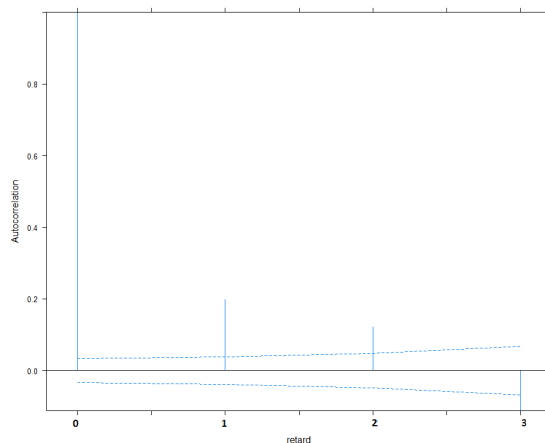


FIGURE 2.8 – Fonction d'autocorrélation empirique des résidus standardisés du modèle (2.10).

c'est pourquoi nous ne les considérons pas dans le modèle.

Ainsi, pour une mesure  $i$  prise en compte à un niveau  $j$  du facteur fertilisation  $F$  ( $j = 1, \dots, 10$ ) et à un niveau  $k$  du facteur site  $S$  ( $k = 1, \dots, 4$ ), avec  $i = 1, \dots, n_{jk} = 1, \dots, 84$ , on considère le vecteur aléatoire :  $\log(Y_{ijk}) = (\log(Y_{ijk1}), \dots, \log(Y_{ijkT}))'$  de  $\mathbb{R}^T$ , avec ici  $T = 4$ . Nous faisons l'hypothèse suivante :

$$\log(Y_{ij}) \sim \mathcal{N}_T(\mu_j, R_k),$$

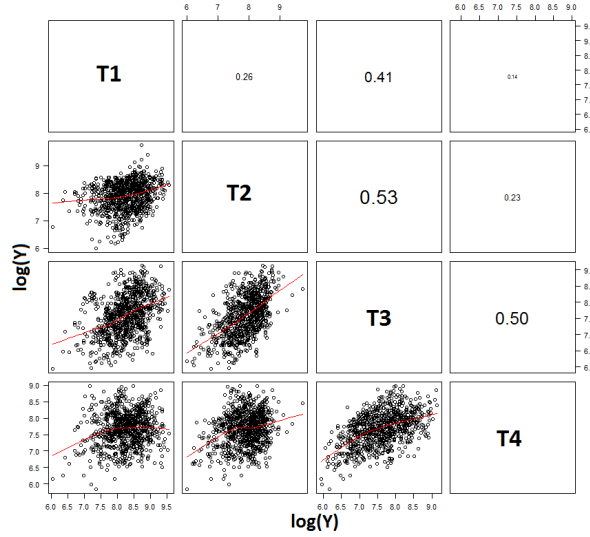
où  $\mu_{jk} = (\mu_{jk1}, \dots, \mu_{jkT})'$ . Par ailleurs, on définit :

$$\mu_{jkt} = \alpha_j + \gamma_t + \theta_k + (\alpha\theta)_{jk} + (\gamma\theta)_{tk}, \quad (2.11)$$

où on distingue comme précédemment :

- $\mu$  l'effet général ;
- $\alpha_j$  l'effet principal du niveau  $j$  du facteur fertilisation  $F$  ;
- $\gamma_t$  l'effet principal de la saison  $t$  ;
- $\theta_k$  l'effet principal du site  $k$  ;




 FIGURE 2.9 – Représentation de la matrice de corrélations estimées sur le  $\log(\text{rendement})$ .

- $(\alpha\theta)_{jk}$  l'effet de l'interaction entre la fertilisation  $j$  et le site  $k$  ;
- $(\gamma\theta)_{tk}$  l'effet de l'interaction entre la saison  $t$  et le site  $k$ .

Contrairement au cas précédent, la structure de variance-covariance dépend du site, afin de distinguer l'hétérogénéité inter-site. Ainsi, nous notons  $R_k$  la matrice de covariance associée au site  $k$ .

Le modèle pour le rendement est défini par :

$$\log(Y_{ijkt}) = \mu_{jkt} + U_{ijkt}, \quad (2.12)$$

où l'erreur  $U_{ijkt}$  est la  $t$ -ième composante du vecteur aléatoire gaussien centré  $U_{ijk}$  de  $\mathbb{R}^T$ , de matrice de variance-covariance  $R_k$ .

En résumé, les coefficients du modèle à estimer sont :

- l'effet général  $\mu$  ;
- les 10 effets principaux du facteur fertilisation  $F$ , avec une contrainte de centrage ;
- les 4 effets principaux du facteur site  $S$ , avec une contrainte de centrage ;
- les 40 ( $S \times J$ ) effets d'interactions, avec  $(S + J - 1) = 13$  contraintes pour le double centrage ;
- les 4 effets principaux du facteur saison  $T$ , avec une contrainte de centrage ;
- les 16 ( $S \times T$ ) effets d'interactions, avec  $(S + T - 1) = 7$  contraintes pour le double centrage.

Ici, les effets inter-sujets concernent les trois premiers (indépendants du temps) et les effets intra-sujets correspondent aux deux derniers (dépendants du temps). L'ajustement du modèle supposant l'homogénéité des variances consiste en l'estimation de  $52 + \frac{T(T+1)}{2} = 62$  paramètres, à partir de  $n * T = 840 * 4 = 3360$  observations.

#### 2.5.4 Choix de la structure de variance-covariance

La comparaison des différentes structures de covariance nous conduit à conserver la modélisation de type autorégressive d'ordre 1. Nous avons vu précédemment qu'il est nécessaire de tenir compte de l'hétérogénéité des variances, en plus de la corrélation des données. Ainsi, pour le site  $k$ , avec  $k = 1, \dots, 4$ , la structure retenue est la suivante :

$$R_k = \begin{pmatrix} \sigma_{k,1}^2 & \sigma_{k,2}\sigma_{k,1}\rho & \sigma_{k,3}\sigma_{k,1}\rho^2 & \sigma_{k,4}\sigma_{k,1}\rho^3 \\ \sigma_{k,2}\sigma_{k,1}\rho & \sigma_{k,2}^2 & \sigma_{k,3}\sigma_{k,2}\rho & \sigma_{k,4}\sigma_{k,2}\rho^2 \\ \sigma_{k,3}\sigma_{k,1}\rho^2 & \sigma_{k,3}\sigma_{k,2}\rho & \sigma_{k,3}^2 & \sigma_{k,4}\sigma_{k,3}\rho \\ \sigma_{k,4}\sigma_{k,1}\rho^3 & \sigma_{k,4}\sigma_{k,2}\rho^2 & \sigma_{k,4}\sigma_{k,3}\rho & \sigma_{k,4}^2 \end{pmatrix}_{4 \times 4}, \quad (2.13)$$

où  $\sigma_{k,t}$  correspond à la dispersion des mesures sur le site  $k$  lors de la saison  $t$ .

L'analyse graphique de la fonction d'autocorrélation indique que la corrélation entre les répétitions est pratiquement expliquée par le modèle (Figure 2.10).

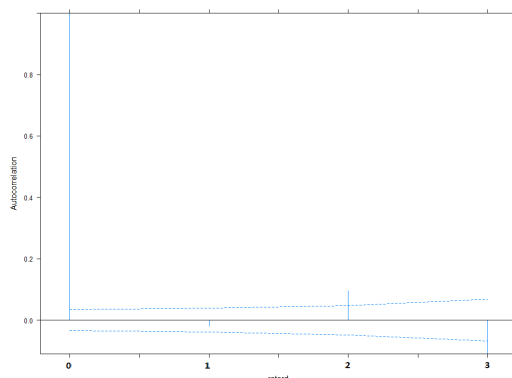


FIGURE 2.10 – Fonction d'autocorrélation empirique des résidus normalisés du modèle 2.12 comportant la structure de variance-covariance 2.13

D'après le test de Levene, au risque 5%, nous ne rejetons pas l'hypothèse d'égalité des variances résiduelles par saison et par site, puisque la p-valeur est de 0.534. Le test de Bartlett nous conduit aussi à la même conclusion : l'homogénéité de ces variances résiduelles n'est pas rejetée au seuil 5%, puisque la p-valeur est de 0.264.

Nous effectuons ensuite deux tests du rapport de vraisemblance : le premier compare les modèles avec et sans la structure de covariance  $AR(1)$ , mais tenant compte de l'hétérogénéité des variances, et le second compare le modèle retenu à celui supposant l'homogénéité des données. Ainsi, comme le montre le tableau 2.6, dans les deux cas nous rejetons l'hypothèse nulle au seuil 5%, ce qui signifie qu'il est mieux de tenir compte à la fois de la corrélation des données, mais aussi de l'hétérogénéité des variances.

Modèle	df	AIC	BIC	logLik
1 : hétérogénéité	68	4216.70	4631.78	-2040.35
2 : $AR(1)$	54	4293.82	4623.44	-2092.91
3 : hétérogénéité + $AR(1)$	69	4089.92	4511.11	1975.97

Test	L.Ratio	p-valeur
1 vs 3	128.77	< .0001
2 vs 3	233.89	< .0001

TABLE 2.6 – Comparaison des modèles retenus.

### 2.5.5 Résultats

Le modèle retenu comporte ainsi cinq effets fixes, correspondants à : la fertilisation, la saison, le site, l'interaction entre le site et la fertilisation et enfin l'interaction entre le site et la saison. La structure de covariance la plus adaptée est de type autorégressive d'ordre 1 ; le coefficient de corrélation est alors estimé par  $\hat{\rho} = 0.232$ . Le modèle suppose aussi l'hétérogénéité saisonnière intra-site. Au total, 16 composantes de la variance sont estimées. Pour cela, la dispersion de référence est l'écart-type obtenu sur le site de MARIANNE lors de la saison 1, c'est à dire  $\hat{\sigma}_{1,1}$ . Les autres variances sont obtenues à partir des coefficients multiplicatifs estimés. Les résultats sont résumés dans la table 2.7.

Le graphique 2.11 nous permet de visualiser les estimations du rendement (en  $kg$  MS/ha/coupe)

site $k$ / saison $j$	saison T1	saison T2	saison T3	saison T4
MAR. $\hat{\sigma}_{1,j}$	0.440	0.462	0.405	0.465
SED. $\hat{\sigma}_{2,j}$	0.383	0.497	0.494	0.379
SIC. $\hat{\sigma}_{3,j}$	0.378	0.434	0.412	0.361
LYC. $\hat{\sigma}_{4,j}$	0.734	0.357	0.433	0.421

TABLE 2.7 – Estimations des écart-types et des facteurs multiplicatifs associés à chaque saison et à chaque site.

par le modèle retenu pour toutes les combinaisons des modalités des facteurs, ainsi que l'intervalle de confiance (pointillées) associé à ces prédictions.

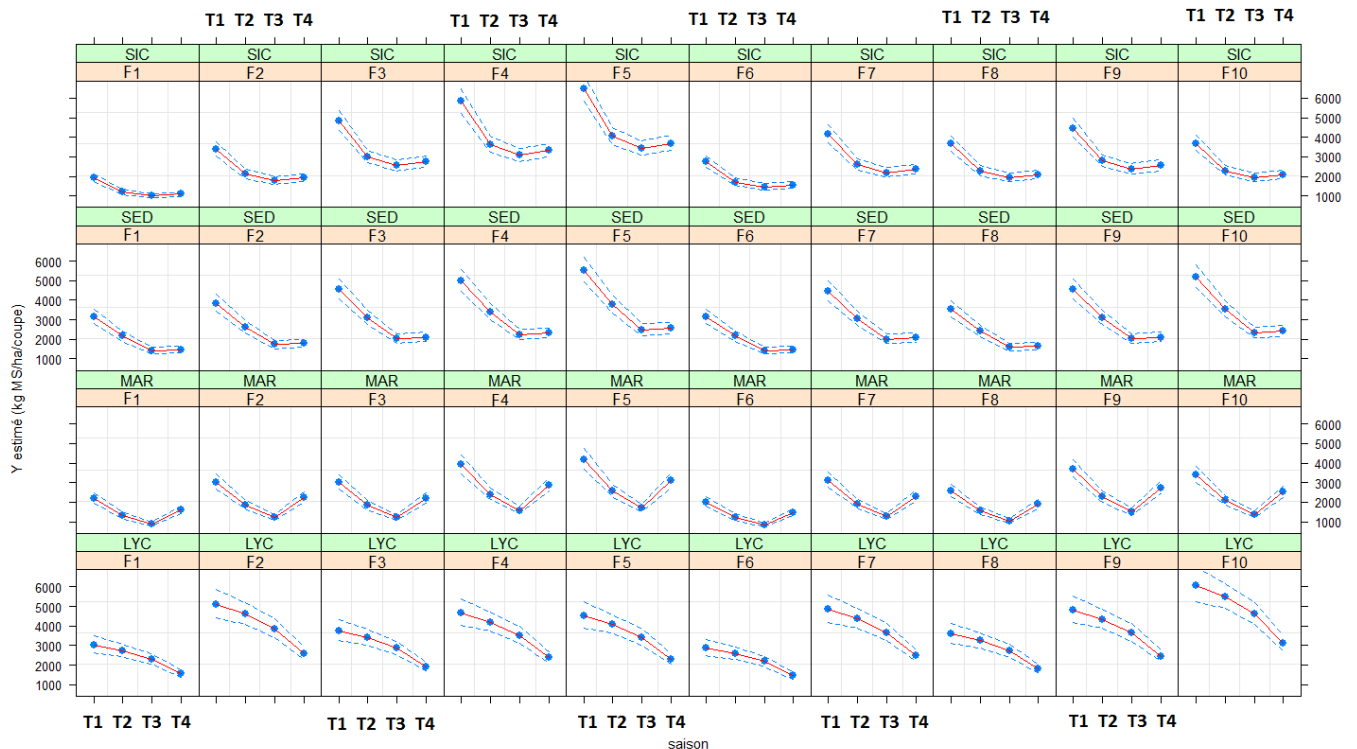


FIGURE 2.11 – Estimations du rendement en  $kg$  MS/ha/coupe par le modèle retenu pour chaque combinaison de modalités des facteurs.

Les sites sont notés par : SIC=SICALAIT, SED=SEDAEL, MAR=MARIANNE, LYC=LYCEE.

## 2.6 Interprétation agronomique des résultats

Suite à cette analyse, nous interprétons les tests de significativité des effets, ainsi que les estimations résultantes. Le test utilisé est le test  $t$  de Student. Un effet est dit significatif (c'est-à-dire que la variable ou l'interaction qui lui est associée a une influence sur la réponse), s'il est, pour le risque 5% fixé, significativement différent d'un niveau de référence choisi. Ce niveau est initialement conditionné par l'estimation du modèle sur le site du LYCEE pour la modalité témoin (F1) lors de la saison T1. Nous faisons varier le niveau de référence lorsque des comparaisons plus ciblées sont nécessaires. Le premier tableau de l'annexe G.1 présente un extrait des résultats obtenus pour les effets principaux (les effets croisés ne figurent pas dans le tableau pour faciliter la lecture). Pour quantifier ces résultats, nous nous appuyons aussi sur les estimations des rendements obtenues pour chaque combinaison de modalités des facteurs (tableaux de l'annexe G.1).

### 2.6.1 Des différences entre les sites

Nous constatons que les rendements fourragers sur le site du LYCEE (climat tropical, sols sableux) pour les modalités de fertilisation témoin (F1) et minérales (F2 et F10), en saison des pluies (T1), sont significativement supérieurs (environ 60%, calculé à partir des estimations des tableaux G.3 et G.2 des annexes) à ceux mesurés sur le site de SICALAIT. D'un point de vue agronomique, ceci s'explique par la supériorité, en termes de capacité de production de biomasse, des espèces fourragères tropicales (*Chloris*) présentes sur le site du LYCEE vis-à-vis des espèces fourragères tempérées (*Ray-grass*, *Dactyle*, *Brome*) utilisées en zone d'altitude (SICALAIT; 1610 m). Les fertilisations minérales apportent des éléments fertilisants (N, P, K) très solubles et rapidement valorisés par les plantes. Au contraire, le rendement fourrager obtenu avec la modalité de fertilisation F5 (lisier à niveau élevé) sur le site du LYCEE est significativement inférieur (de l'ordre de 30%) par rapport à celui mesuré sur le site de SICALAIT.

Nous observons également des différences significatives entre les rendements fourragers obtenus sur les sites du LYCEE et de MARIANNE lors de la saison pluvieuse (T1) et pour les modalités témoin, minérales (F2 et F10) et celles à base de compost (sauf pour F9). En moyenne, sur le site du LYCEE les rendements fourragers sont 55% supérieurs à ceux observés sur celui de MARIANNE (*Dactyle*). Ces résultats sont également liés à la capacité de production fourragère plus importante des espèces tropicales.

Lors de la saison des pluies (T1), les rendements fourragers de toutes les modalités de fertilisation sur le site de SEDAEL ne diffèrent pas significativement de ceux obtenus sur le site du LYCEE. L'espèce fourragère dominante sur le site SEDAEL est le *Kikuyu*, une graminée fourragère tropicale. Les rendements fourragers du *Kikuyu* sont similaires à ceux observés avec le *Chloris* (15 à 20 tonnes de MS/ha).

### 2.6.2 Des différences entre les saisons

En ce qui concerne les différences de rendement intersaisons, nous constatons que les rendements fourragers (toute modalité de fertilisation confondue) sur les sites de SICALAIT, de SEDAEL et de MARIANNE sont significativement supérieurs (environ 80% en moyenne) lors de la saison T1, saison chaude et pluvieuse, par rapport aux autres saisons de l'année. Les conditions climatiques favorables, caractérisées par des températures élevées et une pluviométrie très importante, permettent un développement fourrager nettement supérieur. Sur le site de SICALAIT (d'après les résultats des modèles par site), nous observons que les mesures des rendements lors des saisons T2 (fin de la saison des pluies) et T3 (saison fraîche), ainsi que T4 (fin de la saison fraîche) sont très corrélées. Ceci s'explique par un effet plus étendu (donc moins marqué) des conditions climatiques "hivernales" (saison fraîche) sur la diminution des rendements fourragers lors des deux saisons "proches".

Sur le site du LYCEE, nous constatons que le rendement fourrager de l'ensemble des micro-parcelles est significativement supérieur en saison T1 (saison des pluies; environ 60%), par rapport aux saisons T3 et T4 (saisons fraîches). Cependant, sur ce site, nous n'observons pas de différences significatives entre les deux saisons pluvieuses (T1 et T2) ce qui peut s'expliquer par une pluviométrie (et température) plus importante et constante (climat typiquement tropical) sur ce site par rapport aux autres sites expérimentaux situés en altitude.

### 2.6.3 Des différences entre les modalités de fertilisation

En ce qui concerne l'effet des modalités de fertilisation sur le rendement fourrager, et en considérant toute l'année (quatre saisons), nous observons que, sur le site de SICALAIT, toutes les modalités de fertilisation (organique, minérale et mixte) augmentent significativement le rendement fourrager (2,3 fois plus, en moyenne) par rapport à la modalité témoin (sans fertilisation). Ceci étant expliqué par l'effet de l'apport d'éléments fertilisants (N, P, K) nécessaires au développement des plantes. Parmi les modalités de fertilisation testées, les modalités F4 (lisier + minéral) et F5 (lisier seul à fort niveau) diffèrent significativement des autres modalités (d'après

les résultats des modèles par site). La production fourragère étant la plus élevée pour ces deux modalités (de l'ordre de 60%) par rapport aux autres modalités de fertilisation. Les apports de lisiers peuvent ainsi se substituer totalement aux apports d'engrais minéral. L'apport de compost à faible niveau d'incorporation (modalité F6) à un effet moins marqué sur l'augmentation du rendement fourrager (44% par rapport au témoin).

Sur les sites de SEDAEL et de MARIANNE, les apports de compost seul (modalités F6 et F8) n'augmentent pas significativement les rendements fourragers par rapport à la modalité témoin (sans fertilisation). Nous pouvons conclure que l'effet des éléments fertilisants présents dans le compost n'est pas visible sur le court terme. Curieusement, même si la différence n'est pas significative au seuil de 5%, les rendements fourragers obtenus sur le site de MARIANNE avec la modalité F6 (compost à faible niveau) sont inférieurs à ceux obtenus sans fertilisation (modalité témoin). Sur ces deux sites, les modalités F4 (lisier à faible niveau + engrais minéral), F5 (lisier seul à niveau élevé) et F10 (engrais minéral à fort niveau; uniquement sur le site de SEDAEL) présentent les rendements fourragers les plus élevés par rapport à la modalité témoin (74% en moyenne). Sur le site de MARIANNE, les modalités F4 et F5 ont un effet significatif sur l'augmentation des rendements fourragers (42%) par rapport aux autres modalités de fertilisation (à l'exception de la modalité F9). Sur le site de SEDAEL, les rendements obtenus avec les fertilisations F4 et F10 ne diffèrent pas significativement de ceux obtenus avec les modalités F3 (lisier à faible niveau), F7 et F9 (compost + engrais minéral). Nous pouvons conclure que, comme déjà observé pour le site de SICALAIT, les apports de lisiers (niveau élevé d'incorporation) peuvent ainsi se substituer aux apports d'engrais minéral. Pour des niveaux d'incorporation de lisier plus faibles, l'apport complémentaire d'Azote minéral pendant les premières années permet d'atteindre des niveaux de rendement acceptables et similaires à l'apport 100% minéral.

Sur le site du LYCEE, et en considérant toute l'année, nous observons également que toutes les modalités de fertilisation (à l'exception de la modalité F6; compost à faible niveau) augmentent significativement le rendement fourrager (54% en moyenne) par rapport à la modalité témoin. Parmi les modalités de fertilisation testées, les modalités d'engrais minéraux F2 et F10 présentent les rendements fourragers les plus élevés (84% par rapport au témoin), mais sans différence significative avec les modalités mixtes à base de compost (F7 et F9). Sur ce site, les apports de compost peuvent ainsi se substituer partiellement aux apports d'engrais minéral.

## 2.7 Conclusion

La stratégie de sélection de modèles adoptée ici a été la suivante : nous avons d'abord déterminé les meilleures fonctions "moyennes" dans un modèle à variances homogènes, avec choix des paramètres fixes, significatifs pour chacun des niveaux. En second lieu, nous avons élaboré une structure de corrélation, en comparant différentes possibilités. Enfin, lorsqu'il a été nécessaire, nous avons introduit une fonction de variance résiduelle.

Le modèle ajusté sur les quatre sites comporte les effets principaux : de la fertilisation, du site et de la saison, mais aussi les effets d'interactions du site et de la saison, ainsi que du site et de la fertilisation. La structure de corrélation associée aux observations d'une micro-parcelle sur une année s'est avérée être de type autorégressive d'ordre 1 (AR(1)). L'hétérogénéité des données est prise en compte par site et par saison, grâce à une correction de la matrice de variance-covariance par pondération des observations.

Ainsi, le modèle nous a permis d'estimer les effets principaux des fertilisations sur la biomasse produite, et de constater des différences à la fois entre certains sites mais aussi entre les saisons.

# Chapitre 3

## Analyse de l'évolution de la teneur en Azote et de la CEC

### 3.1 Problématique et description de l'étude

*Dans quelle mesure pouvons-nous évaluer l'impact de la fertilisation organique et/ou minérale sur le "capital-fertilité" des sols à moyen et long terme ?*

Dans cette partie, l'étude de la relation entre la composition chimique du sol, les traitements de fertilisation utilisés et les différents facteurs environnementaux (paramètres climatiques, production de biomasse) a été traitée à l'aide de modèles additifs semi-paramétriques (AM). Ce modèle est une généralisation des techniques de régression multiple (Buja, Hastie et Tibshirani [1], Fang et Chan [5]). Il peut être composé d'une somme de fonctions qui ne sont pas forcément paramétriques. Les modèles de type AM s'utilisent beaucoup pour étudier les liens non-linéaires entre les réponses et les variables explicatives.

