UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline : Biostatistique
Formation Doctorale : Biostatistique
Ecole Doctorale : Information Structures Systèmes

présentée et soutenue publiquement

par

Ibnou DIENG

le 24 janvier 2007

Titre :

Prédiction de l'interaction génotype $\times$ environnement par linéarisation et régression PLS-mixte

JURY

M. Robert SABATIER	Directeur de Thèse
M. Eric GOZÉ	Codirecteur de Thèse
M. Alain CHARCOSSET	Rapporteur
M. Jean-Jacques DAUDIN	Rapporteur
M. Gilles DUCHARME	Examinateur
Mme Christèle ROBERT-GRANIER	Examinateur

à la mémoire de Jean Parriaud

Remerciements

De nombreuses personnes et institutions se sont mobilisées pour cette thèse. Qu'elles trouvent à travers ces lignes mes plus sincères remerciements.

Je voudrais tout d'abord exprimer ma gratitude à M. Robert SABATIER d'avoir accepté de diriger ce travail, de l'avoir constamment guidé et soutenu. Je lui en suis profondément reconnaissant.

Cette thèse n'aurait jamais vu le jour sans la confiance, la patience et la générosité de mon Codirecteur de thèse, M. Eric GOZÉ. Il a suggéré et initié cette orientation en troisième cycle qui a démarré par un Dea. Qu'il soit chaleureusement remercié pour la grande disponibilité, la constante sympathie et tous les conseils dans le cadre et en dehors de cette thèse.

Mes sincères remerciements sont aussi adressés à M. Philippe LETOURMY pour m'avoir accueilli dans son unité et d'avoir suivi et soutenu ce travail depuis le début.

Mon inscription en DEA de biostatistique n'aurait pas été possible sans la bienveillance du Professeur Gilles DUCHARME, responsable de la formation doctorale biostatistique à l'UM II, qui a ensuite facilité le montage institutionnel de cette thèse.

Je voudrais remercier tout particulièrement Mme Danièle CLAVEL, selectionneuse au Cirad pour avoir aussi été à l'origine de ces travaux, d'avoir fourni des données historiques et permis la mise en place d'une expérimentation pilote. Merci également à M. Jean-François RAMI pour les données complémentaires de cette étude.

A MM. Harold ROY-MACAULEY et Serge BRACONNIER, anciennement directeur et responsable scientifique du Ceraas, je voudrais exprimer ma profonde reconnaissance pour m'avoir recruté et accordé la possibilité de faire cette thèse. Ils m'ont facilité les allers et retours entre Montpellier et le Sénégal et m'ont toujours fait confiance.

Je remercie MM. Alain CHARCOSSET et Jean-Jacques DAUDIN pour l'intérêt manifesté à ce travail et pour avoir bien voulu accepter de le rapporter. Que soient aussi remerciés M. Gilles DUCHARME et Mme Christèle ROBERT-GRANIER pour avoir examiné ces travaux.

Luc BAUDOIN, Robert FAIVRE, Hervé MONOD et Danièle CLAVEL ont apporté de nombreux conseils et soutiens dans le cadre du comité de thèse et même au delà. Qu'ils trouvent ici l'expression de ma sincère reconnaissance.

Un grand merci à tous mes anciens collègues du Ceraas au Sénégal et aux nouveaux du Centre régional Agrhymet au Niger pour toutes les discussions professionnelles et non professionnelles. Y est aussi associée toute l'équipe de l'Upr 13 biostatistique du Cirad pour la convivialité.

Pendant toutes ces années, mes parents n'ont cessé de me soutenir et de me ré-armer moralement. Qu'ils soient ici remerciés d'avoir toujours été là. Une grande pensée va à l'endroit de mes amis du Sénégal, de la France et du Niger dont j'ai eu la très grande chance et l'immense plaisir d'avoir un jour croisé le chemin.

Enfin je ne saurais terminer sans remercier les institutions qui ont pris en charge financièrement ces travaux : le gouvernement français, le Cirad et le Ceraas. Une mention toute spéciale pour le Cirad qui, à travers l'Upr 13 et la Desi, n'a ménagé aucun effort pour dégager les ressources financières additionnelles pour que cette thèse puisse se terminer dans les meilleures conditions.

ACRONYMES

ACP	Analyse en Composantes Principales
AMMI	Additive Main effets and Multiplicative Interaction
APLAT	Approximation Par Linéarisation Autour d'un Témoin
ARABHY	Arachide Bilan Hydrique
CERAAS	Ceraas d'Etude Régional pour l'Amélioration de l'Adaptation à la Sécheresse
CILSS	Comité permanent Inter-États de Lutte contre la Sécheresse au Sahel
CIRAD	Centre international de Recherche Agronomique pour le Développement
CSSA	Cadre Stratégique de Sécurité Alimentaire
DHC	Diagnostic Hydrique des Cultures
DVS	Décomposition en valeurs Singulières
EM	Expectation-Maximization
GLS	Generally Least Squares
ICRISAT	International Crops Research Institute for the Semi-Arid Tropics
INERA	Institut de l'Environnement et de Recherches Agricoles
Interaction G×E	Interaction entre génotype et environnement
IRSIS	Irrigation Scheduling Information System
ISRA	Institut Sénégalais de Recherche Agricole
ML	Maximum Likelihood
MSEP	Mean Square Error of Prediction
NIPALS	Nonlinear estimation by Iterative Partial Least Squares
NIRS	Near Infrared Reflectance Spectroscopy
OLS	Ordinary Least Squares
ONG	Organisation Non Gouvernementale
PLS	Partial Least Squares
PRESS	PRediction Error Sum of Squares
R3S	Réseau de Recherche sur la Résistance à la Sécheresse
REML	Restricted or Residual Maximum Likelihood
RUE	Radiation Use efficiency
SARRA	Système d'Analyse Régional des Risques Agroclimatiques
SARRAH	Système d'Analyse Régional des Risques Agroclimatiques - version Habillée
WUE	Water Use efficiency

Table des matières

Chapitre 1

Introduction générale

1.1 Problématique

Au Sahel, les précipitations annuelles ont diminué de l'ordre de 20 à 30 % dans la dernière moitié du 20^{e} siècle (Baterburry et Warren, 2001) et leur déficit y demeure la plus forte contrainte à l'agriculture (Tucker, 1991).

Dans de telles conditions, les paysans, qui constituent la majorité de la population, parviennent difficilement à générer des revenus réguliers tout en gérant durablement les ressources naturelles. L'agriculture, qui est pourtant le principal moteur du développement économique et social du Sahel, ne peut ainsi jouer pleinement son rôle.

Dans ce cadre difficile, le rôle des décideurs (États, ONG, Organismes de coopération, Recherche) est de répondre à la demande sociale, principalement "la promotion d'une agriculture productive, diversifiée, durable [...]". Ce sont les termes employés dans le Cadre stratégique de sécurité alimentaire (CSSA), le document de référence en matière de sécurité alimentaire du Comité permanent inter-états de lutte contre la sécheresse au Sahel (CILSS).

Ce cadre stratégique qui vise aussi "l'amélioration durable des conditions d'accès des groupes et zones vulnérables à l'alimentation", a été traduit en programmes de sécurité alimentaire pour la plupart des pays sahéliens (CILSS, 2000). Ce document est venu confirmer le mandat déjà attribué au Centre d'étude régional pour l'amélioration de l'adaptation à la sécheresse (CERAAS) par le Réseau de recherche sur la résistance à la sécheresse (R3S) du CILSS.

Le CERAAS est un laboratoire national de recherche à vocation régionale sous double tutelle. C'est un laboratoire de l'Institut sénégalais de recherches agricoles (ISRA), qui a la mission d'exécuter le programme national de recherche sur l'adaptation des plantes à la sécheresse. Il est aussi une Base Centre du Conseil ouest et centre africain pour la recherche et le développement agricole (CORAF), chargé de conduire les recherches sur la thématique de l'adaptation des plantes à la sécheresse et celle de la création variétale.

Le CERAAS conduit des recherches suivant quatre axes principaux :

1. la compréhension de la réponse des plantes ;
2. la modélisation du fonctionnement des plantes ;
3. l'amélioration de la méthodologie de la sélection ;
4. l'amélioration des systèmes de culture pour une meilleure adaptation à la sécheresse.

Au troisième axe, l'objectif est d'identifier et de sélectionner du matériel végétal mieux adapté à la sécheresse, stabilisant ainsi le déficit alimentaire dans les pays des régions sèches. Pour atteindre cet objectif, les activités de recherche visent à fournir des solutions techniques pour réduire l'effet dépressif de la sécheresse sur les productions agricoles. Ces solutions consistent à proposer des méthodes de sélection, de suivi des cultures et des itinéraires techniques tenant compte du milieu ciblé, qui est soumis à la contrainte hydrique.

La sélection porte sur des caractères phénologiques, physiologiques et moléculaires permettant une production améliorée en conditions de déficit hydrique par rapport aux cultivars généralement vulgarisés. Elle est validée par des essais en plein champ, réalisés dans des conditions qui reflètent la variabilité du milieu auquel les variétés sont destinées. Ces essais peuvent avoir lieu la même année, sur plusieurs endroits (essai multilocal) ou sur plusieurs années, au même endroit (essai pluriannuel) ou sur plusieurs années, sur plusieurs endroits. Par la suite, nous désignerons tous ces types d'essais sous le nom générique d'essai multienvironnement et emploierons ce terme chaque fois qu'il n'y aura pas d'ambiguïtés et que nous voulons nommer un ensemble d'essais.

Dans cette zone du Sahel à fort risque climatique, il est souvent constaté une variabilité de l'écart entre le rendement des génotypes lors de ces essais multienvironnements de sélection variétale. Cette variabilité est connue des sélectionneurs sous le nom d'interaction génotype $\times$ environnement (G$\times$E) à laquelle nous nous intéressons dans ce travail.

Un cas particulier d'interaction étant un changement de classement direct des génotypes. Si trois génotypes **A**, **B** et **C** sont testés sur plusieurs environnements, nous sommes en présence d'interaction G$\times$E si pour le premier environnement, les génotypes se classent **A**-**B**-**C** selon leur rendement et pour le deuxième environnement, ils se classent par exemple **B**-**A**-**C**. Dans ce cas et même dans le cas plus général où les différences entre génotypes dependent de l'environnement, il sera difficile pour un environnement cible où les génotypes n'ont pas encore été testés, de prédire le meilleur génotype.

Dune manière générale, les essais multienvironnements ne peuvent être assez précis. En effet, en dehors d'une ou de deux variétés témoin généralement reconduites dune année à lautre, chaque variété n'est vue que deux à cinq ans. Au regard de la forte interaction génotype $\times$ année (G$\times$A), ces deux à cinq ans sont un échantillon de taille trop faible. Le sélectionneur doit

souvent extrapoler à partir de ce nombre d'années faible ou d'un essai multi-local où l'interaction génotype × lieu (G×L) ne reproduit qu'imparfaitement l'interaction G×A.

Les interactions G×E gênent donc la sélection variétale et constituent un obstacle aux recommandations éventuellement formulées aux paysans pour l'adoption de cultivars adaptés à leurs milieux. Une solution est de modéliser ces interactions G×E dans le but de les prédire pour une situation nouvelle en fonction de variables environnementales dont on connaît la valeur (ex : nature et profondeur du sol) ou la loi de probabilité (ex : précipitations) (Piepho, Denis et van Eeuwijk, 1998).