Les méthodes d'estimation non-paramétriques sont plus flexibles, elles fournissent des estimateurs ou fonctions de lissage qui se situent entre la régression paramétrique et la pure interpolation entre chaque point. Cette flexibilité se contrôle habituellement par le paramètre de lissage  $\lambda$ .

Malgré une amélioration de la modélisation, nous introduirons ensuite des contraintes sur la matrice de variance-covariance afin de tenir compte de la variabilité des sols, ainsi que des corrélations entre les mesures, par l'intermédiaire des modèles additifs mixtes (AMM).

La première mesure de composition du sol étudiée est sa teneur en Azote total (minéral et organique). Destiné à former les éléments constitutifs des protéines, l'Azote est un facteur essentiel à la croissance des plantes, mais aussi un indicateur de la qualité et de la fertilité des sols.

Nous nous intéressons ensuite à la Capacité d'Échange Cationique (CEC), autre paramètre caractérisant la qualité et la fertilité des sols. Plus la CEC est élevée, plus le sol peut absorber ou libérer des cations qui sont alors mis à disposition des racines. Dans ce contexte, l'intérêt de la fertilisation est d'améliorer la Capacité d'Échange Cationique.

### 3.2 Analyse exploratoire des données longitudinales

L'objectif de l'étude est de proposer deux modèles : le premier explique les variations de la teneur en Azote, et le second concerne l'évolution de la CEC, pour une fertilisation donnée. Comme nous les avons présentées dans la section 1.2, les données de cette partie sont disponibles sous forme de séries temporelles, chacune comportant sept observations.

Graphiquement, l'hypothèse de linéarité de l'évolution de la teneur en Azote (Figure 3.1) ainsi que de la CEC (Figure 3.2) dans le temps est douteuse : les courbes moyennes par site sont irrégulières (courbes rouges). Notons que les échelles des ordonnées sont ajustées aux observations de chaque site afin de mieux visualiser la forme de la courbe. Ainsi, un modèle linéaire simple semblerait peu adapté à ce type d'analyse.

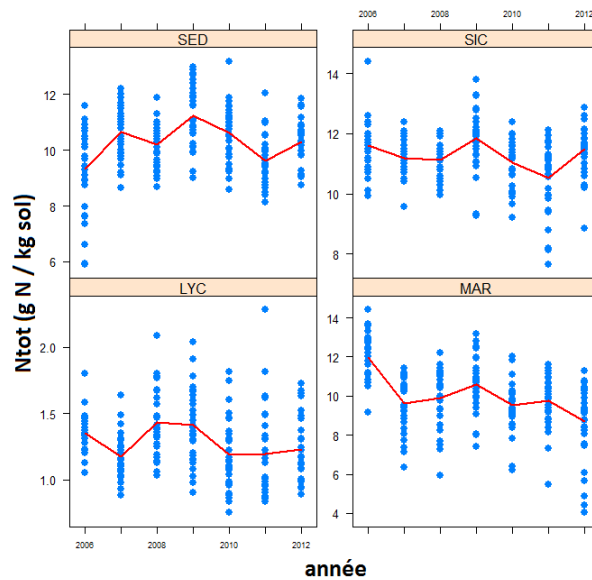


FIGURE 3.1 – Graphiques des observations de la teneur en Azote (en  $g N/kg$  sol) dans le temps selon le site.

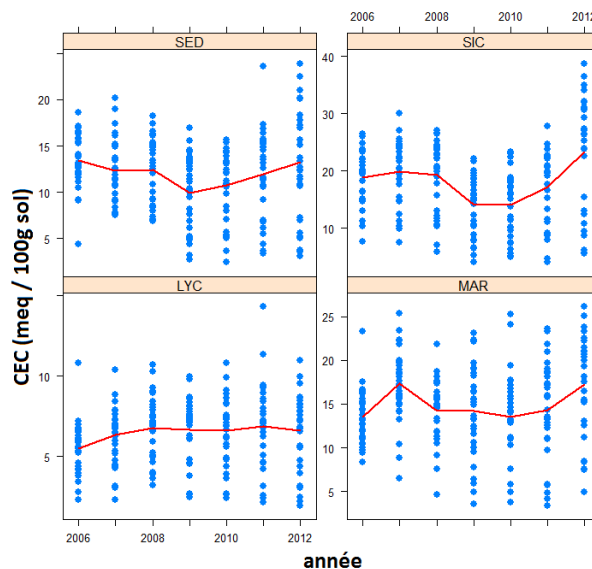


FIGURE 3.2 – Graphiques des observations de la CEC (en  $meq/100 g$  sol) dans le temps selon le site.

Afin de vérifier l’hypothèse de linéarité du temps sur ces variables, une première approche est mise en oeuvre : un test du manque d’ajustement ou “lack-of-fit”.

### 3.2.1 L’hypothèse de linéarité de l’Azote

Dans cette analyse, nous avons des répliquations des différents niveaux des variables prédictives. La somme des carrés des résidus (notée  $SSE(R)$ ) peut alors être séparée en deux composantes : l’erreur pure et le manque d’ajustement. Ceci implique de déterminer la part de la somme des carrés des résidus qui peut être prévue en ajoutant d’autres termes aux variables prédictives du modèle (par exemple, des termes d’interaction), et la part de la somme des carrés des résidus qui ne peut pas être prévue par d’autres termes (c’est-à-dire la somme des carrés de l’erreur pure, notée  $SSE(C)$ ). Il est alors possible d’effectuer un test du manque d’ajustement en utilisant le carré moyen des erreurs pures (noté  $MSE(C)$ ) comme terme d’erreur (J. Neill, [8], D. Wang et M. Conerly, [13]).

*Le test du manque d'ajustement*

Dans un premier temps, par souci d'homogénéité, nous ne prenons pas en compte les données concernant le site du LYCEE, qui montre en permanence des teneurs en Azote nettement inférieures à celles des autres sites.

L'hypothèse nulle testée pour le modèle linéaire multiple est la suivante :

$$H_0 : \mathbb{E}(Ntot_{ij}|annee) = \mu + fert_i + site_j + \beta \times annee, \quad (3.1)$$

où  $Ntot_{ij}$  est une mesure de l'Azote sur le site  $j$  utilisant la fertilisation  $i$ . Le paramètre  $fert_i$  est l'effet de la fertilisation  $i$ , pour  $i = 1, \dots, 10$ , et  $site_j$  est l'effet du site  $j$ , pour  $j = 1, \dots, 4$ . La variable  $annee$  est comprise entre 0 et 6.

Pour effectuer ce test, nous ajustons deux type de modèles afin de juger de la linéarité de la variable  $annee$ . Le modèle "réduit" est un modèle linéaire hierarchique tenant compte de la variable  $annee$  comme une variable continue. Nous obtenons la somme des résidus  $SSE(R) = 935.89$ . Le second modèle ajusté ("complet") considère la variable  $annee$  comme une variable catégorielle plutôt que quantitative, et ajuste donc la moyenne pour chaque année. Dans ce cas, nous obtenons  $SSE(C) = 868.98$ , qui est égale à la somme des carrés de l'erreur pure. La composante de l'erreur relative au manque d'ajustement est donc :  $SSE(LF) = SSE(R) - SSE(C) = 66.91$ .

La statistique de test se calcule de la manière suivante :

$$F^* = \frac{\frac{SSE(LF)}{df_{LF}}}{\frac{SSE(C)}{df_C}} = \frac{MSE(LF)}{MSE(C)}. \quad (3.2)$$

Sous  $H_0$ , cette statistique suit une loi de Fisher à  $(k - 2)$  et  $(n - (p + l + k + 1) + 3)$  ddl, où  $p$  est le nombre de fertilisations testées (10),  $l$  est le nombre de sites (3) et  $k$  est le nombre d'années, soit 7. Nous rejetterons l'hypothèse  $H_0$  au seuil  $\alpha = 0.05$  lorsque  $F^* \geq F_{(5,612,0.95)} = 2.229$ . Ici  $F^* = 9.42$ , nous rejetons donc l'hypothèse nulle, et la p-valeur associée est de  $1.16 \times 10^{-8}$ . Nous concluons qu'il est préférable de séparer les moyennes sur les sept années plutôt que d'ajuster un modèle linéaire sur la variable continue  $annee$ .

Afin de tenir compte d'une éventuelle dynamique commune aux quatre sites, nous incluons les données du site du LYCEE. Pour cela, nous utilisons la transformation logarithmique de la teneur en Azote, afin de réduire la dispersion et de normaliser la distribution des mesures. D'autres fonctions ont été testées (racine carrée, inverse, etc.) mais seul le logarithme a permis d'obtenir une distribution centrée et des quantiles alignés. Le modèle ainsi que les hypothèses restent les mêmes. L'hypothèse  $H_0$  est rejetée au seuil  $\alpha = 0.05$  lorsque  $F^* \geq F_{(5,821,0.95)} = 2.225$ , puisque  $n = 840$  et  $l = 4$ . La statistique de test obtenue est  $F^* = 11,63$ , nous rejetons ici aussi l'hypothèse nulle au risque 5%. La p-valeur associée au test est de  $7.1 \times 10^{-11}$ . Ainsi, l'hypothèse de linéarité de l'évolution de la teneur en Azote dans le temps n'est pas vérifiée.

### 3.2.2 L'hypothèse de linéarité de la CEC

Concernant la CEC, la structure des données est du même type que celle de la teneur en Azote. L'hypothèse de linéarité de la CEC dans le temps peut donc être testée de manière similaire.

*Le test du manque d'ajustement*

Dans cette partie, l'exploration graphique des données (figure F.4) nous conduit à considérer la totalité des observations pour le test. Afin de normaliser la distribution de la CEC et de réduire la dispersion des mesures, différentes transformations de cette variable ont été appliquées : logarithmique, inverse, angulaire, etc. La transformation en racine carrée semble la plus appropriée



et est donc retenue. L'hypothèse nulle testée pour le modèle linéaire multiple est la suivante :

$$H_0 : \mathbb{E}(\sqrt{CEC_{ij}}|annee) = \mu + fert_i + site_j + \beta \times annee, \quad (3.3)$$

où  $\sqrt{CEC_{ij}}$  correspond à une mesure (transformée) de la CEC sur le site  $j$  utilisant la fertilisation  $i$ . De même,  $fert_i$  est l'effet de la fertilisation  $i$ , pour  $i = 1, \dots, 10$ , et  $site_j$  est l'effet du site  $j$ , pour  $j = 1, \dots, 4$ .

Comme précédemment, nous ajustons deux type de modèles : le modèle "réduit" et le modèle "complet". Le premier considère la variable *annee* comme continue. Nous obtenons la somme des résidus  $SSE(R) = 168.01$ . Dans le second modèle, la variable *annee* est catégorielle. Nous obtenons alors la somme des carrés de l'erreur pure  $SSE(C) = 146.32$ . La composante de l'erreur relative au manque d'ajustement est donc :  $SSE(LF) = SSE(R) - SSE(C) = 21.69$ .

La statistique de test se calcule comme précédemment et suit aussi une loi de Fisher à  $(5, 821)$  ddl sous  $H_0$ . Ici  $F^* = 24.34$ , nous rejetons donc l'hypothèse nulle, et la p-valeur associée est inférieure à  $2.2 \times 10^{-16}$ . Comme pour la teneur en Azote, il est mieux de séparer les moyennes par année que d'estimer un coefficient linéaire associé à la variable *annee*.

Ainsi, l'évolution pluriannuelle de l'Azote et de de la CEC ne peut pas être considérée comme un processus linéaire.

### 3.3 Stratégie de la modélisation

L'objectif de cette analyse est de connaître l'effet global des fertilisations sur la teneur en Azote des sols, pour l'ensemble des parcelles expérimentales, représentatives des sols prairiaux de la Réunion. C'est pourquoi nous considérons les quatres sites lors de la modélisation AM, et pour ainsi tenir compte d'une dynamique commune traduisant mieux le phénomène à l'échelle de la Réunion.

#### 3.3.1 Présentation du modèle additif semi-paramétrique

Considérons une variable réelle  $y$  et les variables explicatives  $x_1, \dots, x_p$ . Le modèle additif lissé a la structure suivante :

$$y_i = X_i\beta + f_1(x_{1,i}) + f_2(x_{2,i}) + f_3(x_{3,i}, x_{4,i}) + \dots + \epsilon_i, \quad (3.4)$$

où  $X_i\beta$  est la partie linéaire (paramétrique) du modèle et les  $f_j$  correspondent aux fonctions de lissage. Les  $\epsilon_i$  sont *iid*, possiblement AR(1), tels que  $\mathbb{E}(\epsilon_i) = 0$  et  $\mathbb{V}(\epsilon_i) = \sigma^2$  (l'hypothèse de normalité sera à vérifier si besoin).

L'hypothèse que la contribution de chaque covariable est additive est analogue à celle du modèle linéaire, où chaque composante est estimée séparément. Ainsi, les différents effets présents dans le modèle modélisent les effets partiels, de la même façon que les coefficients du modèle linéaire. Nous détaillons l'approche des modèles additifs par les fonctions splines dans l'annexe E.2.

Tout au long de notre étude, nous choisissons d'approcher les fonctions de lissage par des **splines cubiques de régression pénalisée**. Plusieurs estimations du modèle ont été réalisées avec différents types d'estimateurs et les résultats obtenus sont sensiblement les mêmes. Cependant, le choix des noeuds peut avoir une influence non-négligeable sur les estimations. Le terme de pénalité permet alors de pallier à ce problème, puisqu'il réduit l'influence des coefficients associés aux noeuds afin d'obtenir une courbe plus lisse. C'est l'une des raisons pour lesquelles nous choisissons de travailler avec des splines cubiques de régression pénalisée.

De plus, nous utilisons cette pénalité dans le cadre des modèles mixtes pour l'estimation des coefficients. Cette approche aboutit à un système d'équations où les coefficients pénalisés sont estimés comme des effets aléatoires, variant autour de leur moyenne, tandis que les coefficients non-pénalisés sont traités comme des effets fixes. Cette variabilité (ou pénalité) est déterminée

par  $\lambda$ , qui est équivalent au ratio entre la variance de l'erreur et la variance engendrée par les coefficients de régression. La valeur de  $\lambda$  peut alors être estimée à l'aide de la méthode du Maximum de Vraisemblance Restreinte (REML) que nous introduisons dans l'annexe E.3.

L'utilisation de la fonction `gamm()` du package `mgcv` (S. Wood, [15]) du logiciel R (version 3.0.0, [12]) conduit par défaut à l'estimation du modèle via REML.

### 3.3.2 Critères et outils de comparaison

À chaque modèle obtenu, nous évaluons sa capacité prédictive grâce à l'indicateur d'écart "MAPE" (Mean Absolute Percentage Error), calculé de la façon suivante :

$$MAPE(y, \hat{y}) = \frac{100}{n} \cdot \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i|}.$$

Ce critère est un indicateur pratique de comparaison, puisqu'il se présente sous forme d'un pourcentage. Nous cherchons à minimiser cette moyenne des écarts en valeur absolue par rapport aux valeurs observées.

Un autre critère utilisé dans notre analyse est la racine carrée de l'erreur quadratique moyenne intégrée (RMSE). Nous cherchons à estimer  $\mathbb{E}(Y|X)$  de façon à minimiser :

$$RMSE = \sqrt{\mathbb{E}(\mathbb{E}(Y|X) - s(X))^2}.$$

Cependant, les critères MAPE et RMSE n'ont pas une importance fondamentale dans cette analyse : ils seront surestimés puisque notre base d'apprentissage est aussi celle utilisée pour la validation. Malgré tout, nous les utilisons comme aide à la décision afin de comparer des modèles potentiellement valides. D'autres outils classiques sont aussi utilisés pour comparer les différents modèles étudiés dans cette analyse : le test de rapport de vraisemblance lorsque les modèles sont emboîtés, les critères d'Akaike et de Schwarz (AIC et BIC) dans le cas échéant. Afin de sélectionner le meilleur modèle, nous cherchons à minimiser ces différents critères, tout en conservant un nombre de degrés de liberté convenable par souci de parcimonie.

## 3.4 Le modèle de la teneur en Azote des sols

La non-linéarité de l'évolution de la teneur en Azote dans le temps nous conduit à proposer un modèle semi-paramétrique additif expliquant les relations Azote/fertilisation.

### 3.4.1 Quatre fonctions de lissage et corrélation temporelle

La mise en oeuvre du modèle semi-paramétrique nécessite le choix des fonctions de lissage temporel. En complément des critères d'ajustement exposés dans la partie précédente, l'analyse exploratoire des données nous a permis de mettre en évidence les tendances principales de la teneur en Azote du sol.

#### Choix du modèle

Les graphiques 3.1 et 3.3 montrent un comportement différent de la teneur en Azote moyenne selon le site. La teneur en Azote du sol étant sensible au climat (température, pluviométrie, etc.), ces différences peuvent être expliquées par les variations météorologiques spécifiques au lieu. Nous confirmons cette hypothèse en examinant l'évolution de la teneur en Azote pour chaque traitement de fertilisation et chaque site (Figure 3.3).

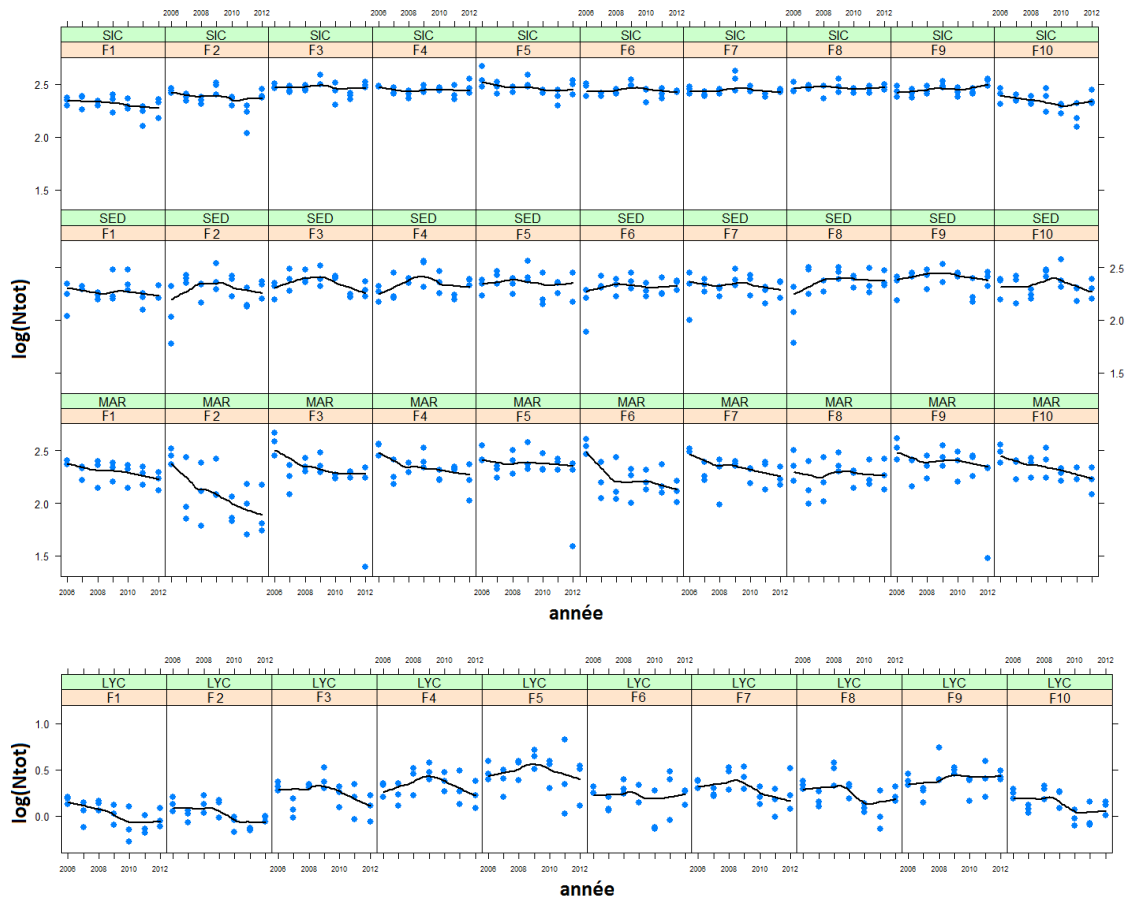


FIGURE 3.3 – Graphiques des observations de la teneur en Azote dans le temps selon le site et la fertilisation et courbe lissée de type *loess* associée.

En moyenne, l'effet fixe associé à la fertilisation sur la teneur en Azote semble alors différent d'un site à un autre. Cette interaction entre le site et la fertilisation peut également être observée sur la figure 3.4. En effet, entre deux sites, les droites relatives aux effets des fertilisations se croisent. Ainsi, en plus des effets principaux liés aux fertilisations et aux sites, il est nécessaire de tenir compte de cet effet d'interaction.

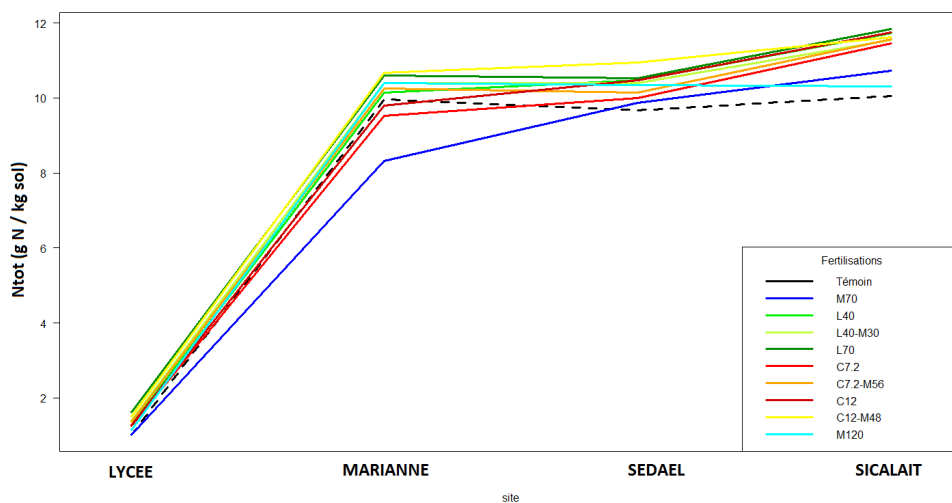


FIGURE 3.4 – Graphiques des interactions entre le site et la fertilisation sur la teneur en Azote.

Afin de réduire l'asymétrie des distributions et de diminuer l'influence des mesures atypiques,

nous considérons la transformation logarithmique de la teneur en Azote. Le modèle retenu appliqué à la teneur en Azote du sol peut s'écrire de la manière suivante :

$$\log(Ntot)_{ijt} = \mu + fert_i + site_j + (fert \times site)_{ij} + f_{1j}(annee_t) + \epsilon_{ijt} \quad (3.5)$$

où  $\log(Ntot)_{ijt}$  est le logarithme d'une mesure d'Azote sur le site  $j$  utilisant la fertilisation  $i$  lors de l'année  $t$ .

Nous distinguons par :

- $\mu$  la constante associée au modèle, représentant la teneur moyenne en Azote mesurée sur le site du LYCEE sans aucune fertilisation ( $i = 1$  et  $j = 1$ ),
- $fert_i$  l'effet de la fertilisation  $i$ , pour  $i = 1, \dots, 10$ ,
- $site_j$  l'effet du site  $j$ , pour  $j = 1, \dots, 4$ ,
- $fert \times site_{ij}$  l'effet de l'interaction entre la fertilisation  $i$  et le site  $j$ ,
- $f_{1j}(annee_t)$  la fonction de lissage associée au site  $j$  et à l'année  $t$ .

Les résidus sont supposé gaussien et de même variance, c'est à dire  $\epsilon_{ijt} \sim \mathcal{N}(0, \sigma^2)$ .

La fonction de lissage temporel  $f_{1j}$  dépend du site  $j$ . L'hypothèse sous-jacente suppose que la teneur en Azote évolue différemment selon le site. Cette fonction de régression partielle permet de capturer les fluctuations des données autour d'une composante annuelle spécifique au site. Comme vu précédemment (section ...), les  $f_{1j}$  sont approximées par des **splines cubiques de régression pénalisée**.

Ainsi, pour chaque site, nous ajustons une spline de régression pénalisée de **degré 3**. Pour chaque micro-parcelle, les mesures sont effectuées à des intervalles de temps égaux. Le nombre d'années n'étant pas très grand, nous avons la possibilité de placer les noeuds à chaque pas de temps. Entre chaque noeud, une fonction cubique est ajustée, pour laquelle la courbe est continue et différentiable. Cette interpolation nous permet de décrire au mieux la variabilité liée à chaque site, qui peut être dûe aux variations météorologiques ou encore aux différents types de sol. Ainsi, les effets des fertilisations pourront être interprétés à l'échelle globale, et indépendamment de ces paramètres.

Les observations des séries temporelles de la teneur en Azote sont proches dans le temps, les aléas  $\epsilon_{ijt}$  ne vérifient pas l'hypothèse d'indépendance, mais sont plutôt corrélés. Ainsi, afin d'assurer la stabilité du modèle, nous modélisons l'autocorrélation des résidus par un processus autorégressif d'ordre 1, ou  $AR(1)$ , tel que :

$$\epsilon_{ijt} = \rho \epsilon_{ij(t-1)} + u_{ijt},$$

où  $\mathbb{E}(u_{ijt}) = 0$ ,  $\mathbb{V}(u_{ijt}) = \sigma_u^2$ ,  $cov(u_{ijt}, u_{ijt'}) = 0$ , et  $|\rho| < 1$ . Cette dernière contrainte est appelée "condition de stationnarité", elle garantit que la variance des  $\epsilon_{ijt}$  tend vers une valeur limite  $\sigma^2$  lorsque  $t$  augmente. Ainsi, pour une mesure effectuée sur une micro-parcelle à une année donnée, une partie de l'aléa correspond à l'aléa de l'année précédente et l'autre partie correspond à l'innovation  $u_{ijt}$ . La structure de corrélation associée est la suivante :

$$cor(\epsilon_{ijt}, \epsilon_{ijt'}) = \begin{cases} 1 & \text{si } t = t' \\ \rho^{|t-t'|} & \text{sinon.} \end{cases}$$

La corrélation entre deux séries temporelles, c'est à dire entre les évolutions de la teneur en Azote sur deux micro-parcelles différentes, est supposée être nulle. Notons que la corrélation est appliquée à l'échelle la plus fine : les observations d'une même micro-parcelle. Nous appliquons ainsi le même coefficient de corrélation  $\rho$  à toutes les séries.

### Validation du modèle

Les graphiques diagnostics des résidus (figure 3.5) montrent que les hypothèses théoriques (supposées être les mêmes que pour la régression linéaire) semblent être assez bien vérifiées. En effet, les résidus sont supposés être indépendants, distribués selon une loi normale et de même variance. La variabilité des résidus selon les différents facteurs est mise en évidence sur

les graphiques 3.6. Nous pouvons tout de même constater une différence de variabilité assez marquée entre les sites (graphique 3.6 B).

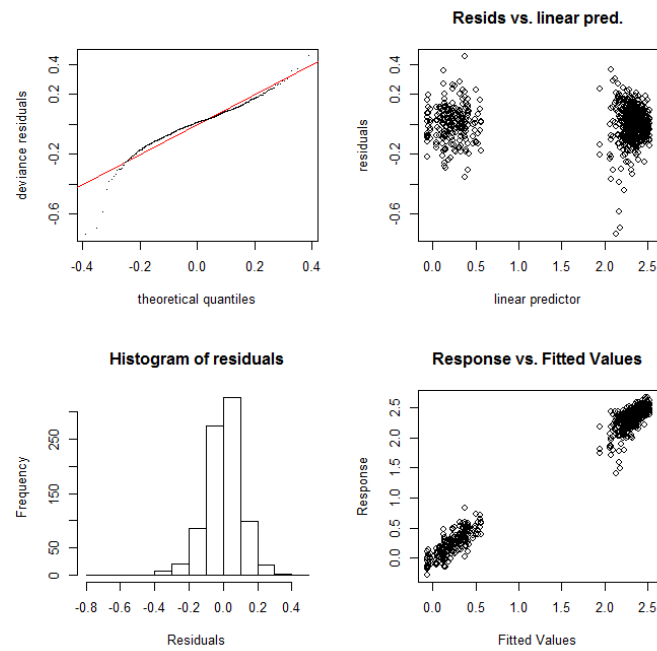


FIGURE 3.5 – Graphiques diagnostics des résidus du modèle de la teneur en Azote.

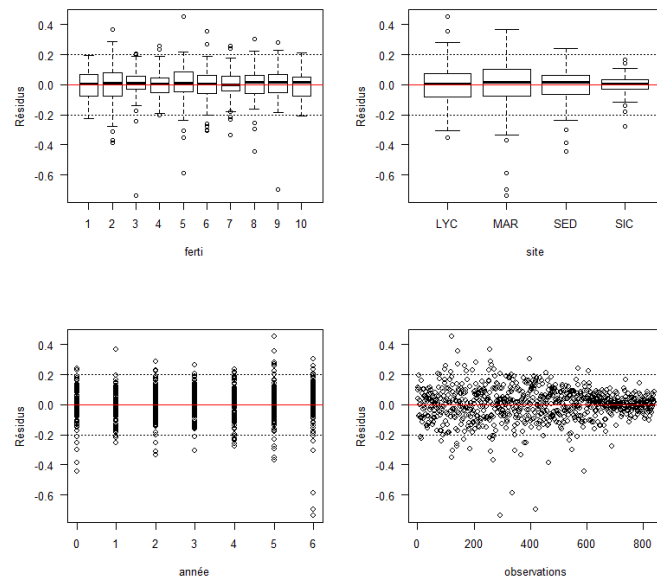


FIGURE 3.6 – Graphiques diagnostics des résidus du modèle de la teneur en Azote en fonction des facteurs et pour chaque observation.

## Résultats

Les degrés de liberté estimés ainsi que les résultats des tests de significativité des termes (test de Fisher) du modèle concernant la variable  $\log(Ntot)$  sont résumés dans le tableau 3.1. Ainsi, tous les facteurs explicatifs ainsi que les termes de lissage sont significatifs à 5% : il existe au moins deux modalités de fertilisation qui diffèrent par leur effet sur la teneur en Azote du sol. De même, il existe au moins un site sur lequel la teneur en Azote moyenne est significativement différente de celle d'un autre site. De plus, l'effet croisé est lui aussi significatif à 5%, ce qui

signifie que l'effet d'au moins une fertilisation diffère d'un site à un autre. Le tableau 3.2 présente les différents critères de comparaison et indicateurs d'écart du modèle calculés sur l'échantillon. Nous obtenons des résultats assez satisfaisants : les erreurs absolues relatives moyennes (MAPE) sont de 8.14% sur notre échantillon.

	df/edf	F	p-valeur
ferti	9.00	15.25	0.00
site	3.00	919.60	0.00
ferti × site	27	2.95	0.00
$f_1(\text{annee})  \text{site.LYC}$	5.64	14.01	0.00
$f_1(\text{annee})  \text{site.MAR}$	5.48	17.35	0.00
$f_1(\text{annee})  \text{site.SED}$	5.37	8.50	0.00
$f_1(\text{annee})  \text{site.SIC}$	4.62	2.63	0.03

TABLE 3.1 – Résultats des tests de significativité des termes paramétriques et non-paramétriques du modèle additif de la teneur en Azote.

AIC	RMSE	MAPE (%)	$R_{aj}^2$
-987.79	0.86	8.14	0.985

TABLE 3.2 – Critères de comparaison et indicateurs d'écart du modèle additif de la teneur en Azote .

L'intervalle de confiance à 95% pour la variance de l'erreur résiduelle  $\sigma$  est :  $[0.113, 0.128]$ , et le coefficient de corrélation  $\rho$  est estimé par  $\hat{\rho} = 0.424$ . Cela signifie que les résidus séparés d'une année ont une corrélation de 0.424, et pour deux années d'intervalle, la corrélation est de  $0.424^2 = 0.180$ , et ainsi de suite.

Les estimations des composantes de la variance associées aux paramètres de lissage sont présentées dans le tableau 3.3.

fonction	$\lambda_j$	$\hat{\sigma}^2/\lambda_j$
$f_1$ LYC	0.049	0.295
$f_1$ MAR	0.076	0.189
$f_1$ SED	0.096	0.150
$f_1$ SIC	0.311	0.046

TABLE 3.3 – Estimations des composantes de la variance associés aux paramètres de lissage du modèle additif.

### Comparaisons des modèles

L'estimation du modèle étant effectuée par des techniques de modélisation linéaire mixte, nous avons accès à la vraisemblance complète du modèle. Les tests basés sur la vraisemblance sont donc applicables dans ce contexte. Le test suivant nous permet de justifier de la pertinence d'une fonction de lissage temporel spécifique au site (tableau 3.4).

Modèle	df	AIC	BIC	logLik	Test	L.Ratio	p-valeur
$f_1$	44	-828.87	-622.81	458.44			
$f_1 site$	50	-987.79	-753.81	543.90	1 vs 2	170.92	< .0001

TABLE 3.4 – Comparaison du modèle avec une fonction de lissage temporel par site et du modèle avec une fonction de lissage commune.

Ainsi, comme le confirme le test du rapport de vraisemblance, il est préférable d'utiliser une

fonction de lissage temporel par site plutôt que de spécifier une tendance générale commune à tous les sites. Notons que le modèle comportant une fonction de lissages spécifique à chaque fertilisation s'est avérée non-significative (toujours au seuil 5% fixé) pour chacune des modalités de traitement.

De même, l'hypothèse d'indépendance entre les observations de la teneur en Azote sur une même micro-parcelle peut être remise en question. Le test résumé dans le tableau 3.5 compare le modèle sans structure de corrélation sur les résidus et le modèle construit précédemment, comportant une structure  $AR(1)$ .

Modèle	df	AIC	BIC	logLik	Test	L.Ratio	p-valeur
$\rho = 0$	49	-880.34	-651.04	489.17			
$\rho \neq 0$	50	-987.79	-753.81	543.90	1 vs 2	109.45	< .0001

TABLE 3.5 – Comparaison du modèle avec autocorrélation des résidus et du modèle supposant l'indépendance des observations.

Encore une fois, l'hypothèse nulle est rejetée au seuil 5%. Selon le test du rapport de vraisemblance et le critère de l'AIC, il est clair qu'une structure de corrélation temporelle permet d'améliorer les estimations.

### Détection des anomalies

D'après les résultats des tests et des critères obtenus, le modèle ainsi construit paraît adapté à nos données. Cependant, malgré la transformation de la variable dépendante, une des hypothèses les plus importantes de la régression linéaire et additive, qui suppose l'homogénéité de la variance, ne semble pas vérifiée. En effet, le graphique des résidus (Figure 3.6B, en haut à droite) témoigne d'une variabilité différente selon le site, particulièrement visible sur celui de MARIANNE. Cette dispersion semble plus réduite à SICALAIT.

Afin de compléter cette inspection visuelle, il est possible d'appliquer un test d'homogénéité de la variance, par exemple le test de Bartlett. Ainsi, l'hypothèse nulle testée suppose que la variance est la même pour tous les sites. Nous obtenons une p-valeur inférieure à  $2.2 \times 10^{-16}$  (avec  $K^2 = 170.54$ ), nous pouvons donc rejeter l'hypothèse d'homogénéité des variances au seuil 5%.

Ignorer l'hétérogénéité des résidus peut entraîner des écart-types incorrects lors de l'estimation des paramètres, conduisant à de mauvaises approximations des statistiques de test. Afin d'améliorer le modèle, la partie suivante introduit un modèle plus complexe, mais comportant une structure de variance plus adaptée.

### 3.4.2 Hétérogénéité intra-site

#### Amélioration du modèle

Le modèle précédent suppose que les résidus sont normalement distribués, de variance égale pour toutes les séries temporelles, et permet une corrélation au sein d'une même série. Dans ce sens, cela suppose aussi que la variance est la même d'un site à un autre. Une solution pour tenir compte de cette hétérogénéité est d'utiliser un modèle supposant une dispersion différente selon le site. Ainsi, nous conservons le modèle additif gaussien précédemment construit, c'est à dire :

$$\log(Ntot)_{ijt} = \mu + fert_i + site_j + (fert \times site)_{ij} + f_{1j}(annee_t) + \epsilon_{ijt}, \quad (3.6)$$

mais nous imposons une structure de variance résiduelle différente, de la forme :

$$\epsilon_{ijt} \sim \mathcal{N}(0, \sigma_j^2), \text{ pour } j = 1, \dots, 4.$$

Par conséquent, nous supposons que  $\mathbb{V}(\epsilon_{ijt}) = \sigma_j^2$ , ce qui permet à chaque site d'avoir une variance différente. La structure de corrélation des données reste similaire à celle du modèle précédent ( $AR(1)$ ).

### Validation du modèle

Après inspection graphique (figures 3.7 et 3.8), les résidus semblent mieux vérifier les hypothèses théoriques, y compris l'homoscédasticité.

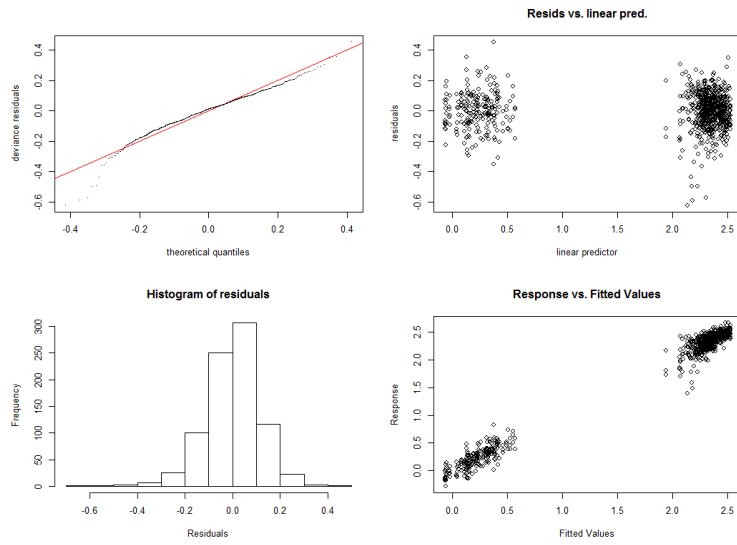


FIGURE 3.7 – Graphiques diagnostics des résidus du modèle.

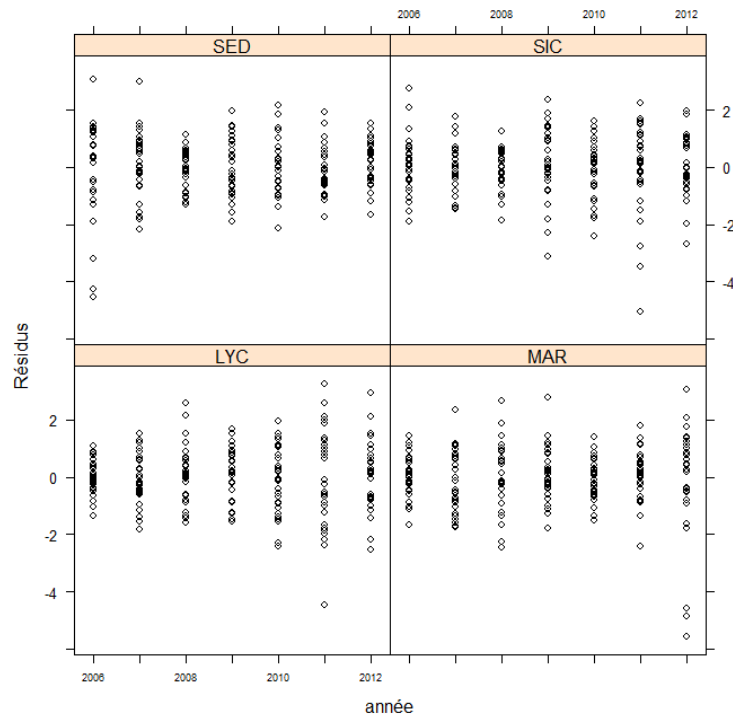


FIGURE 3.8 – Graphiques des résidus du modèle 3.6 selon le site et l'année.



## Résultats

La significativité des différents paramètres du modèle se trouve améliorée, plus particulièrement pour la fonction de lissage temporel associée au site de SICALAIT, comme en témoignent les résultats des tests (tableau 3.6).

	df/edf	F	p-valeur
ferti	9.00	14.60	0.00
site	3.00	1018.10	0.00
ferti $\times$ site	27	3.38	0.00
$f_1(\text{annee})$   site.LYC	5.59	13.39	0.00
$f_1(\text{annee})$   site.MAR	5.18	12.68	0.00
$f_1(\text{annee})$   site.SED	5.41	9.27	0.00
$f_1(\text{annee})$   site.SIC	5.55	11.78	0.00

TABLE 3.6 – Résultats des tests de significativité des termes paramétriques et non-paramétriques du modèle additif de la teneur en Azote.

Le tableau 3.7 présente les différents critères de comparaison et indicateurs d'écart du modèle, mais seul l'AIC est amélioré par rapport au modèle homogène.

AIC	RMSE	MAPE (%)	$R_{aj}^2$
-1132.33	0.86	8.14	0.985

TABLE 3.7 – Critères de comparaison et indicateurs d'écart du modèle additif de la teneur en Azote.

Ainsi, ce second modèle implique l'estimation de trois paramètres supplémentaires. Les variances sont estimées en considérant le site du LYCEE comme étant le site de référence, les variances des trois autres sites étant calculées en multipliant la variance obtenue sur le site de référence par un facteur de variation estimé à partir des données. Les estimations des dispersions résiduelles sont résumées dans le tableau 3.8.

$site_j$	LYC (ref.)	MAR	SED	SIC
$\hat{\sigma}_j$	0.127	0.151	0.114	0.060
$\hat{\sigma}_j/\hat{\sigma}_1$	1	1.188	0.895	0.468

TABLE 3.8 – Estimations des écart-types et des facteurs multiplicatifs associés à chaque site.

Comme attendu, le site de MARIANNE présente le ratio le plus élevé, indiquant une variabilité résiduelle plus importante sur ce site. Les estimations des composantes de la variance associés aux paramètres de lissage sont présentées dans le tableau 3.9.

fonction	$\lambda_j$	$\hat{\sigma}_1^2/\lambda_j$
$f_1$ LYC	0.056	0.291
$f_1$ MAR	0.094	0.172
$f_1$ SED	0.108	0.150
$f_1$ SIC	0.285	0.057

TABLE 3.9 – Estimations des composantes de la variance associés aux paramètres de lissage du modèle additif.

Le coefficient de corrélation est estimé par  $\hat{\rho} = 0.385$ , les résidus obtenus sur une même micro-parcelle pour deux années successives sont encore assez corrélés.

En annexe, le tableau G.6 résume les estimations des effets principaux (paramétriques) du

modèle, ainsi que la significativité associée à chaque niveau des facteurs. Les estimations associées aux effets d'interaction entre le site et la fertilisation sont présentées dans le tableau G.6. Les effets partiels des quatre fonctions de lissage approchées sont représentés sur la figure 3.9.

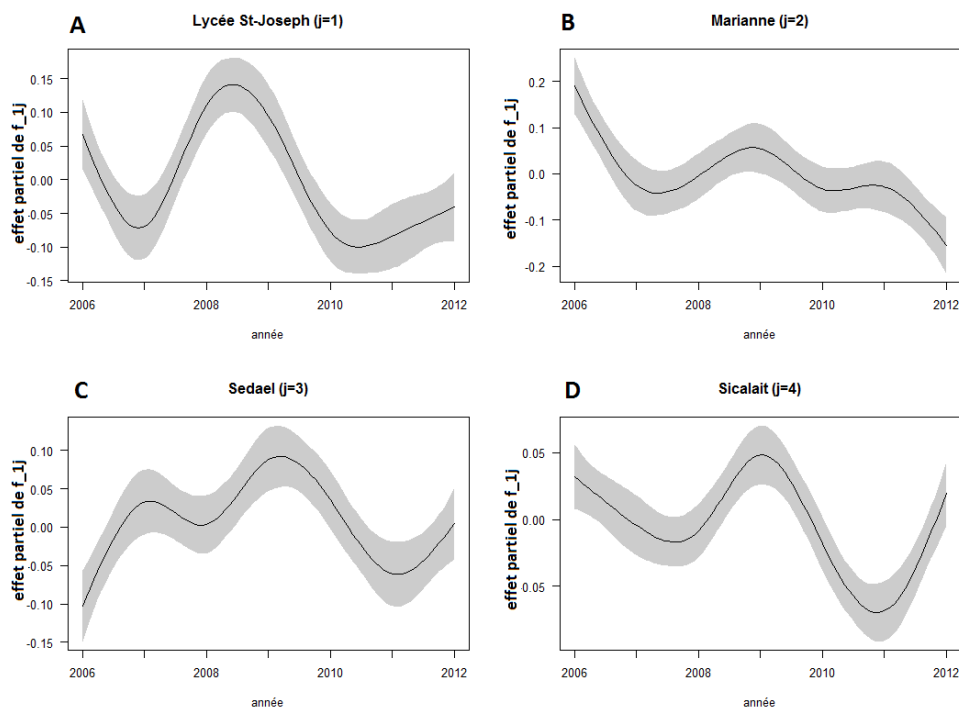


FIGURE 3.9 – Graphiques des quatre fonctions de régression partielles estimées sur le logarithme de la teneur en Azote.

### Comparaisons des modèles

Les deux modèles construits étant imbriqués, nous utilisons le test de la log-vraisemblance afin de les comparer. L'hypothèse nulle testée est :

$$H_0 : \sigma_1 = \dots = \sigma_4$$

et l'hypothèse alternative suppose qu'au moins deux composantes diffèrent.

Modèle	df	AIC	BIC	logLik	Test	L.Ratio	p-valeur
Homogénéité	50	-987.79	-753.81	543.90			
Hétérogénéité	53	-1132.33	-884.31	619.16	1 vs 2	150.54	< .0001

TABLE 3.10 – Comparaison du modèle hétérogène et du modèle homogène.

D'après le test (tableau 3.10), au risque 5%, nous rejetons l'hypothèse nulle qui suppose l'égalité des variances intra-sites, et concluons que le modèle comportant une variance différente par site est meilleur au sens de la vraisemblance. En annexe, le graphique G.1 nous donne une idée des estimations de la teneur en Azote (en  $g N/kg$  sol) par ce modèle pour toutes les combinaisons des modalités des facteurs (fertilisation/site) sur les 7 premières années de l'étude, ainsi que l'intervalle de confiance (pointillées) associé à ces prédictions.

### 3.4.3 Interprétation agronomique des résultats

Le choix du modèle nous conduit à analyser et interpréter les effets principaux du modèle, à l'aide des tests de significativité (tests  $t$ ). Le seuil de significativité reste fixé à 5%. Le niveau de référence choisi est le site du LYCEE pour la modalité témoin (F1).

L'analyse du graphique (Figure 3.1) a montré que l'évolution de la teneur en Azote des sols, toute modalité de fertilisation confondue, est variable selon le site et les années considérées. Les variations (positives ou négatives) de la teneur en Azote des sols dans le temps dépendent du site considéré. Nous avons aussi constaté une non-linéarité de l'évolution de la teneur en Azote des sols dans le temps. Le modèle retenu suppose que les effets des fertilisations sur la teneur en Azote du sol sont stables dans le temps ; seul l'effet du site varie. L'effet du site sur l'évolution de la teneur en Azote des sols s'explique par la variabilité de composition chimique et de structure des sols entre sites (sols plus au moins carencés/riches en éléments fertilisants, perméabilité à l'eau, pH, etc.) ou encore à la variabilité des conditions climatiques (température, pluviométrie, ensoleillement, etc.) qui jouent un rôle direct sur la production de biomasse fourragère et par conséquent sur l'exportation d'éléments fertilisants (dont l'Azote) par les plantes. C'est pourquoi l'analyse de l'évolution de la teneur en Azote des sols est réalisée au niveau de chaque site expérimental.