Nous proposons alors dans ce travail, une méthode de modélisation des interactions G×E, qui permette de tenir compte de l'impact aléatoire de l'environnement induit principalement au Sahel par une variabilité climatique, pour une meilleure prédiction de la réponse des variétés.

Plusieurs méthodes d'analyse des interactions ont été proposées dans la littérature et sont exposées au chapitre 2. Ces méthodes peuvent être rangées, à notre sens, en deux catégories : celles qui utilisent les caractéristiques des environnements et celles qui ne les utilisent pas. Pour ces dernières, dont nous pouvons citer la méthode *Additive Main effects and Multiplicative Interactions* (AMMI) ainsi que la régression conjointe, la critique majeure est qu'elles ne tiennent pas compte justement des environnements cibles pour y prédire le rendement des génotypes. Ce n'est pas adapté dans cette zone du Sahel, car comme nous avions annoncé, les interactions G×E y sont la conséquence principalement de la grande variabilité climatique des environnements. Mais ce qui gêne en réalité, c'est la présence de l'interaction G×A qui est imprévisible, au contraire de l'interaction G×L. Talbot (1997) a établi que l'interaction G×A×L est plus importante que l'interaction G×A, elle-même plus importante que l'interaction G×L. C'est certainement le cas au Sahel, où la variabilité climatique interannuelle est forte. En effet, s'il pouvait

y être possible, pour chaque lieu, de prévoir les conditions climatiques d'une année sur l'autre, il suffirait de mener un essai multilocal sur un ensemble de lieux représentatifs et avec la méthode AMMI par exemple, pouvoir prédire avec assez de précision le rendement des cultures. Mais les importantes variations climatiques d'une année sur l'autre empêchent cette prédiction sans passer par la prise en compte des conditions du milieu.

Parmi les méthodes qui utilisent les caractéristiques environnementales, figure la régression factorielle, dont une limite est qu'elle suppose une action linéaire des environnements sur le rendement. Dans le contexte sahélien, cette méthode utilise le modèle d'analyse de variance à deux facteurs, génotype et environnement, où les interactions G×E sont expliquées par des covariables climatiques mesurées souvent au pas de temps décadaire voire journalier sur chacun des environnements et des covariables mesurées sur chacun des génotypes. En général, les variables climatiques mesurées sur les environnements sont très nombreuses (séries temporelles) et la prise en compte de l'ensemble d'entre elles par cette méthode est impossible.

Les modèles de simulation de cultures sont aussi utilisés comme méthode de prédiction de rendement des cultures tenant compte de l'environnement. Ces modèles ont certes l'avantage d'être plus réalistes et considèrent le rendement d'un génotype dans un environnement particulier comme une fonction non linéaire des paramètres du génotype et des caractéristiques de l'environnement. Ils présentent cependant l'inconvénient de ne pas être applicables à tout génotype. En effet, les paramètres de tels modèles de simulation de cultures ne sont pour la plupart connus que pour un petit nombre de génotypes, car leur évaluation demande une expérimentation spécifique et des mesures coûteuses.

La première méthode proposée : APLAT. Lors d'essais multienvironnements, figure généralement un génotype de référence dont les paramètres

sont bien connus, dans le but de comparer sa performance aux autres génotypes. Tenant compte de l'information souvent disponible pour ce génotype de référence, nous proposons notre première méthode d'estimation qui consiste à linéariser le rendement des génotypes prédit par un modèle de simulation de cultures autour du vecteur de paramètres de ce génotype de référence. Le but étant d'estimer les paramètres de ces génotypes à l'aide des résultats d'essais multienvironnements, sans refaire le travail de "paramétrisation" en station expérimentale nécessaire à l'estimation de ceux du témoin. Cela permet de se ramener approximativement à un modèle linéaire où la matrice des variables explicatives est remplacée par la matrice des dérivées partielles (sensibilités) par rapport aux paramètres.

Cette méthode appelée Approximation par linéarisation autour d'un témoin (APLAT) est décrite au chapitre 3. Pour estimer ainsi la performance de tout génotype i dans un environnement j, nous adjoignons à la performance de ce génotype i prédite par un modèle de simulation de cultures pour l'environnement j linéarisée autour du vecteur de paramètres du témoin, un biais de ce modèle qui ne dépend que de l'environnement et une erreur aléatoire résiduelle. Pour que cette méthode ait un intérêt, il faut que les interactions G×E soient bien reproduites par le modèle de simulation de culture, et que la méthode d'estimation supporte les abandons de génotypes au cours du temps comme cela se pratique habituellement. C'est pourquoi, elle a été testée sur les données d'un essai pluriannuel mené sur la station expérimentale du CERAAS au Sénégal où tous les génotypes n'étaient pas observés tous les ans.

Pour la plupart des modèles de simulations de cultures, il existe un nombre important de paramètres pour les génotypes. Ces paramètres, généralement connus pour le génotype de référence, et que nous cherchons à réestimer pour tout nouveau génotype, conduisent à un nombre important de régresseurs

pour notre méthode proposée. Ils ont été estimés à l'aide la régression *Partial least square* (PLS).

La deuxième méthode proposée : APLAT-Mixte. Au chapitre 4, nous étendons la méthode APLAT au cas d'essais à plusieurs composantes de variance. Pour cette méthode, nous estimons qu'un modèle de simulation de cultures ne permet pas de prendre en compte totalement l'effet aléatoire des interactions G×E, même si nous pouvons concevoir qu'il le permette pour une grande part. Nous rajoutons alors au modèle APLAT un effet résiduel de l'environnement et un effet des interactions G×E aléatoires, dont il faudra estimer les composantes de variance.

Le recours à un modèle de simulation de culture induit, nous l'avons vu, un nombre important de régresseurs. Comme il s'agit également d'estimer la variance de l'effet de l'environnement et de l'effet des interactions G×E supposés aléatoires, nous nous retrouvons avec un modèle avec un nombre important de régresseurs et des composantes de variance. Nous proposons donc dans ce chapitre, une méthode originale d'estimation des paramètres fixes et des composantes de variance dans un modèle où le nombre de régresseurs est important par rapport aux observations. Cette méthode d'estimation, dénommée APLAT-Mixte, se fait par le principe d'une méthode combinée de réduction de dimension et de modèle mixte, que nous avons appelé PLS-Mixte.

Considérant d'une part les algorithmes itératifs d'estimation des paramètres inconnus dans le cadre du modèle mixte, et d'autre part les techniques de réduction de dimension, nous proposons d'imbriquer la régression PLS dans l'algorithme EM. Puisque nos données d'interaction G×E s'appréhendent à l'aide d'un modèle où les erreurs aléatoires sont corrélées, nous appliquons dans un premier temps cette technique à des données de NIRS (*Near infrared spectroscopy*) avec des erreurs indépendantes, où nous ne nous occupons que

de la double contrainte de la dimension du modèle et de la présence des composantes de variance. Par la suite, nous nous intéressons à résoudre le problème supplémentaire des erreurs corrélées qui résultent des données de notre problématique d'interaction G×E.

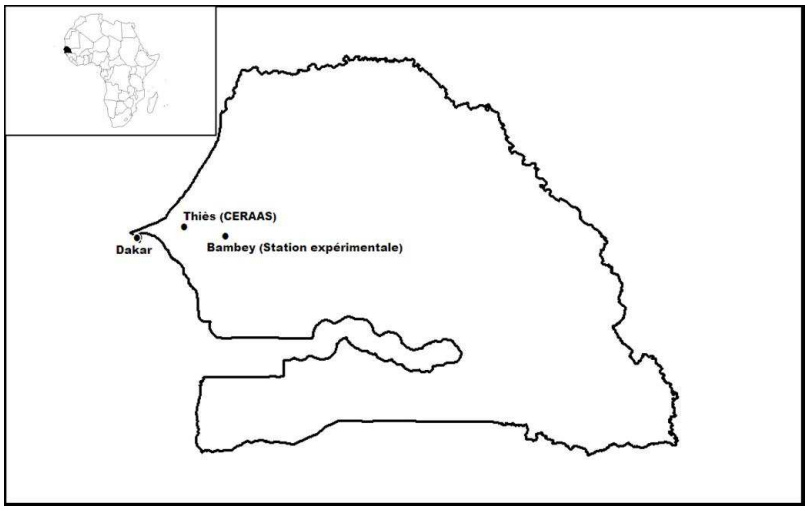
1.2 Présentation des données de l'étude

Les méthodes qui ont été proposées ici se sont appuyées sur des résultats d'essais multienvironnements. Ces expérimentations variétales, sur quoi finalement repose toute cette évaluation du comportement des génotypes en rapport avec les caractéristiques du milieu, sont la base de la sélection variétale. Les objectifs de tels essais peuvent être divers et dépendent des questions auxquelles l'expérimentateur veut apporter des réponses. Mais l'objectif principal est la comparaison des performances, souvent le rendement, de différents génotypes au regard de différentes conditions environnementales.

Plusieurs génotypes sont en général testés dans plusieurs environnements et ces derniers sont choisis avec un souci qu'ils soient aussi divers que possible. Le but recherché est de couvrir la quasi totalité des types d'environnements susceptibles de recevoir les génotypes. Ce but est difficile à atteindre au Sahel tant sont variables, nous l'avons dit, les conditions environnementales d'une année à l'autre.

Les données utilisées dans le cadre de cette étude proviennent de deux types d'expérimentations. D'une part, nous nous sommes servi des résultats d'un essai pluriannuel, mené à la station expérimentale du CERAAS, de 1994 à 1998. D'autre part, un essai multilocal a été mis en place durant l'hivernage 2005, sur 11 localités. L'hivernage désigne la saison des pluies dans les régions tropicales. Au Sénégal, il a lieu chaque année entre mai et octobre et sa durée varie de trois à six mois selon un gradient Nord-Sud.

FIG. 1.1 – Localisation du CERAAS et de la station expérimentale de Bambey au Sénégal.



Toutes ces expérimentations qui ont donc été menées au Sénégal, ont concerné l'arachide (*Arachis hypogaea* L.), qui y demeure la principale culture de rente.

1.2.1 L'essai pluriannuel

Cet essai variétal n'a concerné qu'un seul site, la station expérimentale du CERAAS située à Bambey (14 ° 42 N et 16 ° 28 O). Il est dit pluriannuel et a été mené de 1994 à 1998 (figure 1.1).

Cet essai a été conduit avec 26 génotypes à cycle de développement de 90 jours. Les rendements moyens par génotype et par année sont présentés dans le tableau 1.1. Tous les génotypes ne sont pas observés toutes les années, ce qui est habituel pour les programmes de sélection variétale. En effet, le sélectionneur a le loisir au cours de tels programmes, d'enlever certains génotypes s'il est avéré qu'ils donnent de faibles productions après une ou deux années d'expérimentation. De même, au cours d'un même programme,

le sélectionneur peut tout aussi bien faire entrer de nouveaux génotypes, toujours dans le but de les tester.