Les sols du site du LYCEE présentent les teneurs en Azote les plus faibles (90% moins riches) par rapport à la moyenne des teneurs en Azote observée dans les trois autres sites (SICALAIT, SEDAEL et MARIANNE). En effet, dans le cas du site du LYCEE nous sommes en présence de sols du type sableux qui ont une faible réserve en eau, sont pauvres en éléments nutritifs et ont une faible capacité d'échange anionique et cationique, contrairement aux sols des autres sites qui sont des andosols (sols volcaniques assez jeunes) avec des teneurs fortes en matière organique et des capacités d'échange anionique et cationique élevées. Sur ce site, nous observons que l'évolution de la teneur en Azote au cours du temps est très variable (augmentation, diminution, stabilisation) pour toutes les modalités de fertilisation, à l'exception des modalités F6 et F9 (Figure 3.3). Globalement, toute modalité de fertilisation confondue, nous constatons une légère diminution de la teneur en Azote des sols entre la première et la dernière année de l'étude (Figure 3.9,A). En ce qui concerne les modalités de fertilisation, à l'exception de la modalité F2 (minérale à faible niveau) et dans une moindre mesure F10 (minérale à niveau élevé ; au seuil 1%), toutes les modalités de fertilisation (organiques et mixtes) ont un effet significatif et positif sur la teneur en Azote des sols par rapport à la modalité témoin. La modalité F5 (lisier à niveau élevé) présente les effets les plus importants et significatifs de tous les traitements (à l'exception de F9). Dans le cas de l'apport d'engrais azoté uniquement minéral (F2 et F10), très soluble, et compte tenu de l'absence de stockage que nous observons dans ce type de sol, l'excédent azoté risque fort d'être lessivé vers les aquifères (ou nappes phréatiques). A très fort niveau d'apport minéral, cela peut représenter un gaspillage de près de 40% de l'Azote apporté.

Sur le site de MARIANNE, nous observons que toutes les modalités de fertilisation (à l'exception de T5 et T9) engendrent une diminution de la teneur en Azote au cours du temps (Figure 3.3). L'analyse portée sur l'ensemble des modalités de fertilisation (Figure 3.9,B) montre que la teneur en Azote du sol diminue la première année, augmente pendant les deux années suivantes, diminue à nouveau lors des trois dernières années de l'étude. Sur ce site, nous n'observons pas de différences significatives entre les modalités de fertilisation (à l'exception de F2) sur la teneur en Azote des sols. Même dans le cas des micro-parcelles témoin, qui ne reçoivent aucune fertilisation depuis plus de huit années, leur teneur en Azote ne diffère pas de celles qui reçoivent des quantités considérables de fertilisants (organiques et minéraux). Il faut néanmoins préciser que l'apport de fertilisants (à l'exception des modalités de compost seul ; F6 et F8) augmentent significativement les rendements fourragers par rapport à la modalité témoin. Une bonne partie de l'Azote apporté par la fertilisation est exporté par la plante sous forme de protéine végétale (biomasse fourragère). Curieusement, la modalité de fertilisation F2 (engrais minéral à faible niveau) fait baisser significativement la teneur en Azote des sols, par rapport au traitement témoin. Dans ce cas précis, l'exportation d'Azote par la plante n'est pas entièrement compensée par l'apport azoté de l'engrais minéral et les sols perdent une partie de leur "capital de fertilité" original.

Sur le site de SEDAEL, nous n'observons pas d'évolutions marquées de la teneur en Azote des sols au cours du temps pour chacune des modalités de fertilisation utilisées (Figure 3.3). Globalement, pour toute modalité de fertilisation confondue, nous constatons une augmentation

de la teneur en Azote des sols au cours des trois premières années de l'étude suivie d'une diminution de la teneur en Azote lors des trois années suivantes (Figure 3.9,C). Sur ce site, nous n'observons pas de différences significatives entre les modalités de fertilisation (à l'exception de F9) sur la teneur en Azote des sols. Seule la modalité de fertilisation F9 (compost à niveau élevé + engrais minéral) fait augmenter significativement la teneur en Azote des sols, par rapport aux modalités témoin (F1) et minérale à faible niveau (F2). Dans ce cas, l'Azote organique apporté par le compost de la modalité F9, non utilisé directement par la plante, est en partie stocké dans le sol, contribuant ainsi à accroître le "capital de fertilité" de la micro-parcelle pour produire un "arrière effet" à plus long terme.

Sur le site de SICALAIT, nous observons une légère diminution de la teneur en Azote au cours du temps pour les modalités de fertilisation témoin (F1) et minérale (F2 et F10; 3.3). Globalement, pour toute modalité de fertilisation confondue, nous constatons une stabilisation de la teneur en Azote des sols au cours des trois premières années de l'étude, une forte diminution de la teneur en Azote lors des deux années suivantes, et une augmentation de la teneur en Azote la dernière année de l'étude (Figure 3.9,D). Sur ce site, à l'exception de la modalité F10 (engrais minéral à niveau élevé), toutes les modalités de fertilisation ont un effet significatif et positif sur la teneur en Azote des sols par rapport à la modalité témoin. Les modalités de fertilisation composées exclusivement de lisier (F3 et F5) ont les plus grandes différences (positives) de teneur en Azote par rapport à la modalité témoin. Elles sont suivies de près par les fertilisations à base de compost à niveau élevé (F8). Cependant, les effets positifs des modalités F3 et F5 sur les teneurs en Azote des sols ne sont pas significatifs des autres modalités de fertilisation organiques et mixtes. En parallèle, nous observons que les rendements fourragers moyens des modalités de fertilisation à base de lisier sont significativement les plus élevés par rapport au témoin. Les modalités de fertilisation à base de lisier permettraient, simultanément, un rendement fourrager plus important et une augmentation de la teneur en Azote du sol.

### 3.5 Le modèle de la CEC des sols

Le problème de non-linéarité concerne aussi l'évolution de la CEC dans le temps. Afin de mieux expliquer le phénomène, nous élaborons là encore un modèle semi-paramétrique additif. Nous allons voir dans cette partie que deux choix sont possibles pour modéliser la relation de la CEC avec le temps.

#### 3.5.1 Un lissage temporel selon le type de site

##### Choix du modèle : première approche

Les variations moyennes de la CEC calculées sur les quatre types de sites sont représentées sur la figure 3.2 de la section 3.3. Les différences inter-sites sont moins visibles que pour la teneur en Azote. Lors des dernières années, une tendance à la hausse est commune pour trois sites (SEDAEL, SICALAIT et MARIANNE). Toutefois, l'évolution de la CEC sur le site de la SICALAIT se démarque nettement des trois autres.

Dans un premier temps, nous supposons qu'il existe des différences de variation de la CEC dans le temps selon le site. L'effet du site sur la CEC est justifié par le fait que la CEC dépend beaucoup de la texture du sol, et les différents sites étudiés reflètent la diversité des sol réunionnais. Plus il y a de l'argile dans un sol, plus la CEC est élevée et plus il peut stocker et échanger les éléments nutritifs.

Contrairement à l'Azote, il ne semble pas y avoir un effet significatif de l'interaction entre le site et la fertilisation. Très peu de droites se croisent sur la figure 3.10. Excepté pour les fertilisations F4 et F5 (lisier et lisier+minéral), la majorité des fertilisations conserve le même effet sur la CEC d'un site à un autre. Ainsi, le modèle considéré dans cette partie ne tiendra pas compte de cet effet croisé. Pour certains types de fertilisation, la quantité de biomasse fourragère annuellement produite semble corrélée à la Capacité d'Échange Cationique, comme

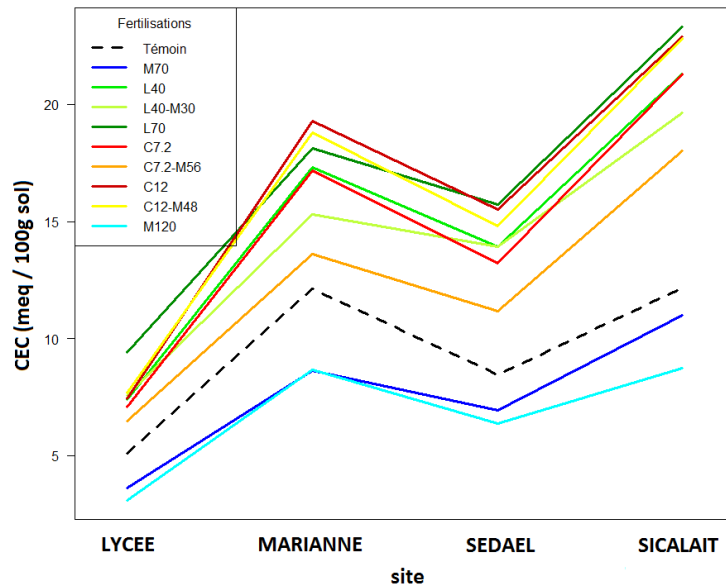


FIGURE 3.10 – Graphiques des interactions entre le site et la fertilisation sur la CEC (en  $meq/100g$ ).

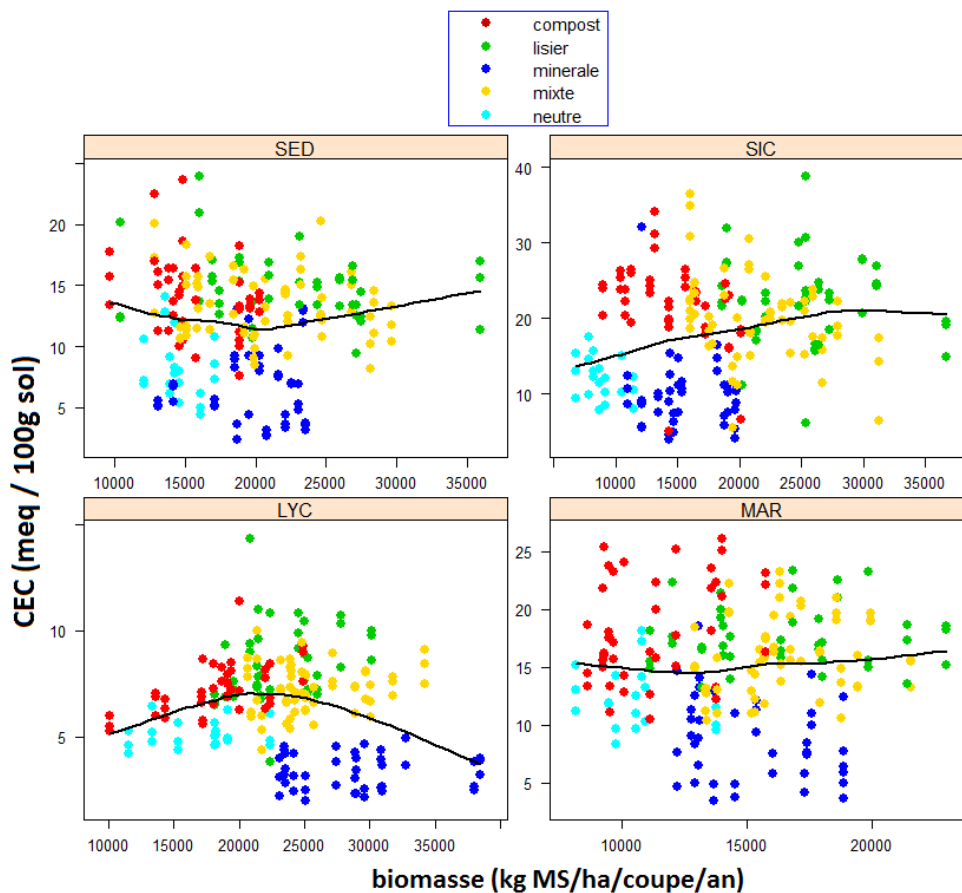


FIGURE 3.11 – Graphiques de la CEC calculée (en  $meq/100g$ ) sol en fonction de la biomasse annuelle (en  $kg MS/ha/coupe/an$ ) selon le site.

le montre le graphique 3.11. Nous observons que des sols sans fertilisation sont associés à une production de biomasse basse, et à l'inverse, des sols recevant une fertilisation minérale produisent un rendement élevé. Cependant, ces deux cas (sans fertilisation et avec une fertilisation

minérale) présentent tous deux une CEC faible. Par ailleurs, les sols recevant du lisier se situent en haut et à droite de la figure, c'est à dire avec une forte production de biomasse et une CEC élevée, et s'opposent aux sols recevant du compost (points rouges). Nous supposons alors qu'il existe un lien entre la biomasse annuelle et la CEC.

Comme nous l'avons dit précédemment, nous avons recours à la transformation en racine carrée de la CEC. De plus, nous utilisons la transformation logarithmique de la biomasse annuelle, afin de stabiliser la variance et de linéariser la relation avec la CEC. Plusieurs transformation ont été comparées (racine carrée, inverse, etc.) mais, tout comme le rendement fourrager saisonnier, le logarithme a permis d'obtenir les meilleurs résultats. Ainsi, la modélisation de la Capacité d'Échange Cationique du sol peut s'écrire de la manière suivante :

$$\sqrt{(CEC)_{ijt}} = \mu + fert_i + site_j + f_{1j}(annee_t) + \beta \times \log(bio)_{ijt} + \eta_{ijt} \quad (3.7)$$

où  $\sqrt{(CEC)_{ijt}}$  est la racine carrée d'une mesure de la CEC sur le site  $j$  utilisant la fertilisation  $i$  lors de l'année  $t$ . Dans ce modèle, nous notons :

- $\mu$  la constante du modèle, représentant la CEC moyenne mesurée sur le site du LYCEE sans aucune fertilisation ( $i = 1$  et  $j = 1$ ),
- $fert_i$  l'effet de la fertilisation  $i$  sur la CEC, pour  $i = 1, \dots, 10$ ,
- $site_j$  l'effet du site  $j$ , pour  $j = 1, \dots, 4$ ,
- $f_{1j}(annee_t)$  la fonction de lissage associée au site  $j$  et à l'année  $t$ .
- $\log(bio)_{ijt}$  le logarithme de la biomasse produite lors de l'année  $t$  sur le site  $j$  avec la fertilisation  $i$ .
- $\eta_{ijt}$  les résidus de loi  $\mathcal{N}(0, \sigma^2)$ .

Comme précédemment, la fonction de lissage temporel  $f_{1j}$  est associée à l'année et au site  $j$ . Cela signifie que la CEC évolue de manière différente selon le type de site. Ainsi, pour un site donné, l'évolution de la CEC dans le temps est supposée être la même quelque soit la fertilisation utilisée, modulo l'effet fixe associé à cette fertilisation.

Tout comme pour le modèle de la teneur en Azote, les fonctions de régression partielle  $f_{1j}$  sont approchées par des splines cubiques de régression pénalisée, dont les noeuds sont associés aux différentes années de l'étude. L'estimation des coefficients fixes et aléatoires est aussi effectuée par l'approche REML.

Les données disponibles pour la CEC sont de nature chronologique, ce qui exclut l'hypothèse d'indépendance entre les différentes mesures. Afin de prendre en compte cette corrélation, les résidus  $\eta_{ijt}$  sont supposés être issus d'un processus autorégressif d'ordre 1 ( $AR(1)$ ), tels que :

$$\eta_{ijt} = \phi \eta_{ij(t-1)} + v_{ijt},$$

où  $\mathbb{E}(v_{ijt}) = 0$ ,  $\mathbb{V}(v_{ijt}) = \sigma_v^2$ ,  $cov(v_{ijt}, v_{ijt'}) = 0$ , et  $|\phi| < 1$ . La structure de corrélation associée est donc de la même forme que pour la teneur en Azote :

$$cor(\eta_{ijt}, \eta_{ijt'}) = \begin{cases} 1 & \text{si } t = t' \\ \phi^{|t-t'|} & \text{sinon.} \end{cases}$$

L'hypothèse d'homogénéité résiduelle est aussi remise en question. Graphiquement, des différences de variabilité des résidus (du modèle homogène) entre les sites sont constatées (fig. 3.12). La dispersion des résidus est relativement faible sur le site du LYCEE par rapport au site de SICALAIT. Le test d'homogénéité de Bartlett confirme nos suspicions, puisque l'hypothèse d'égalité des variances intra-site est rejetée avec une p-valeur inférieure à  $2.2 \times 10^{-16}$  (pour  $K^2 = 131.59$ ). Afin de tenir compte cette hétérogénéité, nous définissons la structure de variance suivante :

$$\eta_{ijt} \sim \mathcal{N}(0, \sigma_j^2), \text{ pour } j = 1, \dots, 4,$$

ce qui revient à poser  $\mathbb{V}(\eta_{ijt}) = \sigma_j^2$ , et ainsi permettre à chaque site d'avoir une variance résiduelle spécifique.

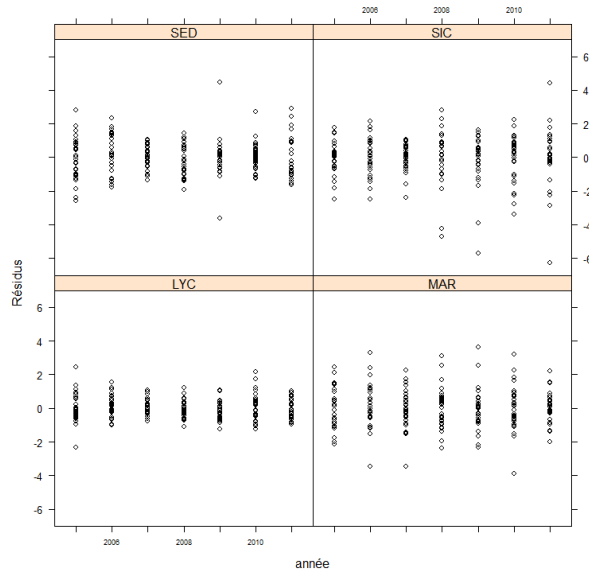


FIGURE 3.12 – Graphiques diagnostics des résidus du modèle homogène selon le site et l’année.

### Validation du modèle

L’inspection graphique des résidus du modèle additif (fig. 3.13 et 3.14), comportant une structure de variance propre au site, semble coïncider avec les hypothèses classiques du modèle linéaire. D’après le graphique 3.14, l’hétérogénéité des données semble mieux respectée.

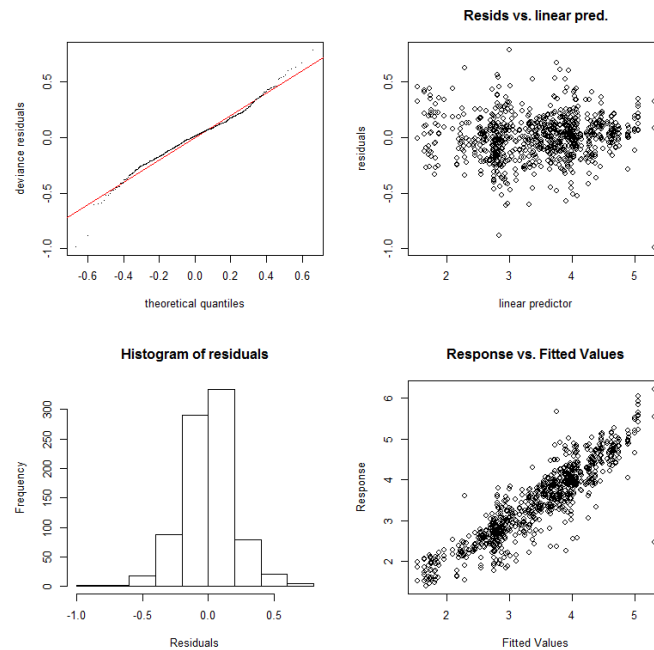


FIGURE 3.13 – Graphiques diagnostics des résidus du modèle hétérogène de la CEC.

### Résultats

Les différents paramètres du modèle sont significatifs au niveau 5% (tableau 3.11). Le tableau 3.12 présente les différents critères de comparaison et indicateurs d’écart du modèle construit. Les erreurs absolues relatives moyennes sont de 18.56% sur notre échantillon. Les estimations des composantes de la variance, ainsi que les facteurs multiplicatifs obtenus, sont présentés dans le tableau 3.13.

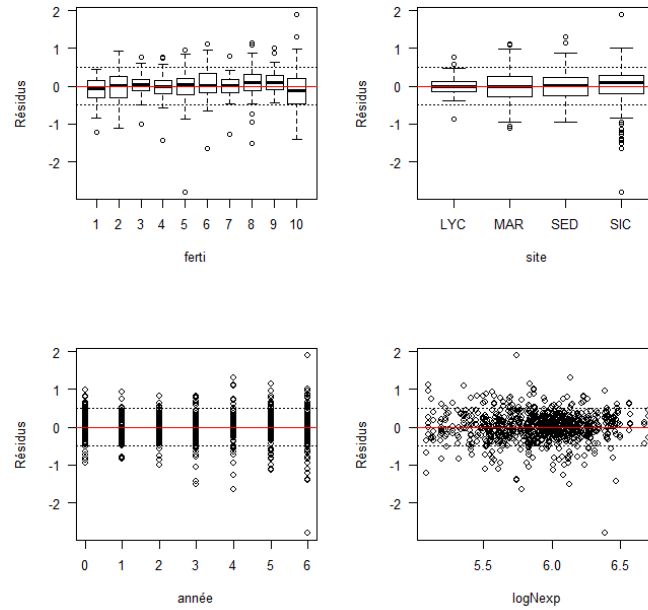


FIGURE 3.14 – Graphiques des résidus du modèle hétérogène de la CEC en fonction des facteurs et pour chaque observation.

	df/edf	F	p-valeur
ferti	9.00	73.07	0.00
site	3.00	249.21	0.00
log(bio)	1.00	6.10	0.01
$f_1(\text{annee})   \text{site.LYC}$	2.92	2.49	0.06
$f_1(\text{annee})   \text{site.MAR}$	5.33	8.90	0.00
$f_1(\text{annee})   \text{site.SED}$	4.89	5.14	0.00
$f_1(\text{annee})   \text{site.SIC}$	4.79	11.79	0.00

TABLE 3.11 – Résultats des tests de significativité des termes paramétriques et non-paramétriques du modèle additif de la CEC.

AIC	RMSE	MAPE (%)	$R_{a_j}^2$
623.45	2.86	18.56	0.855

TABLE 3.12 – Critères de comparaison et indicateurs d'écart du modèle additif de la CEC.

$site_j$	LYC (ref.)	MAR	SED	SIC
$\hat{\sigma}_j$	0.205	0.371	0.435	0.589
$\hat{\sigma}_j / \hat{\sigma}_1$	1	1.807	2.120	2.874

TABLE 3.13 – Estimations des écart-types et des facteurs multiplicatifs associés à chaque site.

Les estimations des écart-types des résidus concordent avec notre analyse graphique : la dispersion résiduelle sur le site du LYCEE est moins marquée que celle obtenue sur le site de SICALAIT.

Le coefficient de corrélation  $\phi$  est estimé par  $\hat{\phi} = 0.55$ . Ainsi, deux mesures successives de la CEC sur une même micro-parcelle sont relativement corrélées. Les effets partiels des quatres fonctions de lissage approchées sont représentés sur la figure 3.15.



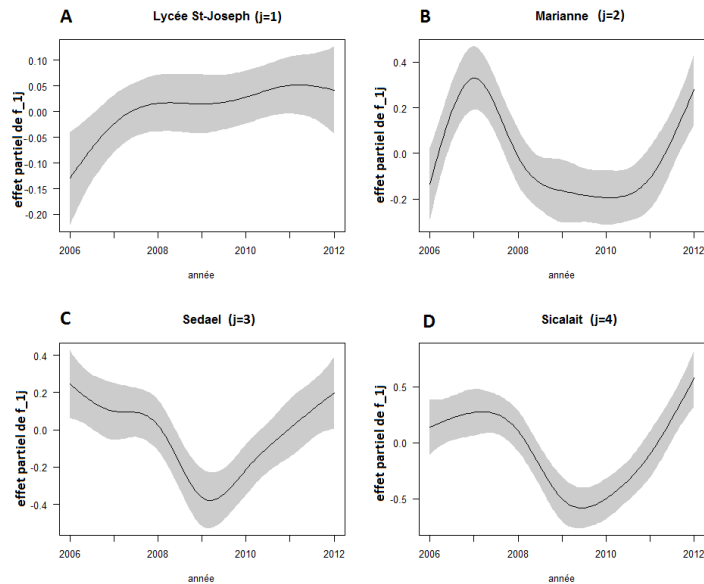


FIGURE 3.15 – Graphiques des quatre fonctions de régression partielles estimées sur la racine carrée de la CEC en fonction de l’année.

### Comparaisons des modèles

Comme précédemment, nous effectuons un test du rapport de vraisemblance afin de justifier de la pertinence d’une fonction de lissage temporel spécifique au site. Ainsi, la p-valeur du test obtenue est inférieure à 5% (tableau 3.14), estimer les quatre fonctions de lissage améliore donc le modèle.

Model	df	AIC	BIC	logLik	Test	L.Ratio	p-valeur
$f_1$	21	743.60	842.63	-350.80			
$f_1 _{site}$	27	623.45	750.67	-284.73	1 vs 2	132.15	< .0001

TABLE 3.14 – Comparaison du modèle avec une fonction de lissage temporel par site et du modèle avec une fonction de lissage commune.

Afin de vérifier la nécessité de l’hypothèse de dépendance entre les données, nous testons si le modèle ne considérant pas la corrélation est meilleur au sens de la log-vraisemblance.

Modèle	df	AIC	BIC	logLik	Test	L.Ratio	p-valeur
$\rho = 0$	26	820.50	943.00	-384.25			
$\rho \neq 0$	27	623.45	750.67	-284.73	1 vs 2	99.05	< .0001

TABLE 3.15 – Comparaison du modèle avec autocorrélation des résidus et du modèle supposant l’indépendance des observations.

Ainsi, d’après le test (tableau 3.15), le modèle construit, impliquant l’estimation d’un paramètre supplémentaire, améliore significativement les estimations au seuil 5%. La structure de corrélation temporelle semble adaptée à nos données.

En utilisant le test de la log-vraisemblance, nous le comparons ensuite au modèle considérant l’homogénéité des résidus. L’hypothèse nulle testée est :

$$H_0 : \sigma_1 = \dots = \sigma_4$$

et l’hypothèse alternative suppose qu’au moins deux composantes sont différentes.

Modèle	df	AIC	BIC	logLik	Test	L.Ratio	p-valeur
Homogénéité	24	802.53	915.61	-377.26			
Hétérogénéité	27	623.45	750.67	-284.73	1 vs 2	185.07	< .0001

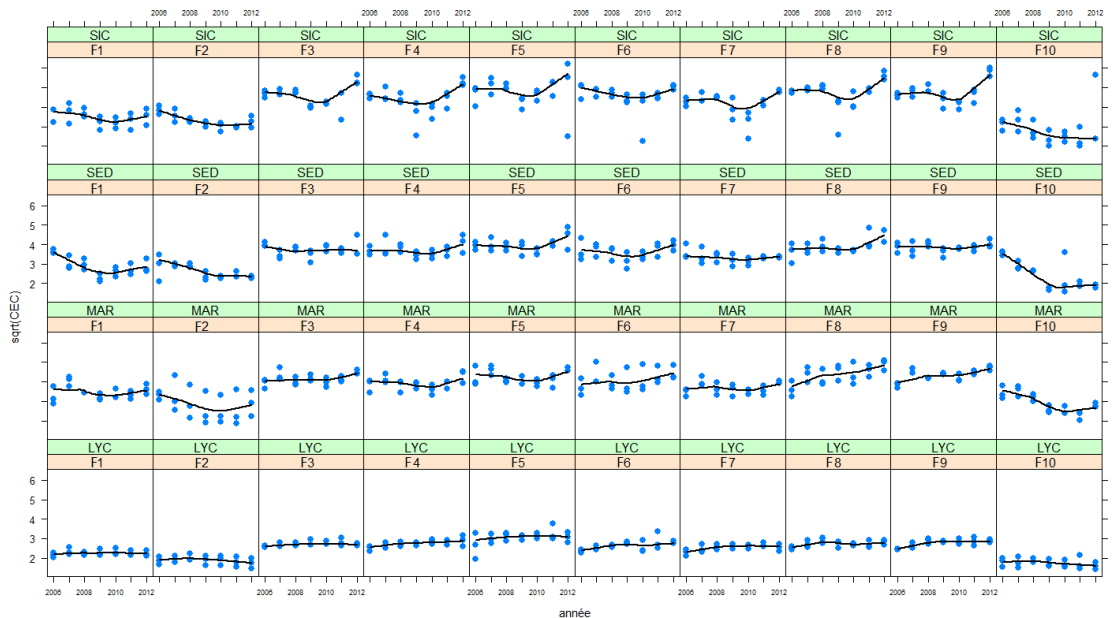
TABLE 3.16 – Comparaison du modèle hétérogène et du modèle homogène.

Au seuil fixé (5%), d'après le tableau 3.16, nous rejetons l'hypothèse nulle qui suppose l'égalité des variances intra-sites, et concluons que le modèle comportant une variance différente par site est plus adapté.

### 3.5.2 Un lissage temporel par fertilisation

#### Choix du modèle : seconde approche

L'hypothèse que nous avons faite sur l'effet partiel du temps peut être remise en question. En effet, pour certaines fertilisations, notamment minérales (traitement de fertilisation  $F2$  et  $F10$ ), nous observons un comportement de la CEC différent dans le temps, comme le montre le graphique 3.16.


 FIGURE 3.16 – Graphiques des observations de la CEC (en  $meq/100g$ ) dans le temps selon le site et la fertilisation.

Il semble plus adapté de distinguer l'évolution de la CEC selon la fertilisation, et non pas selon le site. Comme le montre le graphique 3.16, les formes des courbes lissées par fertilisation sont assez similaires (modulo l'effet fixe associé au type de site). Cette hypothèse est préférable, puisque l'effet du type de sol sur la CEC est fixe, alors que la fertilisation aura des effets plus au moins retardés, visibles à moyen et long terme. Ainsi, nous proposons un second modèle expliquant le comportement de la CEC :

$$\sqrt{(CEC)_{ijt}} = \mu + fert_i + site_j + f_{2i}(annee_t) + \eta_{ijt} \quad (3.8)$$

où  $\sqrt{(CEC)_{ijt}}$  est la racine carrée d'une mesure de la CEC sur le site  $j$  utilisant la fertilisation  $i$  lors de l'année  $t$ . Le paramètre  $fert_i$  est l'effet de la fertilisation  $i$ , pour  $i = 1, \dots, 10$ , et  $site_j$  est l'effet du site  $j$ , pour  $j = 1, \dots, 4$ .

Cette fois, la fonction de lissage temporel  $f_{2i}$  est associée à la fertilisation  $i$ . Ainsi, pour une fertilisation donnée, la variation de la CEC dans le temps est supposée être la même quelque

soit le type de site. Suite à cette nouvelle hypothèse, la biomasse exportée annuellement n'est plus significative au niveau 5%, la variable  $\log(bio)$  est donc exclue du modèle. Nous conservons la même approche lors de l'ajustement des fonctions de lissage (splines cubiques de régression pénalisée / REML).

De même, les résidus  $\eta_{ijt}$  sont supposés être issus d'un processus  $AR(1)$ , tels que :

$$\eta_{ijt} = \phi\eta_{ij(t-1)} + v_{ijt},$$

où  $\mathbb{E}(v_{ijt}) = 0$ ,  $\mathbb{V}(v_{ijt}) = \sigma_v^2$ ,  $cov(v_{ijt}, v_{ijt'}) = 0$ , et  $|\phi| < 1$ .

L'hypothèse d'hétérogénéité intra-site est encore valable ici : le graphique des résidus du modèle homogène est similaire à celui de la figure 3.12, et le test d'homogénéité de la variance de Bartlett est encore très significatif à 5%. Comme précédemment, la structure de variance proposée est

$$\eta_{ijt} \sim \mathcal{N}(0, \sigma_j^2), \text{ pour } j = 1, \dots, 4,$$

ce qui revient à poser  $\mathbb{V}(\eta_{ijt}) = \sigma_j^2$ , et ainsi permettre à chaque site d'avoir une variance résiduelle spécifique.

### Validation du modèle

L'hypothèse de normalité des résidus semble vérifiée, comme en témoignent les diagnostics graphiques de la figure 3.17. De plus, la variabilité des observations selon les sites est respectée.

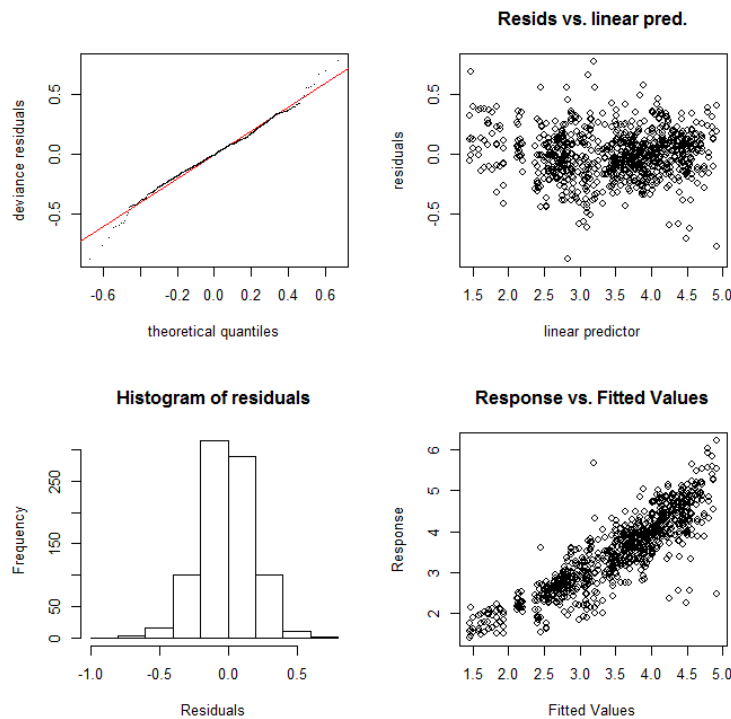


FIGURE 3.17 – Graphiques des résidus du modèle de la CEC selon l'année.

### Résultats

Au niveau des paramètres du modèle, plusieurs remarques peuvent être établies sur les résultats des estimations (tableau 3.17). Tout d'abord, les effets principaux sont encore très significatifs, mais les approximations des fonctions de lissage ne le sont pas toutes. Ainsi, le témoin, la fertilisation F3 et F4 (lisier et lisier+minéral), ainsi que la fertilisation F7 (compost+minéral) n'ont pas d'effet significatif (au seuil de 5%) sur l'évolution de la CEC dans le temps.

La seconde constatation concerne les degrés de liberté estimés. Pour quatre fertilisations (F2,

	df/edf	F	p-valeur
ferti	9.00	78.00	0.00
site	3.00	458.31	0.00
$f_2(\text{annee})$   <b>ferti.1</b>	1.00	0.29	<b>0.59</b>
$f_2(\text{annee})$   ferti.2	1.00	4.33	0.04
$f_2(\text{annee})$   <b>ferti.3</b>	1.00	0.98	<b>0.32</b>
$f_2(\text{annee})$   <b>ferti.4</b>	2.80	2.13	<b>0.10</b>
$f_2(\text{annee})$   ferti.5	1.00	5.13	0.02
$f_2(\text{annee})$   ferti.6	1.00	6.18	0.01
$f_2(\text{annee})$   <b>ferti.7</b>	1.00	1.52	<b>0.22</b>
$f_2(\text{annee})$   ferti.8	3.29	4.76	0.00
$f_2(\text{annee})$   ferti.9	1.00	7.53	0.01
$f_2(\text{annee})$   ferti.10	2.00	6.18	0.00

TABLE 3.17 – Résultats des tests de significativité des termes paramétriques et non-paramétriques du modèle additif de la CEC.

F5, F6 et F9), la relation entre la CEC et le temps est linéaire (et significative). En effet, un seul degré de liberté est obtenu, ce qui signifie que l'ajustement a impliqué l'estimation d'un seul coefficient. Seules deux fertilisations ont une relation non-linéaire entre la CEC et l'année : la fertilisation F10 (minérale) et la fertilisation F8 (compost). Le graphique 3.18 présente les différents effets partiels de ces fonctions estimées. Les linéarités et non-linéarités identifiées précédemment sont clairement visibles ici.

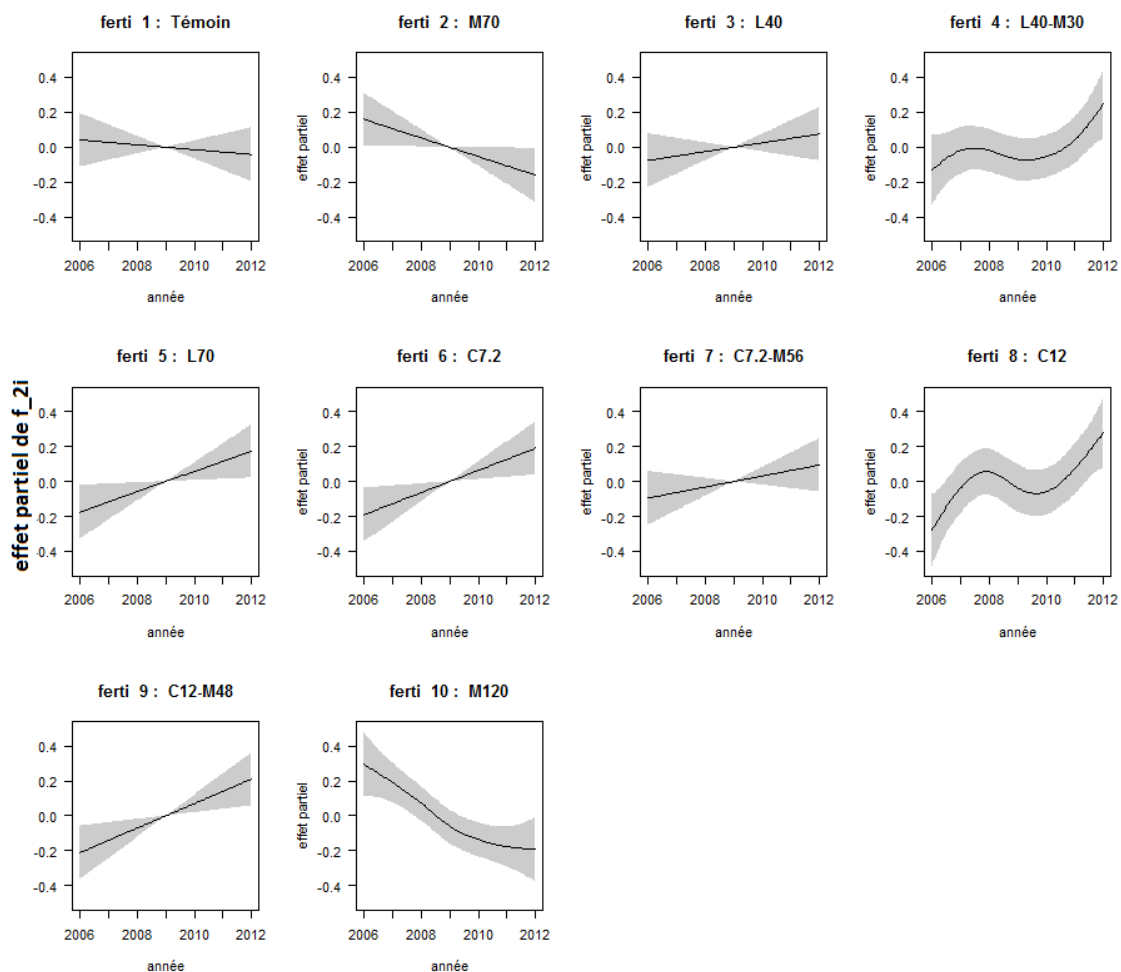


FIGURE 3.18 – Graphiques des dix fonctions de régression partielles estimées sur la racine carrée de la CEC en fonction de l'année.

Le tableau 3.18 présente les différents critères de comparaison et indicateurs d'écart du modèle construit.

AIC	RMSE	MAPE (%)	$R_{aj}^2$
771.86	3.27	20.89	0.833

TABLE 3.18 – Critères de comparaison et indicateurs d'écart du modèle additif de la CEC.

Prendre en compte l'hétérogénéité intra-site fait chuter l'AIC de 937.45 à 771.86. Les erreurs absolues relatives moyennes sont ici de 20.89% sur notre échantillon. Les estimations des variances résiduelles sont présentées dans le tableau 3.19.

$site_j$	LYC (ref.)	MAR	SED	SIC
$\hat{\sigma}_j$	0.207	0.411	0.424	0.657
$\hat{\sigma}_j/\hat{\sigma}_1$	1	1.989	2.052	3.183

TABLE 3.19 – Estimations des écart-types et des facteurs multiplicatifs associés à chaque site.

Les estimations attestent la variabilité observée précédemment : le site du LYCEE présente une dispersion résiduelle inférieure à celle de SICALAIT ( $\hat{\sigma}_1 = 0.207$  et  $\hat{\sigma}_4 = 0.657$ ). La dépendance des observations est encore une fois confirmée par un fort coefficient de corrélation estimé :  $\hat{\phi} = 0.472$ .

### 3.5.3 Interprétation agronomique des résultats

D'après l'interprétation des résultats des premiers modèles statistiques et de l'analyse exploratoire, il nous semble plus adapté de distinguer l'évolution de la CEC selon la modalité de fertilisation (organique, minérale et/ou mixte), et non pas selon le site expérimental. Comme le montre les graphiques de la figure 3.16, les formes des courbes lissées par modalité de fertilisation sont assez similaires. Le modèle retenu suppose donc que certains effets de la fertilisation sur la CEC varient au cours du temps.

Si nous considérons les graphiques des fonctions de régression partielles estimées pour chaque modalité de fertilisation (Figure 3.18), nous constatons que les effets des modalités de fertilisation F1 (témoin), F3 (lisier à faible niveau), F4 (lisier à faible niveau + engrais minéral) et F7 (compost à faible niveau + engrais minéral) ne varient pas dans le temps, c'est-à-dire que leurs effets sur la CEC ne sont pas visibles lors des sept années d'étude, et cela pour tous les sites. Les effets des autres modalités de fertilisation sur la CEC varient dans le temps, pour tous les sites. Pour les modalités de fertilisation minérale F2 (engrais minéral à faible niveau) et F10 (engrais minéral à niveau élevé) nous observons une tendance décroissante (linéaire et quasi-linéaire, respectivement). Pour les modalités de fertilisation organique F5 (lisier à niveau élevé) et F6 (compost à faible niveau), ainsi que pour la modalité de fertilisation mixte F9 (compost à niveau élevé + engrais minéral) nous observons plutôt une tendance linéaire croissante de la CEC au cours du temps. Dans le cas de la modalité F8 (compost à niveau élevé) nous observons un effet variable sur la CEC ; une augmentation lors des deux premières années, la stabilisation de la CEC pendant les deux années suivantes, et une augmentation de la CEC lors des deux dernières années de l'étude. Nous confirmons ainsi que l'utilisation d'engrais organiques (et mixtes) joue un rôle important sur l'augmentation de la capacité du sol à retenir et échanger les éléments nutritifs facilement disponibles pour les plantes. Au contraire, l'utilisation d'engrais minéraux engendre une diminution de cette capacité d'échange des sols qui aura comme conséquence une diminution (plus ou moins significative) des rendements fourragers des parcelles.

Les estimations des effets fixes, ainsi que leurs significativités, obtenues avec le modèle tenant compte d'une variabilité de l'effet du site dans le temps, sont sensiblement les mêmes (tableau G.7). Ainsi, nous préférons utiliser le modèle supposant une variabilité des effets des fertilisations dans le temps pour mieux identifier l'impact sur le long-terme des modalités de fertilisation.

L'analyse de la significativité des effets fixes est effectuée au seuil 5%. Comme prévu, nous observons que les sols des sites de SICALAIT, de MARIANNE et de SEDAEL (andosols, riches en matière organique et des capacités d'échange anionique et cationique élevées) présentent des CEC moyennes supérieures à celle du site du LYCEE (sol sableux, pauvre en matière organique et en CEC). Par ailleurs, la CEC moyenne observée sur le site de SICALAIT est significativement supérieure à celle des autres sites. En ce qui concerne les effets des modalités de fertilisation sur la CEC des sols, nous observons que sur tous les sites, toutes les modalités de fertilisation conduisent à une CEC significativement différente de celle de la modalité témoin (sans apports de fertilisants) : les modalités de fertilisation minérale (F2 et F10) ont un effet négatif sur la CEC par rapport au témoin, tandis que les modalités de fertilisation organique et mixte conduisent à une augmentation de la CEC des sols. Parmi toutes les modalités de fertilisation organique, la modalité F5 (lisier à niveau élevé) présente l'influence positive la plus élevée (et significative) sur la CEC sur tous les sites, suivie de F8 (compost à niveau élevé) et F9 (compost à niveau élevé + engrais minéral). Les effets des modalités F8 et F9 sont semblables et ne diffèrent pas significativement de ceux de F3 (lisier à faible niveau), F4 (lisier à faible niveau + engrais minéral) et F6 (compost à faible niveau).

### 3.6 Conclusion

Dans cette partie, la stratégie de modélisation a consisté tout d'abord à identifier la non-linéarité (temporelle) de la variable à expliquer, graphiquement puis à l'aide d'un test du manque d'ajustement. Le choix du type de fonction de lissage s'est basé essentiellement sur l'analyse graphique des données et sur des tests de comparaison. La détection et la prise en compte d'éventuelles corrélations et spécificités des variances intra-groupes a ensuite permis d'améliorer la modélisation et d'assurer la validité des estimations et donc des tests.

La modélisation de la teneur en Azote nous conduit à des résultats assez satisfaisants : les erreurs absolues relatives moyennes (MAPE) sont de 8.14% sur notre échantillon, et la correction de la structure de variance-covariance fait chuter le critère d'AIC. Concernant la CEC, le critère MAPE est plus élevé (20.89), ce qui nous indique une moins bonne qualité d'ajustement du modèle, mais le  $R_{aj}^2$  reste assez élevé et les contraintes imposées à la structure de variance-covariance permettent tout de même d'abaisser le critère d'AIC.

# Conclusion générale

L'étude réalisée dans ce rapport s'articule autour de l'effet à moyen et long terme de l'efficacité des fertilisations organiques et minérales sur le rendement fourrager et sur la qualité des sols à la Réunion. Les résultats obtenus reposent sur des modèles respectant les contraintes imposées par les données : l'hétérogénéité, la corrélation et la non-linéarité.

L'approche initialement envisagée était celle des "panel data", qui est essentiellement appliquée en économétrie pour exploiter la double dimension des données (temporelle et individuelle). Cette méthode est de type "dynamique" puisqu'elle permet d'estimer des variables dans plusieurs dimensions, à partir de données nombreuses et variables. Cependant, l'estimateur intra-individuel (within), néglige une grande part de l'information contenue dans l'échantillon, et il est surtout très sensible aux erreurs de spécifications, telles que l'omission de variables (qui peut être ici la pluviométrie ou encore le rayonnement solaire). De plus, les points aberrants perturbent gravement les estimations, même avec une grande quantité d'informations disponibles. La recherche d'erreurs de mesures dans notre base de données n'était pas envisageable du fait de l'importante variabilité des quantités étudiées. Enfin, l'aspect non-linéaire des variables conduit à des difficultés méthodologiques et numériques lors de la spécification des erreurs par cette méthode.

Deux approches alternatives ont alors été explorées pour construire ces modèles. La première se base sur une structure de variance-covariance, et la seconde est fondée sur les modèles additifs et les fonctions splines. La théorie des fonctions splines de régression pénalisée est assez complexe, mais les résultats sont facilement interprétables grâce à l'outil graphique du package *mgcv* du logiciel R (S. Wood, [15]).