Génotype	1994	1995	1996	1997	1998	Moyenne
55-113		1 337,4	1 114,9	351,1	937,9	935,3
55-140		1 236,5	908,2		931,8	1 025,5
55-15		1 234,4	1 052,6			1 143,5
55-16		1 268,4	937,7			1 103,1
55-17		1 246,8	991,6			1 119,2
55-437†	571,3	1 081,0	707,1	421,2	1 074,6	771,1
57-111				401,7	964,7	683,2
57-115				556,5	977,9	767,2
57-120				452,4	1064,0	758,2
57-123				596,7	985,7	791,2
57-125				651,6	1 110,8	881,2
57-126				665,2	1 232,3	948,8
57-14			649,4	244,1		446,8
FLEUR11	1 420,1	1 624,9	1 065,3	941,6	1 098,0	1 230,0
S-45		1 194,3	936,6			1 065,5
S-46			1 001,5	641,4	1 081,8	908,2
SR-1-1	657,8					657,8
SR-1-11	492,0					492,0
SR-1-12	437,7					437,7
SR-1-2	772,7	1 170,1				971,4
SR-1-22	692,2	1 113,3	934,7	474,4		803,7
SR-1-23	582,5					582,5
SR-1-4	563,0	1 205,6	870,9			879,8
SR-1-6	672,8					672,8
SR-1-9	572,7					572,7
US-83	957,7	1 307,0				1 132,4
Moyenne	699,4	1 251,7	930,9	533,2	1 041,8	888,8

† génotype de référence

TAB. 1.1 – Données de l'essai pluriannuel de 26 génotypes d'arachide à la station expérimentale de Bambey au Sénégal de 1994 à 1998 (rendement en kg ha^{-1}).

Le génotype de référence choisi est le *55-437*, une variété de 90 jours ; sa longueur de cycle est la même que celles des autres génotypes. Les données de ce génotype sont disponibles pour la durée totale des expérimentations, ce qui autorise sa comparaison aux autres variétés utilisées dans cet essai.

Le tableau 1.1 révèle que les génotypes ont des productions moyennes très variées et certains font même plus que doubler leur production d'une année sur l'autre. Par exemple, le rendement de la variété *55-437* est de 1 081 kg ha^{-1} en 1995 alors qu'il n'est que de 571,3 kg ha^{-1}en 1994, soit une augmentation d'un peu moins de 90% entre deux années consécutives.

Les productions moyennes par année sont aussi très contrastées même si d'une année sur l'autre, nous n'avons pas les mêmes génotypes. L'année 1995 (1251,7 kg ha^{-1}) est l'année la plus productive et les génotypes qui ont les rendements les plus élevés sur l'ensemble des cinq années ont tous été observés cette année. Il s'agit des génotypes *FLEUR11* avec 1 230 kg ha^{-1}, *55-15* avec 1 143,5 kg ha^{-1}, *US-83* avec 1 132,4 kg ha^{-1}, etc. L'année 1997 (533,2 kg ha^{-1}) est l'année la moins productive.

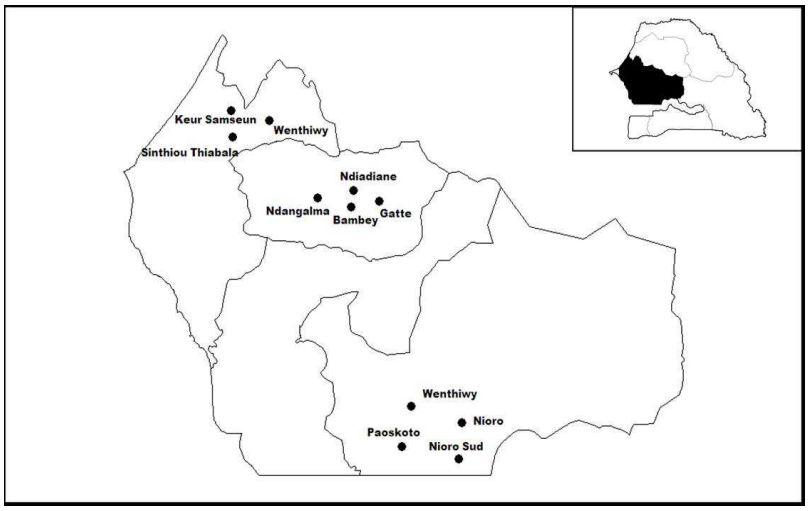
1.2.2 L'essai multilocal

Ces essais ont été menés durant l'hivernage 2005 et visent à mesurer la production et la qualité des graines dans des conditions environnementales différenciées et connues afin de modéliser les interactions G×E.

Pour ces essais, 6 génotypes (*55-437*, *FLEUR11*, *GC 8-35*, *JL24*, *55-128*, *55-33*) ont été testés sur 11 localités. Ces localités ont été choisies dans la principale zone de culture de l'arachide au Sénégal, appelée bassin arachidier. Ce bassin a été divisé en trois zones à l'intérieur desquelles les essais ont été implantés : 3 essais dans la zone Nord, 4 dans la zone Centre dont un à la station expérimentale de Bambey et 4 dans la zone Sud dont un à la station expérimentale de Nioro (figure 1.2). Cet échantillonnage par stratification a été mis en place pour tenir compte de la variabilité climatique constatée dans ce pays selon le gradient Nord-Sud.

Dans chaque localité, un dispositif en blocs complets randomisés avec quatre répétitions, soit 24 parcelles, a été adopté. La parcelle élémentaire est constituée

FIG. 1.2 – Localisation des sites des essais multilocaux dans le bassin arachidier au Sénégal durant l'hivernage 2005.



par 5 lignes de 6 m avec un écartement de 50 cm entre les lignes et de 15 cm entre les poquets. Les trois lignes centrales constituent la parcelle utile à l'intérieur de laquelle les prélèvements de plantes ont été réalisés.

Le semis a été effectué dès l'installation effective de l'hivernage sur des parcelles avec un précédent cultural arachide ou mil. Pour les conditions expérimentales, un labour superficiel et un piquetage ont été effectués et le semis fait à la main quand le sol était humide ; deux graines traitées au granox ont été semées par poquet. De l'engrais NPK 6-20-10 à raison de 150 kg ha^{-1} a été appliqué à la levée. Un démariage à un pied par poquet a été fait vers 10 jours après le semis et enfin, des sarclo-binages et des traitements phytosanitaires ont été réalisés à la demande.

Les rendements en gousses et en fanes ont été mesurés et un suivi journalier des paramètres climatiques (température minimale et maximale, humidité relative minimale et maximale, vitesse du vent, durée d'insolation) effectué. Les rendements moyens par génotype et par localité sont présentés dans le tableau 1.2.

Les mesures des données climatiques pour les essais de Nioro et de Bambey sont celles des postes météorologiques de ces stations. Ces mesures ont servi aussi pour les sites autour de ces deux localités à l'exception des mesures de pluviométrie, réputées plus variables dans l'espace. La localité de Meckhé était équipée d'une station météorologique portable. Chacun des 11 sites a disposé d'un pluviomètre.

Localités	55-128	55-33	55-437	F11	GC-8-35	JL24	Moyenne
Keur Fary	263,0	209,3	189,0	392,5	281,7	239,3	262,5
Keur Samseun	103,7	91,5	119,2	260,9	141,8	151,7	144,8
Sinthiou Thiabala	317,9	198,1	244,0	459,9	275,8	554,2	341,7
Moyenne Nord	228,2	166,3	184,1	371,1	233,1	315,1	249,6
Bambey	248,1	196,6	298,7	472,3	134,0	350,8	283,4
Gatte	121,8	116,0	209,3	261,9	86,3	283,4	179,8
Ndangalma	162,3	202,1	223,9	237,5	183,6	313,8	220,5
Ndiadiane	72,6	107,4	234,9	236,9	159,9	368,7	196,7
Moyenne Centre	151,2	155,5	241,7	302,2	140,9	329,2	220,1
Nioro	928,0	866,3	807,0	856,5	614,1	780,7	808,8
Nioro Sud	1 593,6	1 013,0	670,7	1 115,8	1 115,8	1 255,4	1 127,4
Paoskoto	1 164,0	1 254,2	908,3	1 231,6	1 157,8	712,7	1 071,4
Winthewy	189,2	97,4	369,8	447,3	244,9	528,1	312,8
Moyenne Sud	968,7	807,7	688,9	912,8	783,1	819,2	830,1
Moyenne	469,5	395,6	388,6	543,0	399,6	503,5	450,0

TAB. 1.2 – Données de l'essai multilocal de 6 génotypes d'arachide sur 11 localités au Sénégal durant l'hivernage 2005 (rendement en kg ha^{-1}).

L'examen des données du tableau 1.2 révèle une variabilité de rendement des génotypes selon les trois principales zones de l'étude : le Nord avec une production moyenne de 249,6 kg ha^{-1} et le Centre avec une production moyenne de 220,1 kg ha^{-1} s'opposent au Sud qui affiche la meilleure production moyenne avec 830,1 kg ha^{-1}. A l'intérieur des zones, la variabilité relative est plus importante au Nord où le rendement du génotype le plus productif représente 123,2% de celui du génotype le moins productif et au Centre où ce pourcentage est de 133,6%, qu'au Sud où il passe à 40,6%.

Par ailleurs, il est à noter que le génotype *FLEUR11* qui a la meilleure production moyenne pour les trois zones (543,0 kg ha^{-1}), n'est pas le meilleur partout. Il ne donne le meilleur rendement qu'au Nord alors que le génotype *GC-8-35* domine au Centre et le génotype *55-128* au Sud. Ce constat renforce l'idée que dans cette zone sahélienne marquée par une variabilité environnementale importante, une recommandation pour le choix de génotypes spécifiques aux lieux est préférable et plus pertinente qu'une recommandation unique du génotype le meilleur en moyenne.

Chapitre 2

Les méthodes classiques d'analyse des interactions G×E

Dans ce chapitre, nous parlerons des outils classiques d'analyse des interactions G×E et présenterons les modèles de simulation de cultures comme méthode alternative pour prédire le rendement des cultures. Nous allons également soumettre nos données à ces différentes méthodes présentées et évaluer comment elles prennent en compte les éventuelles interactions décelées. Nous allons toutefois appliquer ces différents modèles uniquement sur les données de l'essai multilocal. En effet, la plupart des méthodes classiques nécessitent des données complètes et seules le sont celles de l'essai multilocal.

Les raisons de la présence des interactions G×E peuvent être de deux ordres. De telles interactions sont, d'une part, attendues en présence d'une large variation des caractéristiques de résistance aux stress des génotypes, le stress hydrique par exemple. D'autre part, en présence, d'une large variation des environnements au niveau de ce même stress. Mais généralement, c'est l'une et l'autre de ces conditions qui les favorisent, même si au Sahel, la grande

variabilité climatique y sévissant dont nous considérons qu'elle caractérise essentiellement les environnements, contribue pour une large part à la présence de ces interactions.

Le terme génotype fait référence à un cultivar, c'est-à-dire un matériel génétique qui peut être homogène ou hétérogène et l'environnement à un ensemble de conditions climatiques, de types de sol et de pratiques culturales d'un essai conduit dans un lieu donné, une année donnée (Annicchiarico, 2002).