Le modèle retenu pour le rendement fourrager comporte les effets fixes de la fertilisation, du site et de la saison. Une structure de corrélation associée aux répétitions saisonnières de type AR(1) a été retenue, ainsi qu'une variance différente selon la saison et le site, afin de palier au problème d'hétérogénéité des observations.

Concernant l'analyse de la qualité du sol en terme d'Azote, la dimension temporelle s'est avérée non-linéaire et différente selon le type de site. Nous avons aussi identifié les effets fixes des différentes modalités de fertilisation, ceux du site ainsi que les effets croisés de ces deux facteurs. Deux possibilités se sont ensuite offertes à nous lors de la modélisation de la CEC. En effet, la relation non-linéaire entre les années et la CEC pouvait s'exprimer soit par le type de site, soit par le type de fertilisation. Notre choix s'est basé à la fois sur la volonté d'expliquer la variabilité de la CEC par l'effet des fertilisations et sur l'hypothèse agronomique suivante : la CEC dépend peu du climat mais principalement de la composition chimique du sol. Ainsi, l'effet du site est fixe, ce qui nous a permis d'expliquer le comportement de la CEC dans le temps par la fertilisation utilisée.

Comme pour les données de rendement fourrager, nous étions en présence de corrélation parmi les observations de la teneur en Azote et de la CEC, et là encore, une dépendance de type AR(1) a été nécessaire afin d'assurer la validité du modèle. De plus, la variabilité observée n'était pas la même d'un site à un autre, ce qui nous a conduit à pondérer les observations suivant le site auquel elles appartenaient.

Les modèles ainsi construits nous ont permis de comparer les différentes fertilisations dans de bonnes conditions. Ainsi, au cours de l'année, une large part des fertilisations ont eu des effets significatifs sur les rendements fourragers par rapport à l'absence de fertilisation. Sur trois des sites expérimentaux (SICALAIT, SEDAEL et MARIANNE), de type andosol, les fertilisations de type lisier semblent pouvoir se substituer totalement aux apports d'engrais minéral.

---

Les résultats obtenus sur le site de type sableux (LYCEE) avec des fertilisations constituées de compost n'ont pas montré de différence significative par rapport aux engrais minéraux qui présentent les rendements les plus élevés.

L'analyse de la teneur en Azote a aussi mis en évidence la supériorité des engrais organiques de type lisier, et surtout, l'impact plutôt faible voire négatif des engrais minéraux sur la présence cet élément dans le sol. La dernière analyse, celle de la CEC, confirme que l'utilisation d'engrais organiques (et mixtes) joue un rôle important sur l'augmentation de la capacité du sol à retenir et échanger les éléments nutritifs facilement disponibles pour les plantes. Au contraire, l'utilisation d'engrais minéraux diminue cette capacité d'échange des sols. Á plus long terme, ce phénomène pourrait alors impacter négativement les rendements fourragers des parcelles.

Ainsi, nous avons identifié les avantages et potentialités de certaines fertilisations, à la fois sur le rendement fourrager et sur la composition du sol, mais dans des modèles distincts. Il serait alors intéressant d'identifier ce lien dans un modèle global.

Une autre amélioration possible serait de prendre en compte d'autres variables, comme par exemple les espèces (tropicales et/ou tempérées) présentes sur les parcelles étudiées, dont les données sont disponibles sous forme de pourcentage de recouvrement.

Suite à ces analyses, certaines recommandations et conseils vont être effectués aux professionnels agricoles. Ainsi, l'élaboration future d'un référentiel fiable sur la fertilisation raisonnée des prairies à l'échelle de la Réunion, c'est-à-dire dans des écosystèmes contrastés, va pouvoir être envisagée. Ces pratiques pourront être résumées dans des fiches techniques mises à disposition des éleveurs.

Les méthodes actuelles vont ensuite pouvoir progresser vers une intensification de la production de biomasse qui soit plus écologique, en recyclant au mieux les matières organiques produites par les animaux, tout en améliorant la fertilité du sol. Loin d'être un déchet dont les éleveurs "se débarrassent", les effluents d'élevage ont une valeur fertilisante réelle qui permet de diminuer fortement l'achat d'engrais minéraux. Au niveau de la Région Réunion, la valorisation des effluents d'élevage peut contribuer à la diminution de la dépendance vis-à-vis de l'importation des intrants pour l'agriculture et pour l'élevage.



# Bibliographie

- [1] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *Econometrica*, 17(2) :453–510, 1989.
- [2] P. Chabalier, V. Van De Kerchove, and H. Saint Macary. *Guide de la fertilisation organique à La Réunion*. CIRAD, 2006.
- [3] M. Davidian and D. Giltinan. *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, 1995.
- [4] P. Diggle, P. Heagerty, K. Liang, and S. Zeger. *Analysis of Longitudinal Data*, volume 2. Oxford University Press, 2002.
- [5] X. Fang and K. Chan. Generalized additive models with spatio-temporal data. 1989.
- [6] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.
- [7] R. McLauren and K. Cameron. *Soil Science : sustainable production and environmental protection*, volume 2. Oxford University Press, Auckland, 1996.
- [8] J. Neill. Testing for lack of fit in nonlinear regression. 16(2) :733–740, 1988.
- [9] B. Normand. *Etude expérimentale et modélisation du devenir de l’azote dans le système sol-plante-atmosphère*. Université Joseph Fourier-Grenoble 1, 1996.
- [10] J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.
- [11] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme : Linear and Nonlinear Mixed Effects Models*, 2013. R package version 3.1-109.
- [12] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [13] D. Wang and M. Conerly. Evaluation of three lack of fit tests in linear regression models. 30(6) :683–696, 2003.
- [14] D. Whitehead. *Grassland Nitrogen*. CAB International, Wallingford, 1995.
- [15] S. Wood. *Generalized Additive Models : an introduction with R*. Chapman and Hall, 2006.
- [16] A. Zuur, E. Ieno, N. Walker, A. Saveliev, and G. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Springer, 2009.

# Table des figures

1.1	Visualisation de l'emplacement géographique de chaque site sur la carte de l'île de la Réunion. . . . .	3
1.2	Représentation schématique du plan expérimental en bloc complet pour un site. . . . .	4
2.1	Graphique des interactions entre l'année et la fertilisation sur le rendement $Y$ (en $kg\ MS/ha/coupe$ ). . . . .	10
2.2	Graphique des interactions entre la fertilisation et la saison sur le rendement $Y$ (en $kg\ MS/ha/coupe$ ). . . . .	11
2.3	Fonction d'autocorrélation empirique des résidus standardisés du modèle (2.5). . . . .	12
2.4	Représentation de la matrice des corrélations estimées sur le logarithme du rendement $Y$ pour chaque couple de saisons. . . . .	13
2.5	Fonction d'autocorrélation empirique des résidus normalisés du modèle 2.8 comportant la structure de variance-covariance 2.9 . . . . .	15
2.6	Graphique des interactions a) entre la fertilisation et la site, b) entre la saison et le site, sur le rendement (en $kg\ MS/ha/coupe$ ). . . . .	17
2.7	Représentation en boîte à moustaches de la dispersion des observations du logarithme du rendement $Y$ par site et par saison. . . . .	18
2.8	Fonction d'autocorrélation empirique des résidus standardisés du modèle (2.10). . . . .	18
2.9	Représentation de la matrice de corrélations estimées sur le $\log(\text{rendement})$ . . . . .	19
2.10	Fonction d'autocorrélation empirique des résidus normalisés du modèle 2.12 comportant la structure de variance-covariance 2.13 . . . . .	20
2.11	Estimations du rendement en $kg\ MS/ha/coupe$ par le modèle retenu pour chaque combinaison de modalités des facteurs. . . . .	21
3.1	Graphiques des observations de la teneur en Azote (en $g\ N/kg\ sol$ ) dans le temps selon le site. . . . .	25
3.2	Graphiques des observations de la CEC (en $meq/100\ g\ sol$ ) dans le temps selon le site. . . . .	25
3.3	Graphiques des observations de la teneur en Azote dans le temps selon le site et la fertilisation et courbe lissée de type <i>loess</i> associée. . . . .	29
3.4	Graphiques des interactions entre le site et la fertilisation sur la teneur en Azote. . . . .	29
3.5	Graphiques diagnostics des résidus du modèle de la teneur en Azote. . . . .	31
3.6	Graphiques diagnostics des résidus du modèle de la teneur en Azote en fonction des facteurs et pour chaque observation. . . . .	31
3.7	Graphiques diagnostics des résidus du modèle. . . . .	34
3.8	Graphiques des résidus du modèle 3.6 selon le site et l'année. . . . .	34
3.9	Graphiques des quatre fonctions de régression partielles estimées sur le logarithme de la teneur en Azote. . . . .	36
3.10	Graphiques des interactions entre le site et la fertilisation sur la CEC (en $meq/100g$ ). . . . .	39
3.11	Graphiques de la CEC calculée (en $meq/100g$ ) sol en fonction de la biomasse annuelle (en $kg\ MS/ha/coupe/an$ ) selon le site. . . . .	39
3.12	Graphiques diagnostics des résidus du modèle homogène selon le site et l'année. . . . .	41
3.13	Graphiques diagnostics des résidus du modèle hétérogène de la CEC. . . . .	41
3.14	Graphiques des résidus du modèle hétérogène de la CEC en fonction des facteurs et pour chaque observation. . . . .	42
3.15	Graphiques des quatre fonctions de régression partielles estimées sur la racine carrée de la CEC en fonction de l'année. . . . .	43
3.16	Graphiques des observations de la CEC (en $meq/100g$ ) dans le temps selon le site et la fertilisation. . . . .	44

3.17	Graphiques des résidus du modèle de la CEC selon l'année. . . . .	45
3.18	Graphiques des dix fonctions de régression partielles estimées sur la racine carrée de la CEC en fonction de l'année. . . . .	46
B.1	Le cycle de l'Azote dans un système agricole (d'après McLauren et Cameron [7]).	59
C.1	Représentation en boîte à moustache de la dispersion des durées selon le site. . .	60
D.1	Diagramme en bâtons du rendement moyen par coupe et par fertilisation, en kilogramme de matière sèche par hectare. . . . .	62
D.2	Représentation de la matrice des corrélations estimées sur les observations du rendement pour chaque couple d'années. . . . .	63
D.3	Histogrammes et boxplots de la variable à expliquer du rendement $Y$ et de sa transformation logarithmique. . . . .	64
D.4	Graphiques des quantiles théoriques et estimés du rendement $Y$ et de sa transformation logarithmique. . . . .	64
D.5	Représentation en boîte à moustaches de la dispersion des observations du logarithme du rendement $Y$ par fertilisation. . . . .	65
D.6	Représentation en boîte à moustaches de la dispersion des observations du logarithme du rendement $Y$ par site et par fertilisation. . . . .	65
D.7	Représentation en boîte à moustache de la dispersion des observations du $\log(\text{rendement})$ selon la saison. . . . .	65
E.1	Illustration d'une fonction de base $b_4(x)$ pour une spline de régression cubique (à gauche) et d'une combinaison de plusieurs fonctions de base lors du lissage d'une courbe (à droite). . . . .	67
F.1	Diagramme en bâtons de la teneur en Azote moyenne obtenue annuellement par fertilisation, en gramme d'Azote par kilogramme d'échantillon de sol sec (g N/kg sol). . . . .	70
F.2	Graphiques des observations du taux d'Azote dans le temps (en g N/100g sol). .	70
F.3	Diagramme en bâtons de la CEC moyenne obtenue annuellement par fertilisation, en milligrammes d'équivalents par 100 grammes d'échantillon de sol sec (meq/100g sol). . . . .	71
F.4	Graphiques des observations de la CEC (en meq/100g sol) dans le temps. . . .	71
G.1	Estimations de la teneur en Azote (en g N/kg sol) par le modèle retenu pour chaque combinaison de modalités des facteurs sur les 7 premières années. . . . .	75
G.2	Estimations de la CEC (en meq/100g sol) par le modèle retenu pour chaque combinaison de modalités des facteurs sur les 7 premières années. . . . .	76

# Liste des tableaux

1.1	Descriptif de la diversité agro-écologique des sites expérimentaux. . . . .	3
1.2	Descriptif des fertilisations minérales et/ou organiques utilisées. . . . .	4
2.1	Résultats des tests de l'ANOVA pour le modèle 2.5. . . . .	11
2.2	Comparaison des différentes structures de variance-covariance. . . . .	14
2.3	Comparaison des modèles retenus. . . . .	15
2.4	Estimations des écart-types et des facteurs multiplicatifs associés à chaque saison. . . . .	16
2.5	Résultats des tests de l'ANOVA pour le modèle 2.10. . . . .	16
2.6	Comparaison des modèles retenus. . . . .	20
2.7	Estimations des écart-types et des facteurs multiplicatifs associés à chaque saison et à chaque site. . . . .	21
3.1	Résultats des tests de significativité des termes paramétriques et non-paramétriques du modèle additif de la teneur en Azote. . . . .	32
3.2	Critères de comparaison et indicateurs d'écart du modèle additif de la teneur en Azote . . . . .	32
3.3	Estimations des composantes de la variance associés aux paramètres de lissage du modèle additif. . . . .	32
3.4	Comparaison du modèle avec une fonction de lissage temporel par site et du modèle avec une fonction de lissage commune. . . . .	32
3.5	Comparaison du modèle avec autocorrélation des résidus et du modèle supposant l'indépendance des observations. . . . .	33
3.6	Résultats des tests de significativité des termes paramétriques et non-paramétriques du modèle additif de la teneur en Azote. . . . .	35
3.7	Critères de comparaison et indicateurs d'écart du modèle additif de la teneur en Azote. . . . .	35
3.8	Estimations des écart-types et des facteurs multiplicatifs associés à chaque site. . . . .	35
3.9	Estimations des composantes de la variance associés aux paramètres de lissage du modèle additif. . . . .	35
3.10	Comparaison du modèle hétérogène et du modèle homogène. . . . .	36
3.11	Résultats des tests de significativité des termes paramétriques et non-paramétriques du modèle additif de la CEC. . . . .	42
3.12	Critères de comparaison et indicateurs d'écart du modèle additif de la CEC. . . . .	42
3.13	Estimations des écart-types et des facteurs multiplicatifs associés à chaque site. . . . .	42
3.14	Comparaison du modèle avec une fonction de lissage temporel par site et du modèle avec une fonction de lissage commune. . . . .	43
3.15	Comparaison du modèle avec autocorrélation des résidus et du modèle supposant l'indépendance des observations. . . . .	43
3.16	Comparaison du modèle hétérogène et du modèle homogène. . . . .	44
3.17	Résultats des tests de significativité des termes paramétriques et non-paramétriques du modèle additif de la CEC. . . . .	46
3.18	Critères de comparaison et indicateurs d'écart du modèle additif de la CEC. . . . .	47
3.19	Estimations des écart-types et des facteurs multiplicatifs associés à chaque site. . . . .	47
C.1	Extrait de la base de données construite. . . . .	60
D.1	Extrait des données disponibles. . . . .	62
G.1	Tests et significativité des effets principaux du modèle du rendement fourrager (le site du LYCEE, la fertilisation F1 et la saison S1 sont les modalités de référence). . . . .	72

G.2	Estimations du rendement moyen par coupe et par saison (en <i>kg</i> MS/ha/coupe) sur le site du LYCEE. . . . .	72
G.3	Estimations du rendement moyen par coupe et par saison (en <i>kg</i> MS/ha/coupe) sur le site de SICALAIT. . . . .	73
G.4	Estimations du rendement moyen par coupe et par saison (en <i>kg</i> MS/ha/coupe) sur le site de MARIANNE. . . . .	73
G.5	Estimations du rendement moyen par coupe et par saison (en <i>kg</i> MS/ha/coupe) sur le site de SEDAEL. . . . .	73
G.6	Estimations des effets principaux et des termes d'interaction du modèle additif hétérogène. . . . .	74
G.7	Estimations des termes paramétriques du modèle additif . . . . .	76



# Annexes

# Annexe A

## Présentation de la structure d'accueil

Créé en 1984, le centre de Coopération Internationale en Recherche Agronomique pour le Développement (Cirad) est un établissement public à caractère industriel et commercial. Deux ministères gèrent ses activités : le ministère de l'Enseignement supérieur et de la Recherche ainsi que le ministère des Affaires étrangères et européennes. Les enjeux internationaux de l'agronomie ainsi que le développement agricole sont les principaux défis de ce centre de recherche, qui dispose d'un réseau mondial de partenaires.

Le second dispositif du Cirad en France se trouve à l'île de la Réunion. Ce pôle concentre ses recherches sur le développement des filières agricoles et agroalimentaires dans un contexte de préservation de l'environnement et de la biodiversité. Trois dispositifs se partagent les activités de ce centre, répartis entre l'île de la Réunion et Mayotte. Le Cirad regroupe des domaines de recherche très variés. Ainsi, les différentes missions peuvent concerner la production végétale, l'élevage, la qualité et la sécurité des aliments, ou encore l'impact et les services agro-environnementaux.

Parmi les différents programmes de recherche du Cirad, l'unité de recherche Systèmes d'Élevage Méditerranéens et Tropicaux dans laquelle je travaille, est en charge du programme CIEEL (Conduite Intégrée des Exploitations et des filières d'Élevage). Ce programme a pour objectif de mettre à la disposition des éleveurs, des filières d'élevage et des prestataires du développement, des outils d'aide à la décision finalisés dans une optique de développement durable. CIEEL intègre un projet issu du Ministère de l'Outre-mer, visant à évaluer et modéliser l'efficacité technique, économique et environnementale de systèmes fourragers intégrant agriculture et élevage laitier à la Réunion et à Mayotte.

L'unité mixte de recherche Systèmes d'Élevage Méditerranéens et Tropicaux regroupe des agents de trois structures : l'INRA (le département Physiologie animale et systèmes d'élevage et le département Sciences pour l'Action et le Développement), le SupAgro de Montpellier (le département Sciences pour l'Environnement, l'Élevage et l'Agriculture) et le CIRAD (le département Environnements et Sociétés et département Performances des Systèmes de Production et de Transformation Tropicaux). Cette UMR s'intéresse aux agro-écosystèmes d'élevage en milieux chauds, favorables ou difficiles.

Une cinquantaine de cadres scientifiques et une trentaine de doctorants composent l'unité. Près d'un tiers d'entre eux sont expatriés en régions tropicales. L'équipe localisée à la Réunion se compose de quatre chercheurs en sciences vétérinaires et en zootechnie, un doctorant et trois techniciens.



# Annexe B

## Le cycle de l'Azote

Cette partie des annexes permet d'expliquer brièvement le cycle de l'Azote (Figure B.1) et à définir les différents paramètres qui contrôlent les transformations.

L'Azote est un élément essentiel pour la croissance et le fonctionnement de tous les organismes dans le sol. Les sols contiennent cependant peu d'Azote, en moyenne 0.1 – 0.6% dans les 15 premiers cm, et la source principale d'Azote, sous forme de gaz inerte  $N_2$ , est l'atmosphère. La plupart des végétaux ne peuvent assimiler l'Azote que sous forme soluble minérale, principalement sous forme nitrique ( $NO_3^-$ ) et éventuellement sous forme ammoniacale ( $NH_4^+$ ). Ces deux formes constituent l'essentiel de la fraction inorganique qui représente généralement moins de 5% de l'Azote du sol. Le reste est présent sous forme organique comme les racines, la faune, la microflore bactérienne et fongique, la litière et les composés humiques. Le cycle de l'Azote ainsi que les principaux processus mis en jeu sont schématisés sur le diagramme de la Figure B.1. On pourra se référer à la littérature pour plus de détails (Whitehead [14], Normand, [9]).

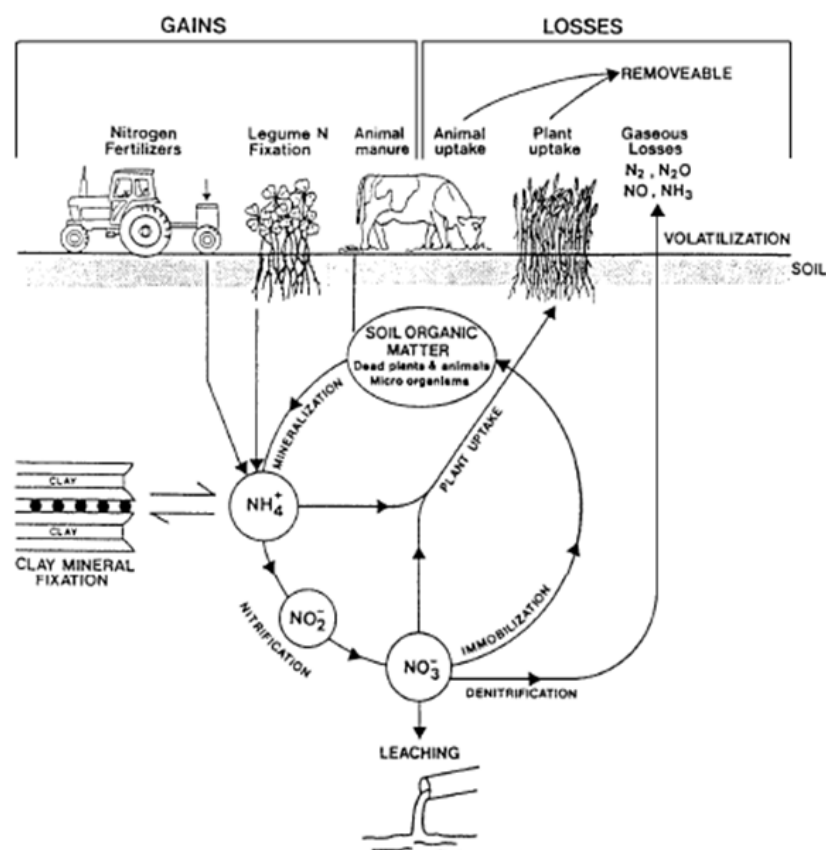


FIGURE B.1 – Le cycle de l'Azote dans un système agricole (d'après McLauren et Cameron [7]).

# Annexe C

## L'intervalle entre coupes de fourrage

Toutes les micro-parcelles d'un même site sont fauchées à la même date (soient 30 coupes le même jour). Ainsi, pour chaque site, nous calculons l'intervalle de temps en jours entre chaque coupe, que nous moyennons par saison et par année. Le tableau C.1 présente un extrait de la base de données obtenue concernant le site du LYCEE.

num. obs.	intervalle (jours)	Y (kg MS/ha/coupe)	site	sais	annee
1	46.0	1325.9	LYC	Q1	2006
2	38.3	3556.7	LYC	Q2	2006
3	56.5	2833.8	LYC	Q3	2006
4	50.0	3267.3	LYC	Q4	2006
5	62.5	2391.2	LYC	Q1	2007
6	55.5	3993.3	LYC	Q2	2007

TABLE C.1 – Extrait de la base de données construite.

Nous souhaitons savoir si ces intervalles moyens entre coupes sont distribués de façon égale pour chacun des sites. L'hypothèse de normalité n'étant pas confirmée sur les données brutes, nous étudions alors la transformation inverse de l'intervalle entre coupes.

Dans un premier temps, nous comparons les intervalles moyens ou durées moyennes selon les sites. Avant d'effectuer une ANOVA, nous vérifions l'homogénéité des variances selon les groupes. Graphiquement, l'hypothèse semble être valide (Figure C.1). Nous confirmons ce postulat à l'aide d'un test de Levene, avec la modification de Brown et Forsythe, qui fait intervenir l'écart absolu à la médiane (plutôt qu'à la moyenne pour la version classique). Pour l'ensemble des saisons, l'hypothèse d'égalité des variances des intervalles entre coupes entre les sites est conservée au seuil 5%, avec une p-valeur de 0.491. Ce résultat indique que les variances des durées sont égales quel que soit le site (toute saison confondue). Nous effectuons ensuite une ANOVA sur l'inverse de l'intervalle entre coupes selon le facteur site, puisque les hypothèses sont à présent vérifiées. La p-valeur obtenue est de 0.623, l'effet du site sur l'intervalles entre coupes n'est donc pas significatif à 5%.