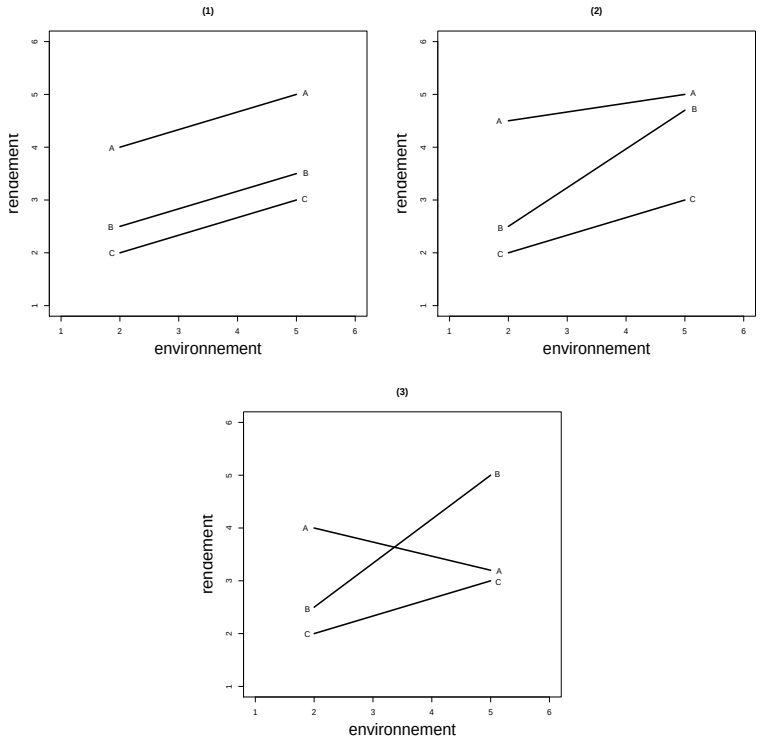
Deux types d'interactions G×E sont à distinguer (figure 2.1). Les interactions sont dites quantitatives ou *noncrossovers*, si les classements des génotypes entre les différents environnements sont conservés mais que l'écart entre les génotypes est modifié. Par contre, elles sont dites qualitatives ou *crossovers* lorsque les classements sont inversés (Baker, 1988 ; Baril, 1992).

Dans les essais multienvironnements, il peut être envisagé de sélectionner les génotypes de plus grande production moyenne sur l'ensemble des environnements testés ou de les choisir en fonction de leurs performances selon les environnements. Pour cela, les informations issues de ces expérimentations, sont étudiées afin d'être synthétisées en dissociant les effets du génotype, de l'environnement et des interactions G×E au travers des modèles statistiques (Brancourt-Hulmel, Biarnès-Dumoulin et Denis, 1997).

Plusieurs modèles des interactions G×E ont donc été proposés. Dans ce qui suit, nous en ferons un tour d'horizon et en présenterons les principaux : le modèle d'analyse de variance à deux facteurs, la régression conjointe, la méthode AMMI, la régression factorielle et les modèles de simulation de cultures.

Mais avant de présenter ces différentes méthodes fondées principalement sur le modèle d'analyse de variance, il est à remarquer qu'il est aussi possible de concevoir, à travers deux statistiques descriptives, l'étude des interactions

FIG. 2.1 – Types d'interactions G×E pour trois génotypes A, B et C. (1) : sans interactions ; (2) : interactions quantitatives ; (3) : interactions qualitatives.



G×E pour décrire le comportement des génotypes sur un échantillon d'environnements.

Pour cela, la variabilité intrinsèque du génotype sur un ensemble d'environnements est étudiée à l'aide de la variance environnementale S_i^2 (Becker, 1981 ; Lin, Binns et Lefkovitch, 1986 ; Piepho, 1998). L'écart à la valeur moyenne des performances du génotype, compte tenu du nombre de milieux sur lequel il est testé, représente une mesure de son instabilité. Cette variance environnementale est estimée par

$$S_i^2 = \sum_{j=1}^{J} (Y_{ij} - Y_{i.})^2/(J-1)$$

où Y_{ij} est la réponse du génotype i de l'environnement j, $Y_{i.}$ la moyenne des réponses du génotype i des différents environnements et J le nombre d'environnements. Par la suite, l'opérateur (.) désigne la moyenne sur l'indice qu'il remplace.

Quant à l'écovalence variétale W_i^2 (Becker, 1981 ; Becker et Léon, 1988), elle est mesurée par la stabilité relative du génotype et est estimée par

$$W_i^2 = \sum_{j=1}^{J} (Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})^2$$

C'est la somme des carrés des termes d'interaction propres au génotype i. A la différence de S_i^2, la somme de carrés W_i^2 n'est pas divisée par les degrés de liberté (ddl) correspondants.

Cependant, la liste des méthodes d'étude des interactions G×E présentée dans ce chapitre n'est pas exhaustive. D'autres méthodes, qui ne sont pas décrites ici, existent par ailleurs :

- structuration de l'interaction (Denis et Vincourt, 1982)
- modèles multiplicatifs (Cornelius, Seyedsadr et Crossa, 1992 ; Crossa, Cornelius, Seyedsadr et Byrne, 1993 ; Crossa, Cornelius, Sayre et Ortiz-Monasterio, 1995) ;
- application de l'analyse canonique (Seif, Evans et Balaam, 1979 ; Calinski, Czajkaet Kaczmarek, 1987) ;
- variantes des modèles de regression factorielle (Denis, 1988 ; van Eeuwijk 1992, 1995 ; van Eeuwijk, Denis et Kang, 1996) ;
- régression *Partial Least Squares* (Aastveit et Martens 1986 ; Talbot et Wheelwright 1989 ; Vargas, Crossa, Sayre, Reynolds, Ramìrez et Talbot, 1998) ;
- une méthode récente fondée sur l'approche bayésienne (Theobald, Talbot et Nabugoomu, 2002).

2.1 Le modèle d'analyse de variance à deux facteurs

2.1.1 Le modèle

Le modèle linéaire mixte généralement considéré sur les moyennes par génotype et par environnement est le suivant

$$Y_{ij} = m + g_i + E_j + (gE)_{ij} + e_{ij} \tag{2.1}$$

où Y_{ij} est la réponse du génotype i de l'environnement j, m la moyenne générale et g_i l'effet fixe du génotype i. L'effet E_j de l'environnement j, l'interaction $(gE)_{ij}$ et le terme d'erreur e_{ij} sont supposés aléatoires, iid et indépendants les uns des autres avec

$$\mathbb{E}(E_j) = \mathbb{E}[(gE)_{ij}] = \mathbb{E}(e_{ij}) = 0 \text{ et } \mathrm{Var}(E_j) = \sigma_E^2, \mathrm{Var}[(gE)_{ij}] = \sigma_{gE}^2 \text{ et } \mathrm{Var}(e_{ij}) = \sigma_e^2$$

où la fonction $\mathbb{E}(\cdot)$ désigne l'espérance et $\mathrm{Var}(\cdot)$ la variance.

Dans l'optique de prédire la performance des génotypes dans les différents environnements considérés, l'option qui consiste à prendre les génotypes comme fixes et les environnements comme aléatoires est argumentée par Denis, Piepho et van Euwijk (1997). En effet, ces auteurs justifient ce choix par le fait qu'il s'agit d'étudier un nombre fini de génotypes, d'où l'effet génotype fixe. Au contraire, les environnements ne sont pas considérés pour eux-mêmes, mais en tant qu'échantillons dans une population plus vaste d'environnements possibles auxquels les variétés sont destinées. Pour nous, cela s'appliquera aux années plutôt qu'aux lieux.

Les effets principaux du génotype et de l'environnement sont considérés par rapport à la moyenne générale, alors que le terme d'interaction du modèle représente la variabilité des performances du génotype avec l'environnement qui n'est pas prise en compte dans les effets additifs du génotype et de l'environnement.

D'après le modèle 2.1, les estimations des effets sont, pour un dispositif équilibré :

$$\widehat{g}_i = Y_{i.} - Y_{..}$$

$$\widehat{E}_j = Y_{.j} - Y_{..}$$

$$\widehat{(gE)}_{ij} = Y_{ij} - Y_{i.} - Y_{.j} + Y_{..}$$

Dans l'estimation des termes du modèle qui portent l'indice j, nous retrouvons $Y_{.j}$. Cette moyenne traduit le potentiel de l'environnement. Or l'environnement étant fortement variable au Sahel, les termes en j ne sont pas bien prévisibles à moins de disposer d'un échantillon de nombreux environnements qui fait généralement défaut. Cependant si nous considérons la différence entre deux variétés i et i', l'imprévisibilité de l'effet environnement E_j disparaît lors de l'estimation de cette différence, si le dispositif est complet. En effet, il viendra :

$$Y_{ij} - Y_{i'j} = g_i - g_{i'} + (gE)_{ij} - (gE)_{i'j}$$

Par contre, le problème demeure pour les interactions qu'il faudra modéliser afin de prédire plus finement la différence des performances des génotypes.

2.1.2 Illustration avec les données de l'essai multilocal

Avec le modèle d'analyse de variance à deux facteurs, génotype et environnement, appliqué aux données de l'essai multilocal, nous sommes intéressés

tout premièrement à tester la significativité des interactions G×E. Dans ce cas, les deux effets principaux sont considérés comme étant fixes.

Nous rappelons, que les données proviennent d'un réseau d'essais variétaux effectués au Sénégal durant l'hivernage 2005 (tableau 1.2, page 16). Six génotypes d'arachide ont été testés sur 11 sites dans le bassin arachidier sénégalais qui est la région principale de production de cette légumineuse.

Le tableau 2.1 fournit les résultats de l'analyse de variance à deux facteurs appliquée à ces données.

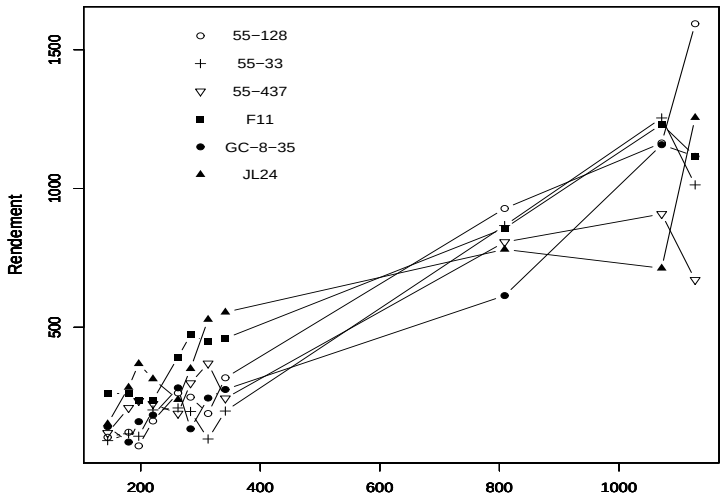
Effet	d.l	Somme de carrés	Carré moyen	Statistique F	Niveau de signification
Génotype	5	23 2765,0	46 553,0	2,42	0,0485
Environnement	10	8 100 921,7	3 810 092,2	42,1	0,0000
Résidus	50	962 311,6	19 246,2		

TAB. 2.1 – Tableau d'analyse de variance des données des essais multilocaux de 6 génotypes d'arachide sur 11 localités au Sénégal durant l'hivernage 2005.

En négligeant dans un premier temps l'interaction, nous concluons qu'au seuil de 5%, les effets génotype et environnement sont significatifs. A l'instar de Denis et Vincourt (1982), nous allons évaluer et comparer l'ordre de grandeur des résidus et l'ordre de grandeur de l'effet génotype. Si CM_r est le carré moyen des résidus, l'ordre de grandeur de ces résidus peut être estimé par $\sqrt{CM_r}$; et si CM_g est le carré moyen du facteur génotype, l'ordre de grandeur de l'effet génotype peut être estimé par $\sqrt{CM_g - CM_r}/J$. Nous notons alors que l'ordre de grandeur des résidus (138,7) est grand par rapport à celui de l'effet génotype (15). Il s'agira alors d'essayer de réduire ces résidus en ajoutant une interaction au modèle additif. L'interaction peut être mise en évidence avec la figure 2.2 où sont représentés les rendements des génotypes de l'essai multilocal. Les rendements moyens par lieu sont rangés par ordre croissant et mis en abscisse. Nous constatons sur ce graphique un changement

de classement des variétés d'un site à un autre. Nous remarquons également que l'écart entre les génotypes augmente avec la moyenne du lieu.

FIG. 2.2 – Variation des rendements des six génotypes de l'essai multilocal. En abscisse sont mis, par ordre croissant, les performances moyennes des lieux.



Pour espérer formuler tout de même des recommandations dans ce milieu très contrasté pour l'adoption de cultivars les mieux adaptés à chaque environnement, la solution consiste à tenter de réduire la part imprévisible de ces interactions en les modélisant ; ce qui peut se faire à travers différentes méthodes que nous allons présenter ci-dessous.

2.2 La méthode de régression conjointe

2.2.1 Le modèle

Ce modèle a été proposé pour la première fois par Yates et Cochran (1938). Il a ensuite été repris par plusieurs autres auteurs dont les principaux sont Finlay et Wilkinson (1963), Eberhart et Russell (1966) et Perkins et Jinks (1968).

Plutôt proposé pour les essais multilocaux que pour les essais pluriannuels, ce modèle pose l'effet des interactions G×E comme une fonction linéaire de l'effet E_j qui représente le potentiel du milieu j pour les génotypes ; une valeur positive de E_j traduisant un potentiel élevé tandis qu'une valeur négative, un faible potentiel.

L'interaction est de la forme :

$$(gE)_{ij} = c_i E_j + d_{ij}$$

où $\widehat{c}_i$ est le coefficient de la régression pour le génotype i ; cette pente caractérise la sensibilité différentielle du génotype i au milieu (Denis et Vincourt 1982). Le terme d_{ij} désigne la déviation du modèle, c'est-à-dire les interactions G×E résiduelles.

Ces coefficients $\widehat{c}_i$ et la performance moyenne des génotypes sont les paramètres d'intérêt pour une analyse de stabilité. En effet, la performance d'un génotype dans un lieu donné dépend de la performance moyenne des génotypes du lieu et des interactions G×E qui y sont espérées. Une valeur largement positive de $\widehat{c}_i$ associée à une performance moyenne d'un site relativement importante, concourent à caractériser ce site comme étant favorable. De l'autre côté, un site est considéré comme non favorable si la valeur de $\widehat{c}_i$ associée à la performance moyenne des génotypes de ce site est négative.

Les coefficients $\widehat{c}_i$ sont obtenus facilement :

$$\widehat{c}_i = \frac{\sum_j \widehat{E}_j \widehat{(gE)}_{ij}}{\sum_j \widehat{E}_j^2}$$

Un cas particulier de ce modèle a été développé par Tukey (1949) : les coefficients c_i s'écrivent Kg_i et l'interaction se réduit à Kg_iE_j.

2.2.2 Illustration avec les données de l'essai multilocal

Les données sont celles du tableau 1.2 de la page 16. Le tableau 2.2 présente les paramètres $\widehat{m} + \widehat{g}_i$ et $1 + \widehat{c}_i$ des différents génotypes des essais.

Génotype	$\widehat{m} + \widehat{g}_i$	$1 + \widehat{c}_i$
55-128	488,9	1,381
55-33	341,3	1,138
55-437	327,2	0,698
FLEUR11	636,0	0,963
GC-8-35	349,2	1,043
JL24	557,1	0,777

TAB. 2.2 – Paramètres estimés par régression conjointe de 6 génotypes d'arachide lors d'essais multilocaux au Sénégal durant l'hivernage 2005.

Le génotype *55-128* qui a la pente $(1 + \widehat{c}_i)$ la plus élevée, est plus sensible aux variations du milieu que la moyenne des génotypes des essais, suivi du génotype *55-33*. En revanche, le génotype de référence *55-437* et le génotype *JL24* semblent moins affectés que la moyenne aux variations du milieu. Quant aux génotypes *GC-8-35* et *FLEUR11* avec des pentes proches de l'unité, ils représentent par leurs valeurs, de bonnes indications des variations du milieu.

2.3 La méthode AMMI

2.3.1 Le modèle

La méthode AMMI, *Additive main effect and multiplicative interaction*, a été introduite par Williams (1952) et reprise par Gollob (1968), puis par Mandel (1961, 1971) et par Bradu et Gabriel (1978). Elle fut développée à l'origine pour les domaines du social et de la physique et son application à la recherche agricole a été proposée par Kempton (1984) et Zobel, Wright et Gauch (1988). Mais il faut attendre Gauch (1992) pour qu'elle devienne répandue. C'est une méthode assez générale et Gauch et Zobel (1996) ont souligné que son champ d'application potentiel va au delà de l'étude des méthodes d'interactions G×E. Beaucoup d'autres auteurs ont étudié les interactions G×E à l'aide de cette méthode :Vargas, Crossa, van Eeuwijk, Ramìrez et Sayre (1999), Yan et Hunt (2001), Ebdon et Gauch (2002a, 2002b), González, Crossa et Cornelius (2003a, 2003b), Gauch (2006).

La méthode AMMI associe l'analyse de variance et l'analyse en composantes principales (ACP). Sont d'abord estimés les effets principaux des variétés et des environnements par une analyse de variance du modèle additif, c'est-à-dire du modèle sans les interactions G×E. Ensuite, la partie non additive du modèle est étudiée par une analyse en composantes principales (Crossa, 1990).

L'interaction est décrite de cette façon

$$(gE)_{ij} = \theta_1\alpha_{1i}\beta_{1j} + \theta_2\alpha_{2i}\beta_{2j} + \cdots + \theta_h\alpha_{hi}\beta_{hj}$$

c'est à dire

$$Y_{ij} = m + g_i + E_j + (\sum_{k=1}^{h} \theta_k\alpha_{ki}\beta_{kj}) + e_{ij}$$

où θ_k est la valeur singulière du k^{e} axe (θ_k^2 étant la valeur propre), α_{ki} est le vecteur propre du i^{e} génotype pour le k^{e} axe, β_{kj} est le vecteur propre du j^{e} environnement pour le k^{e} axe et avec les contraintes

$$\sum_i \alpha_{ki}^2 = \sum_j \beta_{kj}^2 = 1 \quad \text{et} \quad \sum_i \alpha_{ki}\alpha_{k'i} = \sum_j \beta_{kj}\beta_{k'j} = 0$$

Les paramètres du terme d'interaction sont donc estimés par décomposition en valeur singulière (DVS) de la matrice des résidus obtenus après ajustement des deux effets principaux.

Le nombre h de paramètres pour chaque terme multiplicatif de l'interaction qui constitue le nombre d'axes principaux peut être déterminé soit par validation croisée (Gauch et Zobel, 1988 ; Crossa, 1990, Piepho, 1994) où les répétitions sont tour à tour retirées et non les lieux, soit par des tests statistiques (Gollob, 1968 ; Cornelius, 1993 ; Piepho, 1995).

Le nombre d'axes principaux retenu est généralement compris entre zéro, on parle dans ce cas de AMMI-0 c'est-à-dire du modèle additif et le minimum entre $(I-1)$ et $(J-1)$, où I constitue le nombre de génotypes et J le nombre d'environnements. Le modèle complet (AMMI-F, F faisant référence à *full* pour *full model*), avec tous les axes principaux, fournit une estimation parfaite. Mais généralement, lorsque les interactions G×E sont significatives, les modèles avec un (AMMI-1) ou deux (AMMI-2) axes principaux sont les plus utilisés à cause de leur simplicité.

Les tests statistiques proposés pour déterminer le nombre optimal d'axes sont tous fondés sur la statistique t_k^2/s^2 où t_k^2 est l'estimation de la valeur singulière θ_k obtenue par DVS et s^2, avec f degrés de liberté, est le carré moyen des résidus du modèle additif divisé par le nombre de répétitions par environnement (Piepho, 1995).

Sous l'hypothèse nulle ($\text{H}_0 : \theta_k = 0$), les statistiques de tests sont les suivantes :

1. t_k^2/s^2 suit une loi de Fisher à $(J+I-1-2k)$ et f degrés de liberté (Gollob, 1968)
2. $F_{GH1} = gt_k^2/h_1 f s^2$ suit une loi de Fisher avec h_1 et g degrés de liberté, où $h_1 = 2v_1u_1/v_2$, $g = 2 + 2(f-2)v_1/v_2$, $v_1 = u_2^2 + u_1^2 + (f-4)u_1$ et $v_2 = (f-2)u_2^2 + 2u_1^2$ (Cornelius, 1993) ; u_1 et u_2 sont des approximations fournies par Cornelius (1980).
3. $F_{GH2} = t_k^2/u_1 s^2$ est distribuée selon la loi de Fisher avec h_2 et f degrés de liberté, où $h_2 = 2u_1^2/u_2^2$ (Cornelius, 1993).

En outre, une statistique de test F_R, plus simple à calculer, a été proposée par Cornelius et al. (1992). Ce test utilise la somme des carrés résiduels après ajustement du modèle AMMI avec q axes principaux. Sachant que sous l'hypothèse nulle, la somme de carrés résiduels est approximativement une variable de chi-deux, la statistique suivante F_R

$$F_R = \left[\sum_i \sum_j (Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})^2 - \sum_{k=1}^{q} t_k^2 \right] / f_2 s^2$$

suit une distribution de Fisher avec $f_2 = (J-1-q)(I-1-q)$ et f degrés de liberté. La statistique F_R significative révèle qu'il y a au moins un axe principal supplémentaire à prendre en compte en plus des q déjà utilisés.

Dans le cas des essais multienvironnements, il est supposé que les erreurs du modèle sont indépendantes et normalement distribuées avec des variances entre les environnements homogènes. Même si l'indépendance des erreurs peut être assurée par une randomisation des niveaux des facteurs et que les erreurs peuvent être considérées comme gaussiennes, les variances résiduelles sont généralement hétérogènes d'un environnement à l'autre au Sahel. Or, en présence d'erreurs expérimentales hétérogènes, Piepho(1995) a montré que la statistique de test F_R est plus robuste que F_{GH1}, F_{GH2} et celle de Gollob.

2.3.2 Illustration avec les données de l'essai multilocal

Les données de l'essai multilocal (6 génotypes et 11 sites) ont été soumises à la méthode AMMI dans le but d'obtenir l'estimation des performances des génotypes dans les différents environnements. Il est espéré que l'estimation d'un génotype dans un environnement particulier soit plus précise que la simple moyenne des performances de ce même génotype dans les autres environnements où il a été observé.

Pour ces données, le nombre maximum d'axes principaux à retenir est égal à cinq. Pour cela, cinq modèles (AMMI-0 à AMMI-4) vont être testés étape par étape à l'aide de la statistique de test F_R. Si le modèle AMMI-1 est significatif, c'est-à-dire la statistique associée au premier axe principal est significative, le modèle AMMI-2 est alors testé à son tour et ainsi de suite jusqu'au modèle AMMI-q non significatif où l'on devra s'arrêter. Le meilleur modèle est alors le modèle AMMI-q.