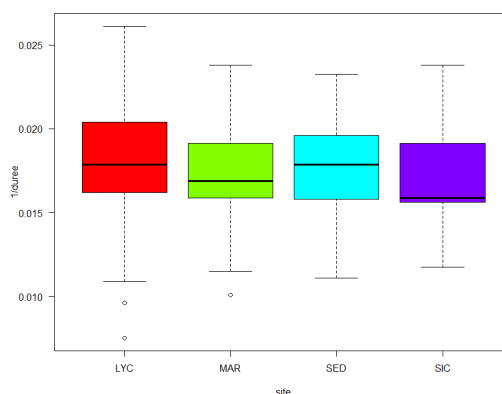


FIGURE C.1 – Représentation en boîte à moustache de la dispersion des durées selon le site.

La seconde hypothèse à vérifier est l'existence d'un effet du site sur l'intervalle entre deux coupes successives selon la saison. Les tests d'égalité des variances des durées par saison ne rejettent pas l'hypothèse nulle à 5%. En effet, les p-valeurs obtenues avec le test de Levene pour les 4 saisons sont respectivement : 0.570, 0.306, 0.535 et 0.378. Ces résultats nous permettent d'effectuer une ANOVA par saison, afin de tester l'effet du site comme précédemment. Ainsi, aucun des 4 tests n'est significatif à 5% (les p-valeurs sont de 0.956, 0.330, 0.711 et 0.531); l'hypothèse d'égalité des moyennes des durées par site selon la saison n'est donc pas rejetée.

Par conséquent, nous pouvons supposer que l'intervalle entre les coupes est le même d'un site à un autre pour une saison donnée, et pour toutes les années. Ainsi, pour la suite des analyses, nous pouvons étudier le rendement saisonnier exprimé en kg de MS/ha/coupe, indépendamment de l'intervalle entre coupe (ou temps de repousse). Dans le cas échéant, c'est à dire si les intervalles entre deux coupes successives avaient été significativement différents entre les sites pour une même saison, nous aurions calculé un rendement journalier moyen entre deux coupes successives, pour ensuite se ramener à l'échelle de la saison.

# Annexe D

## Analyse exploratoire du rendement fourrager

Le tableau D.1 présente un extrait des données sur lesquelles nous travaillons dans lors de l'analyse du rendement fourrager. Le graphique D.1 représente le rendement moyen obtenu par

num. obs.	Y (kg MS/ha/coupe)	site	ferti	bloc	sais	annee
358	3293.75	LYC	F4	1	T2	2011
359	3790.89	LYC	F4	1	T3	2011
360	1975.85	LYC	F4	1	T4	2011
361	8122.25	LYC	F4	1	T1	2012
362	2127.69	LYC	F4	1	T2	2012
363	3682.80	LYC	F4	1	T3	2012
364	1978.58	LYC	F4	1	T4	2012
365	3855.40	SED	F4	1	T1	2006
366	4687.50	SED	F4	1	T2	2006
367	2762.50	SED	F4	1	T3	2006
368	2974.00	SED	F4	1	T4	2006

TABLE D.1 – Extrait des données disponibles.

coupe selon le traitement de fertilisation à titre illustratif, sur la durée de l'essai.

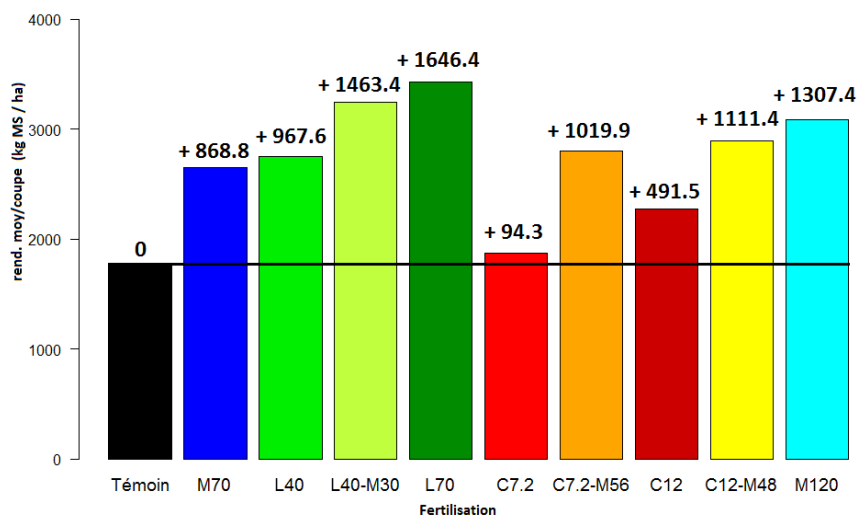


FIGURE D.1 – Diagramme en bâtons du rendement moyen par coupe et par fertilisation, en kilogramme de matière sèche par hectare.

Les observations effectuées ne semblent pas très dépendantes entre les années, comme le montre le graphique des corrélations D.2. Lors de l'analyse du rendement fourrager, nous considérons que les observations sur une année d'une même micro-parcelle sont indépendantes à celles de l'année suivante.

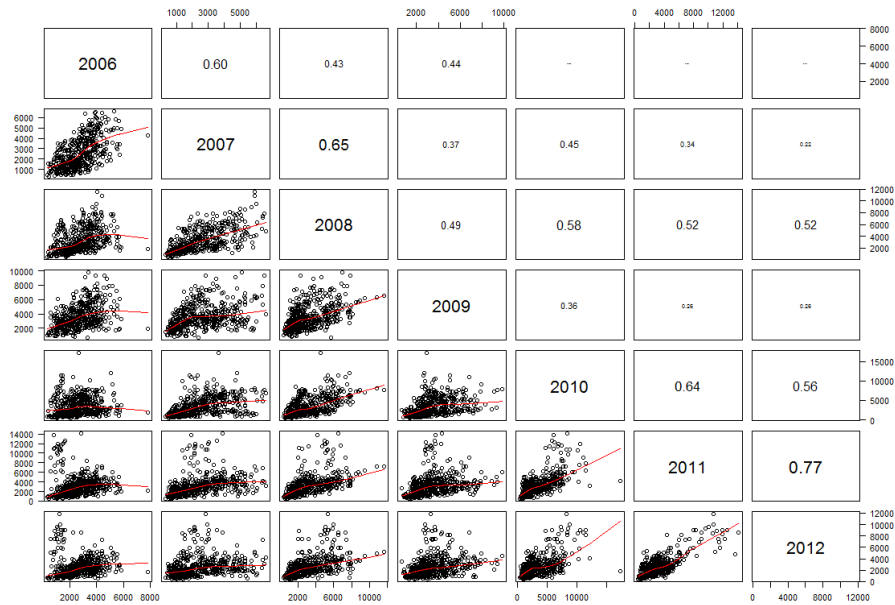


FIGURE D.2 – Représentation de la matrice des corrélations estimées sur les observations du rendement pour chaque couple d'années.

La variable à expliquer du rendement, notée  $Y$ , ne vérifie pas l'hypothèse de normalité. Une transformation logarithmique semble palier au problème de normalité des observations, comme le montre les histogrammes et diagrammes de dispersion de la figure D.3. La densité estimée semble mieux centrée, et la variabilité des données est réduite. De plus, la figure D.4 montre que les quantiles théoriques et les quantiles estimés sont mieux alignés lorsqu'on a recours au logarithme.

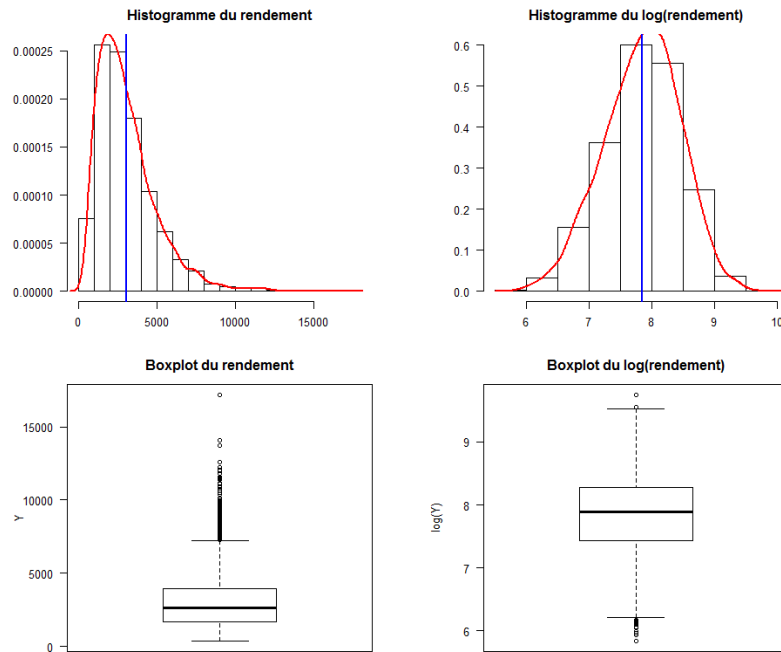


FIGURE D.3 – Histogrammes et boxplots de la variable à expliquer du rendement  $Y$  et de sa transformation logarithmique.

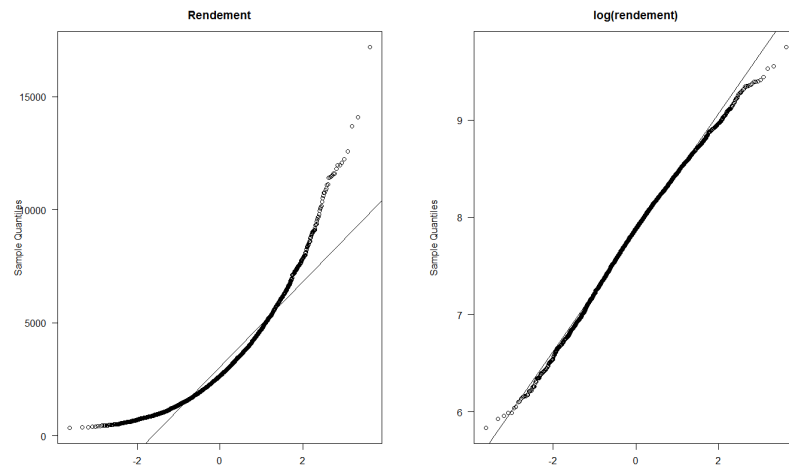


FIGURE D.4 – Graphiques des quantiles théoriques et estimés du rendement  $Y$  et de sa transformation logarithmique.

Les graphiques en boîtes à moustaches des figures D.5 et D.6 permettent de se faire une idée de la dispersion du rendement pour chaque modalité du facteur fertilisation globalement et selon le site, ainsi que de la dispersion au sein de chaque saison (D.7).

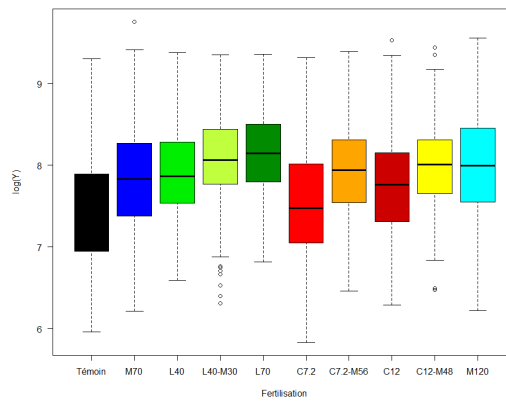


FIGURE D.5 – Représentation en boîte à moustaches de la dispersion des observations du logarithme du rendement  $Y$  par fertilisation.

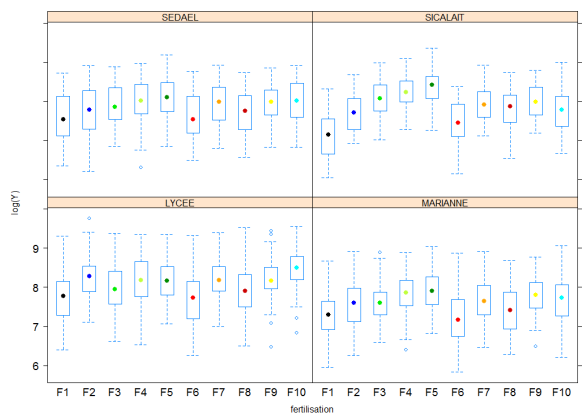


FIGURE D.6 – Représentation en boîte à moustaches de la dispersion des observations du logarithme du rendement  $Y$  par site et par fertilisation.

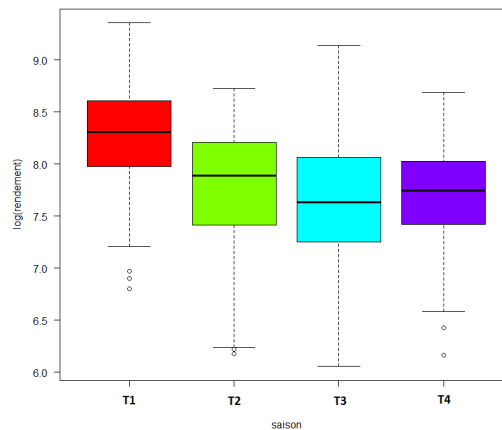


FIGURE D.7 – Représentation en boîte à moustache de la dispersion des observations du  $\log(\text{rendement})$  selon la saison.

# Annexe E

## Le modèle additif et les splines

### E.1 Définition du modèle

Pour introduire les splines et la méthode d'estimation du modèle, nous nous plaçons dans le cadre de l'estimation de la fonction de régression du modèle suivant :

$$Y = f(x) + \epsilon,$$

avec  $\mathbb{E}(\epsilon) = 0$ ,  $\mathbb{V}(\epsilon) = \sigma^2$ , et où

$$f(x) = \sum_{j=1}^J b_j(x)\beta_j,$$

sous la forme matricielle :

$$f(X) = X\beta,$$

avec  $X$  la matrice du modèle. Ainsi, pour un ensemble de valeurs de  $x$ , sur lesquelles on évalue la fonction spline, il est possible de construire la matrice de coefficients  $\beta$ . Nous définissons la matrice de pénalité  $S$ , généralement semi-définie positive, telle que :

$$\int f''(x)^2 dx = \beta^T S \beta.$$

### E.2 Les fonctions splines

Afin d'estimer les fonctions de lissage  $f_i$ , nous utilisons une approche non-paramétrique, les splines de régression pénalisée. Cette méthode se situe entre les splines de régression et les splines de lissage. Nous nous plaçons dans le cas cubique, où les fonctions de puissance tronquées sont toutes de degré 3.

#### Les splines de régression cubiques

L'approche des splines de régression est paramétrique. La spline est une combinaison linéaire d'un polynôme (ici de degré 3) et de fonctions de puissance tronquées :

$$f(x) = \beta_0 + \beta_{0,1}x + \beta_{0,2}x^2 + \beta_{0,3}x^3 + \sum_{k=1}^K \beta_k(x - \kappa_k)_+^3,$$

$$\text{avec } (x - \kappa_k)_+^3 = (x - \kappa_k)^3 \mathbb{1}(x \geq \kappa_k),$$

où  $(\beta_0, \beta_{0,1}, \dots, \beta_K)$  est un vecteur de coefficients inconnus,  $(\kappa_1, \kappa_2, \dots, \kappa_K)$  est un vecteur de  $K$  noeuds fixés. L'estimation des coefficients peut être effectuée par l'approche des moindres carrés. Le choix des noeuds reste l'inconvénient majeur de cette méthode. La figure E.1 illustre une fonction spline de régression cubique (à gauche) et une combinaison de ces fonctions lors de l'approximation d'une courbe non-linéaire (à droite). Dans le cadre des modèles additifs, l'ajustement peut se faire par maximisation de la vraisemblance par la méthode IRLS (P. McCullagh and J. Nelder, [6]). Une fois le nombre et la position des noeuds fixés, la méthode IRLS consiste à définir une matrice de design appropriée.

#### Les splines de lissage



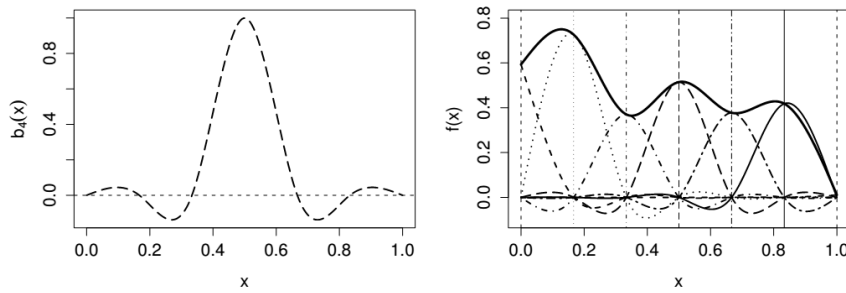


FIGURE E.1 – Illustration d’une fonction de base  $b_4(x)$  pour une spline de régression cubique (à gauche) et d’une combinaison de plusieurs fonctions de base lors du lissage d’une courbe (à droite).

Une solution non-paramétrique consiste à utiliser un grand nombre de noeuds et d’appliquer un terme de pénalité au problème des moindres carrés, pour ainsi éviter tout sur-ajustement. Les coefficients sont alors estimés par minimisation de la somme des carrés pénalisée :

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int f''(x) dx,$$

où  $\lambda \int f''(x) dx$  pénalise la courbure de la fonction ajustée et  $\lambda$  est la paramètre de lissage. Le problème du choix de l’ensemble de noeuds est remplacé par celui de la sélection du paramètre de lissage  $\lambda$ . En choisissant un petit  $\lambda$ , la solution au problème sera une spline cubique qui interpolera les données. Augmenter la valeur de  $\lambda$  aboutira à une approximation plus lisse.

L’estimation requiert donc de maximiser la vraisemblance pénalisée, par l’intermédiaire d’algorithmes tels que le ”backfitting” (Hastie and Tibshirani, [1]), qui sont assez coûteux en terme de temps de calcul. L’algorithme du backfitting présente cependant quelques inconvénients, notamment en termes de convergence, et une sous-estimation des erreurs standards.

### Les splines de régression pénalisées

Ainsi, cette dernière approche utilise les deux précédentes, ce qui nécessite de spécifier l’ensemble des noeuds et le terme de pénalité. La méthode directe de Wood ([15]), moins coûteuse, permet de maximiser la vraisemblance pénalisée tout en spécifiant un vecteur de noeuds. Dans notre étude, l’estimation du modèle additif se fait via une représentation mixte. En effet, il est possible d’écrire n’importe quel lissage de régression pénalisée comme des composants fixes et aléatoires d’un modèle mixte. En réécrivant la spline de régression cubique précédente :

$$f(x) = \sum_{j=0}^2 \beta_{0,j} x^j + \sum_{k=1}^K \beta_k (x - \kappa_k)^3,$$

sous la forme matricielle :

$$f(x) = X_F \beta_F + T \beta,$$

où  $X_F = [1, x, x^2]$ ,  $\beta_F = (\beta_0, \beta_{0,1}, \beta_{0,2})'$ ,  $T = [(x - \kappa_1)^3, \dots, (x - \kappa_K)^3]$  et  $\beta = (\beta_1, \dots, \beta_K)'$ , nous pouvons facilement identifier une formulation du modèle de type mixte. Cette approche nous permet d’estimer le paramètre de lissage  $\lambda$ , tout en réduisant la variance des estimations.

### E.3 Estimation via la modélisation mixte

Les modèles GAM admettent une représentation de type mixte, où les composantes du lissage sont traitées comme des effets fixes et aléatoires, ce qui nous permet d’estimer les paramètres via REML.

Considérons le terme de lissage associé au vecteur de paramètres  $\beta$ , à la matrice du modèle  $X$ , à la matrice de pénalité  $S$  et à la matrice de contraintes  $C$ . Le modèle peut être estimé par minimisation en  $\beta$  de l'expression de régression pénalisée :

$$s(\beta) = \|y - X\beta\|^2 + \lambda\beta^T S\beta, \text{ sous contrainte } C\beta = 0.$$

La contrainte peut être absorbée par décomposition QR :  $QR = C^T$ , où  $Q$  est une matrice orthogonale et  $R$  est une matrice triangulaire supérieure. Posons  $Z$ , une matrice égale à  $Q$  sans les  $n_c$  premières colonnes, où  $n_c$  est le nombre de lignes de  $C$ . En notant  $\beta = Z\beta_z$ , l'expression précédente devient :

$$s = \|y - XZ\beta_z\|^2 + \lambda\beta_z^T Z^T S Z \beta_z.$$

Afin d'identifier la partie fixe et la partie aléatoire du lissage, nous commençons par chercher une décomposition en valeurs propres de  $Z^T S Z$  :  $Z^T S Z = U D U'$ , où  $U$  est une matrice ortho-normale, dont les colonnes sont les vecteurs propres de  $Z^T S Z$ , et  $D$  est une matrice diagonale contenant les valeurs propres correspondantes, arrangées par ordre décroissant sur la diagonale principale. En posant  $\beta_u = U^T \beta_z$ , nous pouvons ainsi réécrire l'expression à minimiser :

$$s = \|y - XZU\beta_u\|^2 + \lambda\beta_u^T D\beta_u.$$

Généralement, les derniers éléments de la diagonale de  $D$  sont nuls. Nous considérons donc  $D^+$  la sous-matrice de  $D$  dont les éléments de la diagonale sont non-nuls. La partition  $\beta_u$  est telle que  $\beta_u^T = [b_u^T, \beta_F^T]$ , correspondant aux effets aléatoires et effets fixes, et telle que  $\beta_u^T D\beta_u = b_u^T D^+ b_u$ . De manière analogue, nous partitionnons les colonnes de  $XZU$  en  $[X_u, X_F]$ . En notant  $b = \sqrt{D^+} b_u$  et  $X_R = X_u(\sqrt{D^+})^{-1}$ , l'expression à minimiser devient :

$$s = \|y - X_F\beta_F - X_R b\|^2 + \lambda b^T b.$$

Il est à présent simple de montrer que, pour un  $\lambda$  donné, les estimations obtenues par minimisation de  $s$  correspondent à l'espérance de  $b$  et à l'estimation de  $\beta_F$  sachant  $y$  sous le modèle mixte suivant :

$$y = X_F\beta_F + X_R b + \epsilon,$$

où  $\epsilon \sim \mathcal{N}(0, I\sigma^2)$  et  $b \sim \mathcal{N}(0, I\tau)$ , avec  $\lambda = 1/\tau$ . Cette méthode conduit à estimer  $1/\lambda$  par REML.

Nous pouvons élargir cette approche au contexte additif. Il s'agit ainsi de produire un  $X_F$  et un  $X_R$  pour chacun des termes de lissage du modèle. Les colonnes des  $X_F$  obtenus sont combinées dans une seule matrice d'effets fixes, notée  $X_\Sigma$ , à laquelle nous pouvons ajouter des colonnes pour la constante du modèle ainsi que pour la partie strictement paramétrique. Le modèle résultant devient :

$$y = X_\Sigma\beta_\Sigma + X_{R1}b_1 + X_{R2}b_2 + \dots + \epsilon,$$

où  $\epsilon \sim \mathcal{N}(0, I\sigma^2)$ , et les  $b_i \sim \mathcal{N}(0, I\tau_i)$ . Les "quantités" de lissages pour chaque courbes sont indépendantes les unes des autres. Les estimations simultanées des composants de la variance et des paramètres de lissage se font de la même façon que dans le cas d'un terme unique.

## E.4 Interactions avec les termes de lissage

Très souvent, dans les modèles additifs, la forme fonctionnelle des effets d'une covariable varie selon les groupes définis par les niveaux d'une variable catégorielle. Tout comme pour le modèle linéaire, il est possible d'inclure un ou plusieurs termes d'interaction entre les variables catégorielles et les variables quantitatives, pour lesquelles une tendance lissée est spécifiée. Le modèle ajuste une courbe distincte pour chaque catégorie du facteur. Les fonctions de lissage sont toutes soumises à des contraintes de centrage, il est donc nécessaire d'ajouter les effets fixes du facteur au modèle.