Les valeurs singulières calculées sur les données de l'essai multilocal sont égales à 723,8 pour le premier axe, 577,6 pour le second, 259,2 pour le troisième, 182,1 pour le quatrième et 66,5 pour le cinquième tandis que le tableau 2.3 montre les vecteurs propres des facteurs pour les quatre premiers axes principaux

Les résultats de la méthode AMMI sont présentés au tableau 2.4. Le test pour le deuxième axe principal est significatif, ce qui signifie qu'il y a au moins un axe supplémentaire intéressant dont il faut tenir compte. Et comme le test pour le troisième axe n'est pas significatif, nous adoptons, pour ces données, le modèle AMMI-3. A la figure 2.3 sont représentés les scores des génotypes et des environnements du deuxième axe principal en fonction de ceux du premier. Ce graphique double, est plus connu sous le nom de *biplot* (Kempton, 1984). Sur ce graphique, un génotype proche de l'origine présente

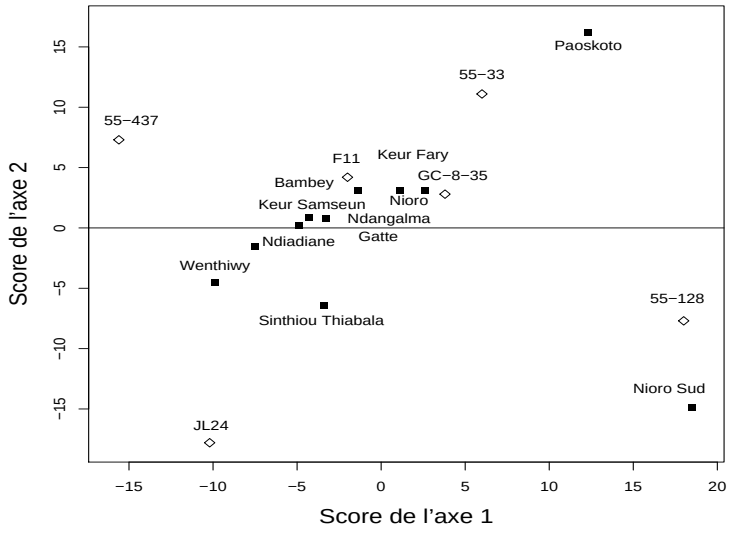
	Axe 1	Axe 2	Axe 3	Axe 4
Vecteurs propres des génotypes pour les 4 axes				
55-128	18.0	-7.7	-5.5	1.8
55-33	6.0	11.1	-5.0	-5.0
55-437	-15.6	7.3	-5.6	-0.8
F11	-2.0	4.2	3.7	11.2
GC-8-35	3.8	2.8	12.6	-4.9
JL24	-10.2	-17.8	-0.2	-2.3
Vecteurs propres des environnements pour les 4 axes				
Wenthiwy	-9.9	-4.5	3.2	3.3
Niorosud	18.5	-14.9	0.5	-0.4
Paoskoto	12.3	16.2	3.9	0.5
Nioro	2.6	3.1	-13.2	-0.2
Keur Samseun	-1.4	3.1	3.2	-0.1
Keur Fary	1.1	3.1	4.6	1.2
Sinthiou Thiabala	-3.4	-6.4	2.9	2.1
Ndiadiane	-7.5	-1.5	1.4	-6.1
Gatte	-4.9	0.2	-2.4	-1.1
Ndangalma	-3.3	0.8	-0.9	-7.6

TAB. 2.3 – Vecteurs propres des facteurs génotype et environnement pour l'essai multilocal.

une faible interaction tandis qu'un génotype qui s'en éloigne est au contraire interactif.

Comme avec la régression conjointe, les génotypes *55-33* et *55-128* présentent de fortes interactions et les génotypes *GC-8-35* et *FLEUR11* semblent les moins interactifs c'est-à-dire les plus conformes au comportement moyen de l'ensemble des génotypes. DE même, les variétés *JL24* et *55-437* sont décrites comme étant moins aptes à profiter des conditions favorables du milieu

FIG. 2.3 – Scores des génotypes et des environnements pour le deuxième axe en fonction de ceux du premier.



2.4 La régression factorielle

2.4.1 Le modèle

C'est le modèle additif auquel trois séries de termes de régression sont ajoutés (Denis, 1980). Ces régresseurs sont des covariables décrivant les génotypes (ex : poids des graines) et les environnements (ex : pluviométrie en août) et aussi l'interaction par les produits deux à deux de ces régresseurs.

Un modèle similaire de celui de régression factorielle a été utilisé par Freeman et Perkins (1971) qui n'avaient considéré qu'une covariable associée au facteur milieu non calculée sur les données. Wood (1976), lui, a utilisé une combinaison linéaire de covariables élémentaires liées au milieu.

Effet	d.l	Somme de carrés	Carré moyen	Statistique F	Niveau de signification
Génotype	5	23 2765,0	46 553,0	2,42	0,0485
Environnement	10	8 100 921,7	3 810 092,2	42,1	0,0000
G×E	50	962 311,6	19 246,2		
AMMI axe 1	14	523 953,0	37 425,2	3,1	0,003
AMMI axe 2	12	333 578,0	27 798,1	6,4	0,000
AMMI axe 3	10	67 186,2	6 718,62	2,5	0,057
AMMI axe 4	8	33 172,6	4 146,6	5,6	0,026
Résidus G×E	6	4 422,5			
Total	65	929 600			

TAB. 2.4 – Tableau d'analyse des données de l'essai multilocal avec la méthode AMMI.

Nous allons présenter le modèle de régression factorielle de deux manières différentes. Dans la présentation matricielle, la plus simple, le but est d'expliquer la matrice des observations $\mathbf{Y}$ de dimension $I \times J$ à partir de deux matrices de covariables associées aux deux facteurs étudiés. Dans la présentation indicielle, la plus générale, les observations $\mathbf{Y}$ sont dans un vecteur de longueur IJ.

Nous pouvons disposer en général d'un tableau $\mathbf{X}$ de p covariables associées aux génotypes et d'un tableau $\mathbf{Z}$ de q covariables associées aux environnements. Les matrices $\mathbf{X}$ et $\mathbf{Z}$ s'écrivent dans ce cas :

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_2^1 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & & \vdots \\ x_I^1 & x_I^2 & \cdots & x_I^p \end{pmatrix} \qquad \mathbf{Z} = \begin{pmatrix} z_1^1 & z_1^2 & \cdots & z_1^q \\ z_2^1 & z_2^2 & \cdots & z_2^q \\ \vdots & \vdots & & \vdots \\ z_J^1 & z_J^2 & \cdots & z_J^q \end{pmatrix}$$

La matrice $\mathbf{X}$ est de dimension $I \times p$ et $\mathbf{Z}$ de dimension $J \times q$. Par souci de simplicité dans la présentation, nous ne considérons alors que le cas d'une unique observation par génotype pour un environnement donné, qui peut être une moyenne ou une moyenne ajustée sur le dispositif de cet environnement. Les matrices $\mathbf{X}$ et $\mathbf{Z}$ sont considérées par la suite centrées.

Alors le modèle de régression factorielle peut se présenter sous forme matricielle, plus commode à manipuler

$$\mathbf{Y} = \mathbf{1}_I m \mathbf{1}'_J + (\mathbb{g}_1 + \mathbf{X}\mathbb{g}_2)\mathbf{1}'_J + \mathbf{1}_I(\mathbb{E}'_1 + \mathbb{E}'_2\mathbf{Z}') + \mathbf{X}\mathbf{\Gamma}\mathbf{Z}' + \mathbb{g}_3\mathbf{Z}' + \mathbf{X}\mathbb{E}'_3 + \mathbf{e} \qquad (2.2)$$

- $\mathbf{Y}$ de dimension $I \times J$ est la variable réponse
- m est la moyenne générale
- $(\mathbb{g}_1 + \mathbf{X}\mathbb{g}_2)\mathbf{1}'_J$ est l'effet principal du génotype où ($\mathbb{g}_2$ est le vecteur des coefficients des covariables 1 à p dans la régression de l'effet génotype et ($\mathbb{g}_1$ les I résidus de cette régression.
- $\mathbf{1}_I(\mathbb{E}'_1 + \mathbb{E}'_2\mathbf{Z}')$ est l'effet principal de l'environnement décomposé de la même façon
- $\mathbf{X}\mathbf{\Gamma}\mathbf{Z}'$ est la partie de l'interaction expliquée par le produit des deux covariables $\mathbf{X}$ et $\mathbf{Z}$
- $\mathbb{g}_3\mathbf{Z}'$ est la partie de l'interaction expliquée par la covariable $\mathbf{Z}$, une fois tenu compte de l'explication fournie par la covariable $\mathbf{X}$
- $\mathbf{X}\mathbb{E}'_3$ est la partie de l'interaction expliquée par la covariable $\mathbf{X}$, une fois tenu compte de l'explication fournie par la covariable $\mathbf{Z}$

Ce modèle étant linéaire, les procédures d'estimation usuelles sont employées, ce qui donne les résultats suivants pour l'estimation des paramètres inconnus (voir Annexe A) :

- $\widehat{m} = \mathbf{1}'_I\mathbf{Y}\mathbf{1}_J/IJ$
- $\widehat{\mathbb{g}}_2 = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}\mathbf{1}_J/J$
- $\widehat{\mathbb{E}}'_2 = \mathbf{1}'_I\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}/I$
- $\widehat{\mathbb{g}}_1 = (\mathbf{I}_I - \mathbf{1}_I\mathbf{1}'_I/I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}')\mathbf{Y}\mathbf{1}_J/J$
- $\widehat{\mathbb{E}}'_1 = \mathbf{1}'_I\mathbf{Y}(\mathbf{I}_J - \mathbf{1}_J\mathbf{1}'_J/J - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}')I$
- $\widehat{\mathbf{\Gamma}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}$
- $\widehat{\mathbb{g}}_3 = (\mathbf{I}_I - \mathbf{1}_I\mathbf{1}'_I/I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}')\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}$

– $\widehat{\mathbb{E}}'_3 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{I}_J - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')$

Cependant, cette façon d'écrire le modèle et les calculs qui en découlent sont liés à la structure des données. En effet, dans ce que nous avons présenté, les observations sont supposées être dans une matrice à I lignes et J colonnes, ce qui prévoit cette matrice de données complète. Cela s'interprète dans le cas d'un essai multienvironnement par exemple, par le fait que chaque génotype doit être présent dans chaque environnement. Or la réalité est souvent autre.

Pour les essais où toutes les variétés ne sont pas dans tous les environnements, cette méthode, qui est ici généralisée avec p covariables associées au facteur génotype et q covariables liées au facteur environnement, peut se présenter sous forme indicielle.

$$Y_{ij} = m + (\sum_{p} x_i^p \beta_p + g_i) + (\sum_{q} z_j^q \alpha_q + E_j) + \sum_{p,q} x_i^p z_j^q \gamma_{pq} + e_{ij} \tag{2.3}$$

où g_i représente la part des effets moyens non expliquée par les covariables $\mathbf{X}$ et E_j celle non expliquée par les covariables $\mathbf{Z}$; x_i^p étant la valeur de la p^{e} covariable associée au génotype i et z_j^q la valeur de la q^{e} covariable associée à l'environnement j.

Les vecteurs $\gamma = \begin{pmatrix} \gamma_{11} \\ \vdots \\ \gamma_{1q} \\ \vdots \\ \gamma_{p1} \\ \vdots \\ \gamma_{pq} \end{pmatrix}$, $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_q \end{pmatrix}$ et $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$

sont estimés par régression du vecteur $\mathbf{Y}$ des observations, de longueur IJ, respectivement sur le produit semi-tensoriel ligne (Dieng, 2003) des deux

matrices des covariables, sur la matrice de covariable attachée au facteur environnement et sur la matrice de covariable attachée au facteur génotype. Nous reprenons ci-après la définition du produit semi-tensoriel ligne.

Définition 1 ***Produit semi-tensoriel ligne***

Si A *est une matrice rectangulaire* (m,r)*,*

Si B *est une matrice rectangulaire* (m,s)*,*

$A \ominus B$ *est dit produit semi-tensoriel ligne de* A *par* B*. C'est une matrice rectangulaire* (m,rs) *qui se présente ainsi*

$$\begin{pmatrix} a_{11} \cdot B[1,] & a_{12} \cdot B[1,] & \cdots & a_{1r} \cdot B[1,] \\ a_{21} \cdot B[2,] & a_{22} \cdot B[2,] & \cdots & a_{2r} \cdot B[2,] \\ \vdots & \vdots & & \vdots \\ a_{m1} \cdot B[m,] & a_{m2} \cdot B[m,] & \cdots & a_{mr} \cdot B[m,] \end{pmatrix}$$

où $B[m,]$ *est la* m^e *ligne de* B*. Autrement dit,* $A \ominus B$ *est la juxtaposition de tous les produits termes à termes possibles entre une colonne de* A *et une colonne de* B*.*

Nous nous servons du produit semi-tensoriel ligne pour obtenir la matrice d'incidence pour l'interaction à partir des matrices d'incidence des effets simples.

Avec cette présentation indicielle, la régression factorielle est effectuée en deux étapes : d'abord la régression des observations sur le produit semi-tensoriel ligne des deux matrices de covariables et celle sur les deux matrices des covariables auxquelles est ajoutée par la suite l'estimation des résidus de ces régressions.

Nous présentons ci-dessous un modèle, plus général, où tous les paramètres sont estimables en une seule étape.

$$\begin{aligned}\mathbf{Y} &= m\mathbf{1}_{IJ} + \mathbf{X}\cdot\mathbb{a}_1 + \mathbb{X}\cdot\mathbb{a}_2 + \mathbf{Z}\cdot\mathbb{b}_2 + \mathbb{Z}\cdot\mathbb{b}_2 + \mathbf{X}\ominus\mathbf{Z}\cdot(\mathbb{ab})_{11} \\ &\quad + \mathbb{X}\ominus\mathbf{Z}\cdot(\mathbb{ab})_{21} + \mathbf{X}\ominus\mathbb{Z}\cdot(\mathbb{ab})_{12} + \mathbb{X}\ominus\mathbb{Z}\cdot(\mathbb{ab})_{22} + \mathbf{e}\end{aligned}$$

- $\mathbf{Y}$ est le vecteur des observations, de longueur IJ
- $m\mathbf{1}_{IJ}$ est la moyenne générale des observations
- $\mathbf{X}\mathbb{a}_1$ est la régression sur les covariables génotypes
- $\mathbb{X}\mathbb{a}_2$ est l'écart des effets génotype à la régression sur les covariables génotypes, $\mathbb{X}$ étant la matrice d'incidence des génotypes
- $\mathbf{Z}\mathbb{b}_2$ est la régression sur les covariables environnements
- $\mathbb{Z}\mathbb{b}_2$ est l'écart des effets environnements à la régression sur les covariables environnements, $\mathbb{Z}$ étant la matrice d'incidence des environnements
- $\mathbf{X}\ominus\mathbf{Z}(\mathbb{ab})_{11}$ est l'effet des covariables environnements modulé par les covariables génotypes
- $\mathbb{X}\ominus\mathbf{Z}(\mathbb{ab})_{21}$ est l'effet des covariables environnements modulé par les génotypes non expliqué par les covariables génotypes
- $\mathbf{X}\ominus\mathbb{Z}(\mathbb{ab})_{12}$ est l'effet des covariables génotypes modulé par les environnements non expliqué par les covariables environnements
- $\mathbb{X}\ominus\mathbb{Z}(\mathbb{ab})_{22}$ est l'interaction G×E expliquée ni par les covariables génotypes, ni par les covariables environnements.

Ainsi, les paramètres $\mathbb{a}_1$, $\mathbb{a}_2$, $\mathbb{b}_2$, $\mathbb{b}_2$, $(\mathbb{ab})_{11}$, $(\mathbb{ab})_{21}$, $(\mathbb{ab})_{12}$, $(\mathbb{ab})_{22}$, sont estimés par simple régression linéaire sur les matrices et vecteurs adéquats.

Exemple d'écriture pour un essai multilocal : Pour un essai multilocal de 3 génotypes effectué en 2 lieux différents où nous avons 2 covariables associées au facteur lieu et 1 covariable associée au facteur variété, les vecteurs et matrices se présentent comme suit :

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{12} \\ Y_{22} \\ Y_{32} \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} X_{111} \\ X_{211} \\ X_{311} \\ X_{121} \\ X_{221} \\ X_{321} \end{pmatrix} \quad \mathbf{Z} = \begin{pmatrix} Z_{111} & Z_{112} \\ Z_{121} & Z_{122} \\ Z_{211} & Z_{212} \\ Z_{221} & Z_{222} \\ Z_{311} & Z_{312} \\ Z_{321} & Z_{322} \end{pmatrix}$$

$$\mathbb{X} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \mathbb{Z} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$$\mathbf{X} \ominus \mathbf{Z} = \begin{pmatrix} X_{111}Z_{111} & X_{111}Z_{112} \\ X_{211}Z_{121} & X_{211}Z_{122} \\ X_{311}Z_{211} & X_{311}Z_{212} \\ X_{121}Z_{221} & X_{121}Z_{222} \\ X_{221}Z_{311} & X_{221}Z_{312} \\ X_{321}Z_{321} & X_{321}Z_{322} \end{pmatrix}$$

$$\mathbb{X} \ominus \mathbf{Z} = \begin{pmatrix} Z_{111} & Z_{112} & 0 & 0 & 0 & 0 \\ 0 & 0 & Z_{121} & Z_{122} & 0 & 0 \\ 0 & 0 & 0 & 0 & Z_{211} & Z_{212} \\ Z_{221} & Z_{222} & 0 & 0 & 0 & 0 \\ 0 & 0 & Z_{311} & Z_{312} & 0 & 0 \\ 0 & 0 & 0 & 0 & Z_{321} & Z_{322} \end{pmatrix} \quad \mathbf{X} \ominus \mathbb{Z} \begin{pmatrix} X_{111} & 0 \\ X_{211} & 0 \\ X_{311} & 0 \\ 0 & X_{121} \\ 0 & X_{221} \\ 0 & X_{321} \end{pmatrix}$$

$$\mathbb{X} \ominus \mathbb{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Jusqu'à maintenant les covariables sont supposées continues. Toutefois, cette méthode est encore valable dans le cas où elles sont de type qualitatif. Il suffit de la même manière que l'on passe de la régression à l'analyse de variance, de remplacer chaque colonne de la matrice de covariables par des colonnes indicatrices des niveaux de la colonne de covariable qualitative considérée (Denis, 1980).

Dans le cas où les covariables sont trop nombreuses, différentes méthodes de sélection sont présentées dans (Denis, 1980).

2.4.2 Illustration avec les données de l'essai multilocal

Les données sont celles du tableau 1.2 de la page 16.

Des variables climatiques ont été mesurées quotidiennement sur les sites des essais. Ces variables (pluie, rayonnement solaire, vitesse du vent, etc.) se trouvent en fait plus nombreuses que les essais, ce qui rend impossible leur complète utilisation dans un modèle de régression factorielle. Pour décrire les différents lieux, nous avons uniquement retenu la pluviométrie totale sur le cycle de culture qui, à notre sens, permet de bien les caractériser.

Pour la covariable génotype, nous pouvons utiliser des indices de tolérance ou de sensibilité qui font intervenir le rendement en conditions de sécheresse et le rendement en conditions optimales. Mais le mode de calcul de ces indices pénalise très fortement les génotypes à rendement potentiel élevé ; des valeurs très élevées caractérisent plutôt des génotypes rustiques à faible productivité que des génotypes à rendement stable et élevé (Belhassen, This et Monneveux, 1995). Nous privilégions le taux de croissance de culture (C) décrit par Turner, Wright et Siddique (2002) comme étant :

$$C = (PF + (1,65 \cdot PG))/T$$

où PF est le poids des fanes, PG le poids des gousses et T la durée floraison-récolte. Ce taux permet effectivement de caractériser les génotypes selon leur sensibilité à la sécheresse car plus une variété produit de biomasse totale et de gousses en un temps relativement court, mieux elle est adaptée à la sécheresse. Le calcul de C s'est fait sur des données qui n'ont pas servi à la modélisation de l'interaction G×E.

Les résultats de la régression factorielle tenant compte de la covariable C attachée aux génotypes et de la covariable *Pluie* liée aux environnements, sont présentés dans le tableau 2.5. La somme des carrés résiduels (751 771,62) est ce qui reste des interactions G×E une fois que nous avons utilisé les covariables pour les interpréter. Sans ces covariables, cette somme des carrés était égale à 962 311,6 (modèle additif, tableau 2.1 de la page 24).

Effet	d.l	Somme de carrés	Carré moyen	Statistique F	Niveau de signification
Génotype	5	232 764,98	46 552,99	2,23	0,0725
Environnement	10	8 100 921,68	810 092,17	38,79	0,0000
C×Pluie	1	8 896,50	8 896,50	0,43	0,5181
C×Env.	9	162 944,57	18 104,95	0,87	0,5624
Génotype×Pluie	4	38 698,94	9 674,74	0,46	0,7622
Résidus	36	751 771,62			

TAB. 2.5 – Régression factorielle des données de l'essai multilocal.

Nous notons tout de même que les covariables associées aux effets principaux ne sont pas significatives (0,5181 pour le produit des deux covariables, 0,5624 pour la covariable liée à l'environnement et 0,7622 pour celle liée au facteur génotype). Cela indique que la décomposition de l'interaction n'est pas pertinente avec de telles covariables et qu'il faudrait en sélectionner d'autres. Ce qui pose le problème du choix de covariables caractérisant le mieux le climat des environnements. En effet, face au nombre important de variables climatiques quotidiennement mesurées sur les lieux, se pose la difficulté de

les résumer efficacement. Dans ce cas, l'utilisation de modèles de simulation de culture pour les synthétiser peut être une alternative.

2.5 Un modèle de simulation de cultures : SarraH

Les modèles de simulation de cultures permettent de prédire la production de l'ensemble du peuplement végétal cultivé d'une parcelle. Pour cela, ils reproduisent le fonctionnement du système sol-plante-atmosphère et permettent la comparaison de divers itinéraires techniques en fonction de l'espèce utilisée, du type de sol, des données climatiques, etc. Moyennant un changement d'echelle convenable, ces modèles peuvent alors être utilisés à diverses fins : prévisions des récoltes et planification de la politique agricole et alimentaire d'un pays, analyse prospective des changements climatiques, protection de l'environnement, etc. (Affholder, 1997).

Au Sahel, les modèles précurseurs fondés sur une estimation du bilan hydrique des cultures en cours de cycle, ont été écrits avant le début des années 1980 (Franquin et Forest, 1977). Le bilan hydrique se situe à l'échelle de la plante ou du champ et permet de comparer les quantités d'eau fournies et les quantités d'eau utilisées par la plante. Il tient aussi compte de la constitution de réserves d'eau et des prélèvements ultérieurs sur ces réserves. Les apports d'eau fournis par les précipitations sont mesurés. Les pertes se composent de la combinaison de l'évaporation et de la transpiration des plantes, plus connue sous le nom d'évapotranspiration.