Par simplicité, nous considérons dans un premier temps l'interaction d'un seul facteur avec un seul prédicteur continu lissé. Soient  $(x - i, y_i, z_i)$ ,  $1 \leq i \leq n$ , où  $x_i$  est un prédicteur continu,  $y_i$  est la réponse et  $z_i \in \{1, \dots, L\}$  représente un facteur catégoriel. Le type de modèle que nous souhaitons ajuster ici est :

$$y_i = f_{z_i}(x_i) + \epsilon_i,$$

où les  $\epsilon_i$  sont iid de loi  $\mathcal{N}(0, \sigma^2)$ , et les  $f_1, \dots, f_L$  sont  $L$  fonctions différentes qui dépendent de la valeur de  $z_i$ .

Soient  $\kappa_1, \dots, \kappa_K$  un ensemble de noeuds distincts, suffisamment nombreux, appartenant à l'ensemble des  $x_i$ , et  $x_+ = \max(0, x)$ . Définissons les variables binaires  $z_{il}$  telles que :

$$z_{il} = \begin{cases} 1 & \text{si } z_i = l \\ 0 & \text{sinon} \end{cases}$$

pour  $l = 1, \dots, L$ . Considérons un modèle de type spline de régression cubique pénalisée tel que :

$$y_i = \beta_0 + \beta_{0,1}x_i + \beta_{0,2}x_i^2 + \beta_{0,3}x_i^3 + \sum_{k=1}^K b_k(x_i - \kappa_k)_+^3 \\ + \sum_{l=2}^L z_{il}(\gamma_{0l} + \gamma_{1l}x_i + \gamma_{2l}x_i^2 + \gamma_{3l}x_i^3) + \sum_{l=1}^L z_{il} \left\{ \sum_{k=1}^K c_k^l(x_i - \kappa_k)_+^3 \right\} + \epsilon_i,$$

où les  $b_k$  sont iid de loi  $\mathcal{N}(0, \sigma_b^2)$  et les  $c_k^l$  sont aussi iid de loi  $\mathcal{N}(0, \sigma_{cl}^2)$ , pour  $l = 1, \dots, L$ .

Comme précédemment, nous utilisons la formulation du modèle mixte :

$$y = X_F \beta_F + X_R b + \epsilon,$$

où

$$\beta_F = (\beta_0, \dots, \beta_{0,3}, \gamma_{02}, \dots, \gamma_{0L}, \gamma_{12}, \dots, \gamma_{1L})^T, \\ b = (b_1, \dots, b_K, c_1^1, \dots, c_K^1, c_1^2, \dots, c_K^L),$$

et

$$\begin{pmatrix} b \\ \epsilon \end{pmatrix} \sim \mathcal{N} \left( 0, \begin{bmatrix} G & 0 \\ 0 & \sigma_\epsilon^2 I \end{bmatrix} \right),$$

avec  $G = \text{diag}(\sigma_b^2 1_K, \sigma_{c_1}^2 1_K, \dots, \sigma_{c_L}^2 1_K)$ . Les méthodes utilisées dans le cadre du modèle mixte linéaire peuvent alors être appliquées, comme l'estimation par la méthode BLUP (Best Linear Unbiased Predictor).

Ce modèle spécifie ainsi une fonction de lissage différente pour chaque sous-ensemble d'observations définis par les niveaux du facteur  $z$ . Le problème peut être facilement élargi au cas du modèle additif, où l'interaction est estimée de la même façon.

# Annexe F

## Analyse exploratoire de la teneur en Azote et de la CEC des sols

### F.1 Analyse exploratoire de la teneur en Azote

Le graphique de la figure F.1 permet de visualiser la teneur en Azote moyenne obtenue annuellement par fertilisation. Les taux d'azote observés sur chaque micro-parcelle lors des sept

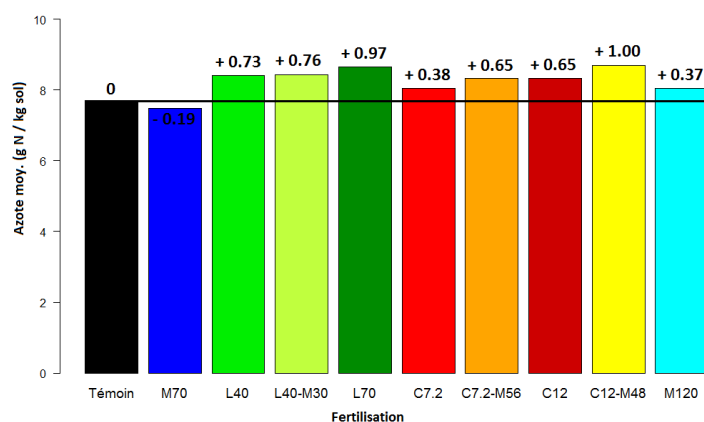


FIGURE F.1 – Diagramme en bâtons de la teneur en Azote moyenne obtenue annuellement par fertilisation, en gramme d'Azote par kilogramme d'échantillon de sol sec (g N/kg sol).

années sont représentés sur la figure F.2.

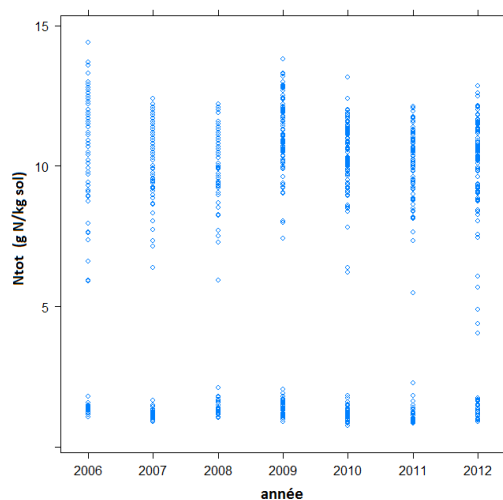


FIGURE F.2 – Graphiques des observations du taux d'Azote dans le temps (en g N/100g sol).

## F.2 Analyse exploratoire de la CEC

La CEC annuelle moyenne obtenue par fertilisation est illustrée sur le graphique de la figure F.3. Le graphique de la figure F.4 nous donne une idée de la dispersion des mesures de la CEC

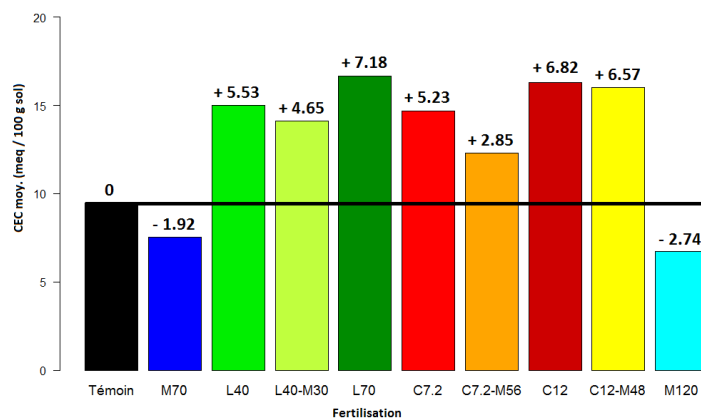


FIGURE F.3 – Diagramme en bâtons de la CEC moyenne obtenue annuellement par fertilisation, en milligrammes d'équivalents par 100 grammes d'échantillon de sol sec (meq/100g sol).

sur ces micro-parcelles lors des 7 années d'essai.

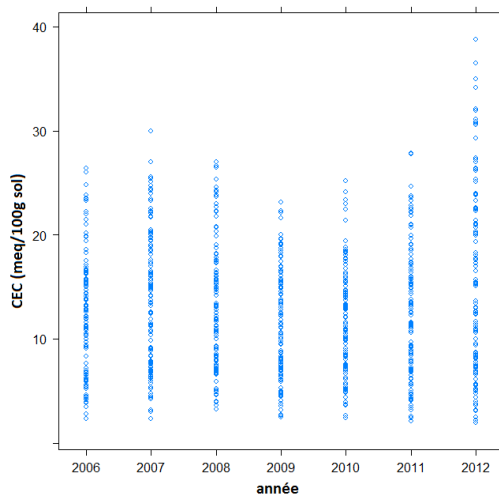


FIGURE F.4 – Graphiques des observations de la CEC (en meq/100g sol) dans le temps.

# Annexe G

## Estimations des modèles

### G.1 Le rendement fourrager

Le tableau G.1 présente les résultats des tests et les effets principaux (sur le logarithme de la variable  $Y$ ) du modèle du rendement fourrager (le site du LYCEE, la fertilisation F1 et la saison T1 sont les modalités de référence). Les tableaux suivants résument les estimations du modèle

	valeur	erreur estimée	t-valeur	p-valeur
constante	8.02	0.07	108.61	0.00
site.MAR	-0.33	0.10	-3.42	0.00
site.SED	0.04	0.09	0.46	0.65
site.SIC	-0.46	0.09	-5.03	0.00
T2	-0.10	0.05	-1.98	0.05
T3	-0.28	0.06	-4.83	0.00
T4	-0.67	0.06	-11.59	0.00
ferti 2	0.52	0.08	6.45	0.00
ferti 3	0.21	0.08	2.67	0.01
ferti 4	0.43	0.08	5.31	0.00
ferti 5	0.40	0.08	4.96	0.00
ferti 6	-0.06	0.08	-0.69	0.49
ferti 7	0.46	0.08	5.80	0.00
ferti 8	0.17	0.08	2.09	0.04
ferti 9	0.46	0.08	5.72	0.00
ferti 10	0.69	0.08	8.65	0.00

TABLE G.1 – Tests et significativité des effets principaux du modèle du rendement fourrager (le site du LYCEE, la fertilisation F1 et la saison S1 sont les modalités de référence).

“global” (comportant l’effet du site) pour chaque combinaison de modalités des facteurs.

Fertilisation \ saison T	T.1	T.2	T.3	T.4
	saison 1	saison 2	saison 3	saison 4.
F1 / Témoin	3039	2747	2304	1550
F2 / M70	5095	4607	3863	2600
F3 / L40	3762	3402	2852	1920
F4 / L40M30	4649	4203	3524	2372
F5 / L70	4522	4089	3428	2308
F6 / C7.2	2875	2600	2180	1467
F7 / C7.2M56	4834	4371	3665	2466
F8 / C12	3593	3249	2724	1833
F9 / C12M48	4806	4346	3644	2452
F10 / M120	6074	5492	4605	3099

TABLE G.2 – Estimations du rendement moyen par coupe et par saison (en  $kg$  MS/ha/coupe) sur le site du LYCEE.

Fertilisation \ saison T	T.1	T.2	T.3	T.4
	saison 1	saison 2	saison 3	saison 4.
F1 / Témoin	1917	1194	1014	1091
F2 / M70	3399	2117	1798	1935
F3 / L40	4818	3001	2548	2743
F4 / L40M30	5830	3631	3084	3320
F5 / L70	6489	4041	3432	3695
F6 / C7.2	2751	1713	1455	1566
F7 / C7.2M56	4153	2587	2197	2365
F8 / C12	3655	2276	1933	2081
F9 / C12M48	4470	2784	2364	2545
F10 / M120	3683	2294	1948	2097

TABLE G.3 – Estimations du rendement moyen par coupe et par saison (en *kg* MS/ha/coupe) sur le site de SICALAIT.

Fertilisation \ saison T	T.1	T.2	T.3	T.4
	saison 1	saison 2	saison 3	saison 4.
F1 / Témoin	2185	1338	883	1611
F2 / M70	3037	1859	1227	2239
F3 / L40	3000	1837	1212	2213
F4 / L40M30	3918	2399	1583	2890
F5 / L70	4189	2565	1692	3089
F6 / C7.2	2016	1234	814	1487
F7 / C7.2M56	3125	1913	1262	2304
F8 / C12	2567	1571	1037	1893
F9 / C12M48	3713	2273	1500	2738
F10 / M120	3405	2085	1376	2511

TABLE G.4 – Estimations du rendement moyen par coupe et par saison (en *kg* MS/ha/coupe) sur le site de MARIANNE.

Fertilisation \ saison T	T.1	T.2	T.3	T.4
	saison 1	saison 2	saison 3	saison 4.
F1 / Témoin	3171	2159	1422	1471
F2 / M70	3845	2618	1725	1783
F3 / L40	4542	3093	2037	2106
F4 / L40M30	4983	3393	2235	2311
F5 / L70	5534	3769	2483	2567
F6 / C7.2	3172	2160	1423	1471
F7 / C7.2M56	4469	3043	2005	2072
F8 / C12	3542	2412	1589	1643
F9 / C12M48	4541	3092	2037	2106
F10 / M120	5190	3534	2328	2407

TABLE G.5 – Estimations du rendement moyen par coupe et par saison (en *kg* MS/ha/coupe) sur le site de SEDAEL.

## G.2 La teneur en Azote

Le tableau G.6 résume les estimations des effets principaux (paramétriques) du modèle retenu pour la teneur en Azote, ainsi que la significativité associée à chaque niveau des facteurs. Les estimations associées aux effets d'interaction entre le site et la fertilisation sont aussi présentées dans ce tableau. Le graphique de la figure G.1 nous donne une idée des estimations de la teneur

	Estimation	Erreur Std.	t-valeur	Pr(> t )
<b>(constante)</b>	0.02	0.04	0.44	0.66
<b>ferti 2</b>	-0.00	0.05	-0.01	0.99
ferti 3	0.21	0.05	3.78	0.00
ferti 4	0.30	0.05	5.50	0.00
ferti 5	0.44	0.05	8.11	0.00
ferti 6	0.19	0.05	3.59	0.00
ferti 7	0.28	0.05	5.15	0.00
ferti 8	0.22	0.05	4.00	0.00
ferti 9	0.38	0.05	6.98	0.00
<b>ferti 10</b>	0.11	0.05	1.98	0.05
site.MAR	2.28	0.06	38.19	0.00
site.SED	2.25	0.05	43.66	0.00
site.SIC	2.29	0.04	53.94	0.00

	Estimation	Erreur Std.	t-valeur	Pr(> t )
ferti 2 : site.MAR	-0.20	0.08	-2.34	0.02
ferti 3 : site.MAR	-0.21	0.08	-2.49	0.01
ferti 4 : site.MAR	-0.26	0.08	-3.05	0.00
ferti 5 : site.MAR	-0.40	0.08	-4.77	0.00
ferti 6 : site.MAR	-0.24	0.08	-2.84	0.00
ferti 7 : site.MAR	-0.25	0.08	-2.94	0.00
ferti 8 : site.MAR	-0.23	0.08	-2.79	0.01
ferti 9 : site.MAR	-0.34	0.08	-4.00	0.00
<b>ferti 10 : site.MAR</b>	-0.07	0.08	-0.80	0.42
<b>ferti 2 : site.SED</b>	0.00	0.07	0.07	0.95
<b>ferti 3 : site.SED</b>	-0.13	0.07	-1.73	0.08
ferti 4 : site.SED	-0.22	0.07	-3.08	0.00
ferti 5 : site.SED	-0.35	0.07	-4.82	0.00
ferti 6 : site.SED	-0.17	0.07	-2.27	0.02
ferti 7 : site.SED	-0.23	0.07	-3.16	0.00
ferti 8 : site.SED	-0.15	0.07	-2.13	0.03
ferti 9 : site.SED	-0.25	0.07	-3.49	0.00
<b>ferti 10 : site.SED</b>	-0.04	0.07	-0.53	0.59
<b>ferti 2 : site.SIC</b>	0.07	0.06	1.15	0.25
<b>ferti 3 : site.SIC</b>	-0.05	0.06	-0.79	0.43
ferti 4 : site.SIC	-0.15	0.06	-2.58	0.01
ferti 5 : site.SIC	-0.27	0.06	-4.53	0.00
<b>ferti 6 : site.SIC</b>	-0.06	0.06	-1.06	0.29
ferti 7 : site.SIC	-0.14	0.06	-2.35	0.02
<b>ferti 8 : site.SIC</b>	-0.06	0.06	-1.01	0.31
ferti 9 : site.SIC	-0.23	0.06	-3.85	0.00
<b>ferti 10 : site.SIC</b>	-0.08	0.06	-1.30	0.19

TABLE G.6 – Estimations des effets principaux et des termes d'interaction du modèle additif hétérogène.

en Azote (en  $g N/kg$  sol) par le modèle pour toutes les combinaisons des modalités des facteurs (fertilisation/site) sur les sept premières années de l'étude, ainsi que l'intervalle de confiance (pointillées) associé à ces prédictions.



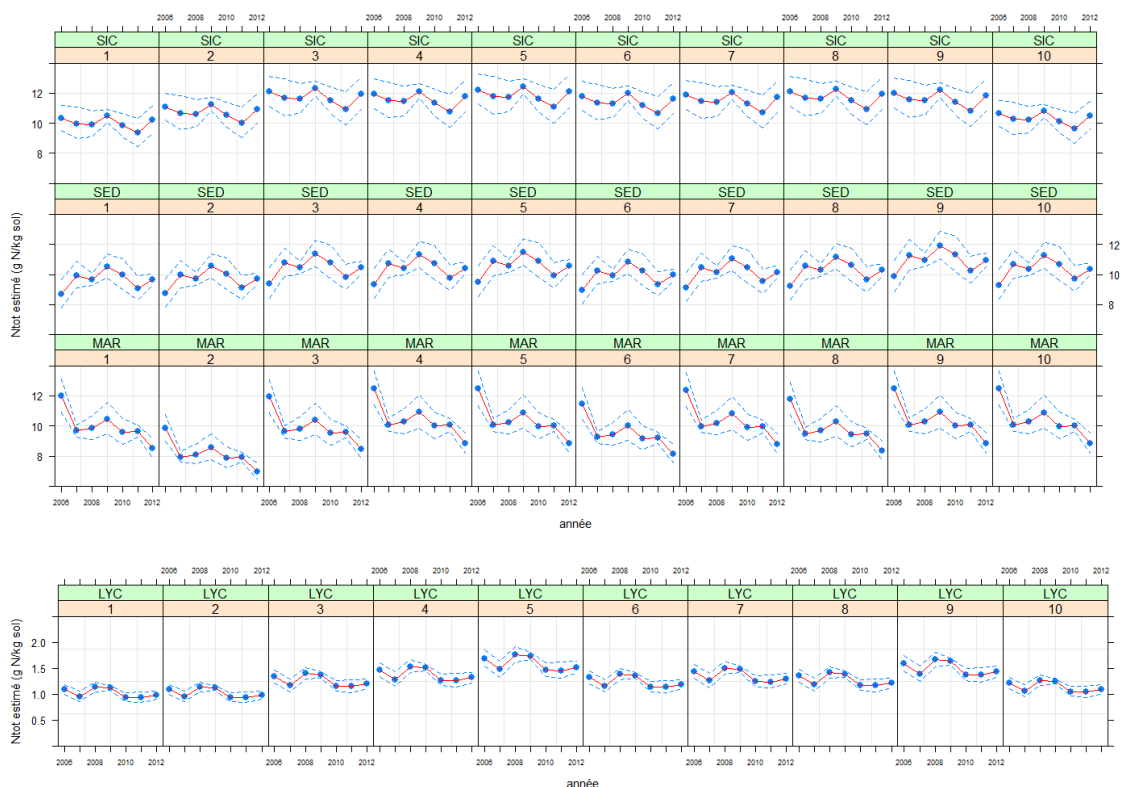


FIGURE G.1 – Estimations de la teneur en Azote (en  $g N/kg$  sol) par le modèle retenu pour chaque combinaison de modalités des facteurs sur les 7 premières années.

Les sites sont notés par : SIC=SICALAIT, SED=SEDAEL, MAR=MARIANNE, LYC=LYCEE.

Les fertilisations sont notées par : 1=F1, 2=F2, 3=F3, 4=F4, 5=F5, 6=F6, 7=F7, 8=F8, 9=F9, 10=F10.

### G.3 La CEC

Le tableau G.7 permet de comparer les estimations des effets principaux des deux modèles construits sur la CEC, à savoir le modèle comportant une fonction de lissage par fertilisation et le modèle comportant une fonction de lissage par le site. Les résultats des effets moyens associés aux fertilisations et aux sites sont très proches, ce qui s'explique par le fait que les effets fixes (de la fertilisation et du site) sont évalués pour l'année de référence, l'année 0 = 2006, soit au démarrage de l'essai. Le graphique de la figure G.2 nous donne une idée des estimations de la CEC (en  $meq/100g$  sol) par le modèle comportant un lissage par fertilisation pour toutes les combinaisons des modalités des facteurs (fertilisation/site) sur les sept premières années de l'étude, ainsi que l'intervalle de confiance (pointillées) associé à ces prédictions.

Modèle	Lissage par site		Lissage par ferti.		
	Terme	Estimation	Pr(> t )	Estimation	Pr(> t )
constante		-0.31	0.76	2.16	0.00
ferti 2		-0.50	0.00	-0.38	0.00
ferti 3		0.52	0.00	0.60	0.00
ferti 4		0.44	0.00	0.56	0.00
ferti 5		0.71	0.00	0.85	0.00
ferti 6		0.53	0.00	0.53	0.00
ferti 7		0.20	0.02	0.32	0.00
ferti 8		0.62	0.00	0.67	0.00
ferti 9		0.57	0.00	0.68	0.00
ferti 10		-0.65	0.00	-0.52	0.00
site.MAR		1.42	0.00	1.33	0.00
site.SED		0.95	0.00	0.95	0.00
site.SIC		1.73	0.00	1.74	0.00
log(bio)		0.26	0.01	-	-

TABLE G.7 – Estimations des termes paramétriques du modèle additif

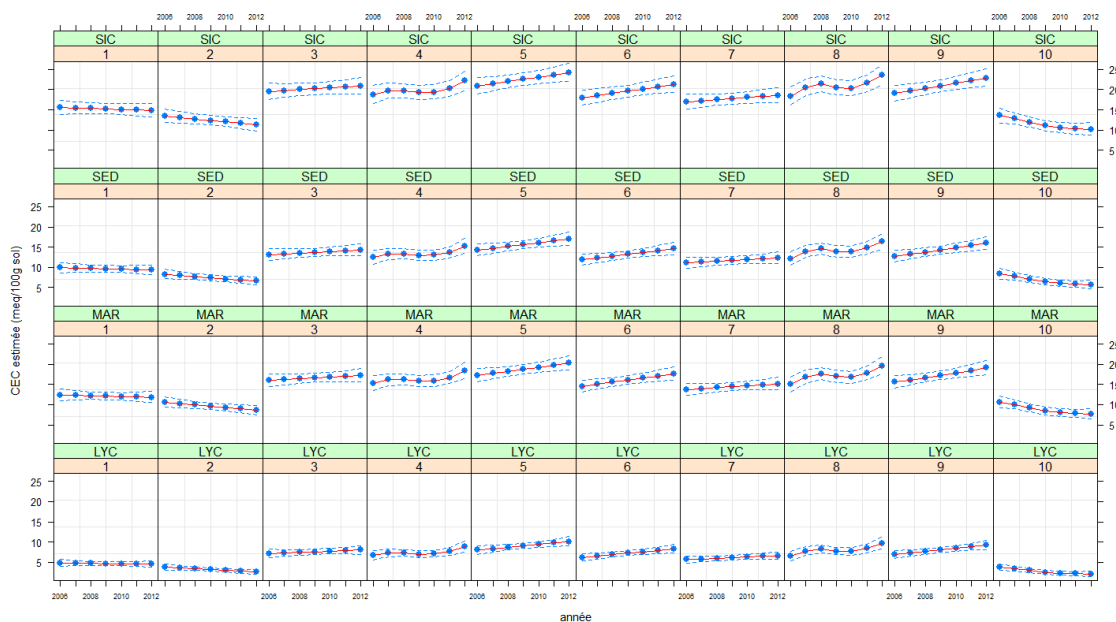


FIGURE G.2 – Estimations de la CEC (en  $meq/100g$  sol) par le modèle retenu pour chaque combinaison de modalités des facteurs sur les 7 premières années.

Les sites sont notés par : SIC=SICALAIT, SED=SEDAEL, MAR=MARIANNE, LYC=LYCEE.

Les fertilisations sont notées par : 1=F1, 2=F2, 3=F3, 4=F4, 5=F5, 6=F6, 7=F7, 8=F8, 9=F9, 10=F10.