L'équation du bilan hydrique, sous sa forme la plus générale, se fonde sur l'équation de conservation de la masse

$$P = \Delta S + D + ETR + R$$

où P désigne la pluviométrie, ΔS la variation de stock d'eau dans le sol, D le drainage, ETR l'évapotranspiration et R les ruissellements.

L'estimation de l'évapotranspiration peut se faire en deux étapes. Pour une culture, l'évapotranspiration en conditions non limitantes d'alimentation hydrique ou évapotranspiration maximale (ETM) est d'abord calculée à l'aide de l'évaporation Bac Classe A (EVA) et du coefficient cultural $K'c$ caractéristique de la culture et de son stade de développement.

$$ETM = K'c \times EVA$$

Ensuite, Eagleman (1971) a proposé une méthode d'estimation de l'évapotranspiration réelle qui est fonction de l'evapotranspiration maximale et de l'humidité relative du sol.

Doorenbos et Jassam (1979) ont développé une approche de modélisation des rendements réels fondée sur les consommations réelles et maximales d'eau et la productivité potentielle des cultures. Dans les zones où les flux hydriques sont les plus déterminants, la modélisation du bilan hydrique et la mise en relation de ses paramètres avec la production aboutit à une estimation bien adaptée des rendements des cultures.

Le modèle Arachide bilan hydrique (Arabhy) constitue la première génération de modèle semi déterministe développé par le CERAAS en collaboration avec le CIRAD, pour l'estimation de la production de l'arachide au Sénégal tenant compte du bilan hydrique (Annerose et Diagne, 1990, 1994). Il ajoute à la simulation du bilan hydrique des cultures, celle des réactions de la plante aux stress hydriques. Ce modèle estime en fin de cycle la production potentielle.

Le modèle de Système d'analyse régional des risques agroclimatiques (Sarra) simule des indicateurs hydriques de production (Forest et Clopes, 1991). Ces indicateurs sont fondés sur la consommation en eau réelle et maximale de la plante en relation avec les techniques culturales. Pour ce modèle qui est

utilisé pour mesurer l'impact du climat sur une culture annuelle, il est supposé que la performance d'une culture est une fonction simple d'une combinaison d'indicateurs hydriques cumulés au cours d'un cycle végétatif.

Cette première génération de modèles paramétrés pour les zones sahéliennes, utilise en entrée une base de données à l'echelle locale de la parcelle sur le climat, le sol et les cultures. Ces modèles sont assez simples et sont écrits sans module de bilan de carbone ni prise en compte de la notion de densité du peuplement.

Le bilan de carbone est fondé sur l'assimilation du carbone selon l'interception de l'énergie lumineuse en fonction du taux de couverture foliaire et la conversion de la fraction de rayonnement interceptée en matière sèche. Associé au bilan hydrique, le bilan de carbone constitue un élément essentiel dans le processus de croissance et de productivité des cultures.

Le modèle de simulation de cultures SarraH (Baron, 2002) est quant à lui, une version plus détaillée de Sarra et fait intervenir, en plus de la simulation du bilan hydrique, plusieurs autres effets tels que celui de la température et celui de la radiation solaire sur la production. Il est fondé sur la notion de l'effet multiplicatif de l'efficience de l'eau (WUE, *Water use efficiency*) et de l'efficience de lénergie radiative (RUE,*Radiation use efficiency*) pour simuler l'élaboration des biomasses.

Le modèle SarraH fonctionne au pas de temps journalier et la production de trois types de plantes peuvent y être simulée : mil, arachide et palmier. Nous avons pris l'arachide comme plante test pour cette étude.

De par la complexité inhérente au système sol-plante-atmosphère, les processus simulés sont formulés par des ensembles d'équations que l'on peut regrouper par grand ensembles de (bilan hydrique, phénologie, etc.) et optimiser par sous ensembles de modules (équations décrivant une étape d'un

processus). Le lien entre les modules et les processus s'établissant au moyen de variables d'états décrivant la plante, le sol et l'environnement climatique.

Développé sous cette approche modulaire, SarraH permet de simuler

- la biomasse initiale qui est fonction de la densité de semis et du poids sec moyen d'un grain ; ce qui permet de déduire la densité de peuplement disposant du taux de levée ;
- la phénologie de la plante qui se définit en plusieurs phases en relation avec la somme de température : phase végétative, phase reproductive, phase de maturation des graines. Chaque phase, d'une durée constante, est exprimée en temps thermique qui est fonction des températures maximale, minimale et de base. La température de base est celle au-dessous de laquelle la plante ne se développe pas. Les durées des phases ainsi que la température de base sont des caractéristiques variétale des plantes simulées (pour la variété d'arachide qui est simulée dans SarraH, elle est égale à 13C ;
- le bilan hydrique qui détermine la dynamique de l'eau dans le sol en fonction des contraintes climatiques (évaporation du sol) et de la consommation en eau de la plante (transpiration). Cette consommation en eau de la plante permet de définir un frein hydrique qui est le rapport entre la consommation réelle et la demande en eau de la plante ;
- le bilan de carbone qui détermine la fabrication journalière d'assimilats en fonction du coefficient de conversion de l'espèce, du rayonnement interceptée et donc de taux de couverture foliaire. A cette production potentielle d'assimilat est appliqué le frein hydrique ;
- la répartition des assimilats, racine, feuille, tige et organes reproducteurs qui évolue en fonction des phases phénologiques de la plante, ces lois de répartition sont fondées sur des relations allométriques.

2.6 Limites des méthodes classiques d'étude des interactions G×E

Le modèle 2.1, qui correspond à un modèle d'analyse de variance à deux facteurs, permet uniquement de tester la significativité des interactions, mais il ne permet aucunement de les explorer. Dans ce cas, l'information contenue dans ces interactions serait inexploitée si aucune analyse supplémentaire n'était faite (Crossa, 1990). Cette logique sous-tend la modélisation des interactions G×E qui permet alors d'améliorer la prédiction des performances des génotypes dans de nouveaux environnements en contrôlant la variation importante de ces interactions et en l'enlevant de la partie significative du modèle (Gauch, 1990, 1992).

La régression conjointe parait peu adaptée à cette forte variation d'une année sur l'autre ou d'un site à l'autre. Cette méthode permet d'interpréter l'interaction par le potentiel du milieu, estimé par l'effet moyen du milieu. Pour un nouveau milieu non encore couvert par une expérimentation, nous n'avons pas d'estimation du potentiel, donc pas de prédiction de l'interaction. De ce fait, elle permet de décrire uniquement les résidus du modèle additif et n'utilise aucune information supplémentaire du milieu pour modéliser l'interaction.

La méthode AMMI est un outil permettant de comprendre des données complexes, notamment celles obtenues dans le cadre des interactions G×E. Cependant, elle n'a qu'un grand intérêt descriptif et constitue une technique appropriée uniquement dans une perspective d'analyse préliminaire. Étant donné qu'une ACP est effectuée sur les résidus du modèle additif, la méthode AMMI permet tout juste d'étudier les corrélations entre composantes principales et covariables de l'environnement et du génotype, alors que le but est de prédire les unes à partir des autres (Yan et al., 2001). Elle péche ainsi par le fait que cette modélisation de l'interaction se fait sans l'utilisation

des données climatiques des environnements, car, de même que l'analyse de variance, les effets estimés pour les environnements sont imprévisibles.

La régression factorielle, elle, tient compte des conditions climatiques des environnements pour prédire la réponse des génotypes. Mais elle suppose que l'action de l'environnement sur la production est linéaire, ce qui n'est pas certain. De plus, le nombre important de variables climatiques généralement mesurées sur les lieux d'essais fait qu'elles ne peuvent pas être totalement prises en compte par cette méthode.

Il s'agira alors, d'améliorer la prédiction de la performance des génotypes en réduisant l'impact de la variabilité climatique sur la précision de cette estimation. Une solution serait de modéliser les variations de la réponse des génotypes en fonction de l'environnement par l'utilisation de modèles de simulation de cultures tels que Diagnostic hydrique des cultures (DHC, Forest et Cortier, 1990), Irrigation scheduling information system (IRSIS, Fao, 1987), SarraH. Seulement, les paramètres de tels modèles ne sont connus que pour un petit nombre de génotypes.

Chapitre 3

La méthode APLAT

Ce chapitre traite de notre première méthode proposée qui consiste à linéariser la performance des génotypes prédite par le modèle SarraH au voisinage d'un génotype de référence. Une fois la linéarisation effectuée, l'estimation des paramètres, du fait du nombre important de régresseurs dont nous disposions, s'est faite par régression *Partial least squares*. Nous commencerons alors par présenter à la section 1.1 cette technique de régression et terminerons ce chapitre en présentant à la section 1.2, la méthode d'estimation APLAT.

3.1 La régression Partial least squares

La régression PLS, *Partial least squares* est devenue aujourd'hui, une méthode très utilisée dans le cas des régressions sur données corrélées. Aussi, est-elle une bonne alternative s'il y a plus de régresseurs que d'observations (Wold, Albano, Dunn, Esbensen, Hellberg, Johansson, Sjöström 1983 ; Tenenhaus, 2001).

Un petit nombre de variables appelées "facteurs" ou "variables latentes" sont construites l'une après l'autre de façon itérative et permettent de remplacer l'espace initial des nombreux régresseurs par un espace de plus faible dimension. Ces facteurs deviennent les nouvelles variables explicatives dans un modèle de régression linéaire classique.

Les facteurs sont orthogonaux, et sont des combinaisons linéaires des variables explicatives initiales. A ce titre, ils renvoient aux composantes principales de la RCP, Régression sur composantes principales. Mais alors que ces dernières ne sont calculées qu'à partir des variables explicatives (et donc sans référence à la variable à expliquer), les facteurs de la régression PLS maximisent les corrélations successives entre les régresseurs et la variable à expliquer, tout en maintenant la contrainte d'orthogonalité avec ceux déjà construits.

La régression PLS s'effectue selon le principe de l'algorithme NIPALS, *Nonlinear estimation by iterative partial least squares* développé par Herman Swold (1966) pour l'analyse en composantes principales. Cette régression s'inspire de l'approche PLS (Wold, 1975) pour l'estimation des modèles d'équation structurelles reliant plusieurs blocs de variables entre eux.

A présent, pour décrire cette méthode, nous nous plaçons dans le cadre du modèle linéaire classique. Le vecteur des observations $\mathbf{Y}$ de dimension $n \times 1$ est supposé suivre le modèle suivant

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{3.1}$$

où le vecteur $\boldsymbol{\beta}$ d'ordre p est le paramètre inconnu à estimer, $\mathbf{X}$ la matrice de dimension $n \times p$ des variables explicatives, et le vecteur $\mathbf{e}$ un terme d'erreur aléatoire.

Nous supposerons qu'il n'y a pas de données manquantes et qu'il n'y a qu'une seule variable à expliquer pour une explication plus claire de la méthode. L'algorithme PLS calcule les variables latentes $\mathbf{t}_1, \cdots, \mathbf{t}_h$ étape par étape. Ces variables latentes $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h$ sont des combinaisons linéaires des $\mathbf{X}$ qui sont orthogonales entre elles et qui maximisent $\mathrm{Cov}(\mathbf{t}_h, \mathbf{Y})$ sous la contrainte $\| \mathbf{w}_h \| = 1$.

A l'étape 1, $\mathbf{w}_1 = (w_1^1 \cdots w_1^p)'$ est solution du problème d'optimisation

$$\begin{cases} \max \mathrm{Cov}(\mathbf{X}\mathbf{w}_1, \mathbf{Y}) \\ \| \mathbf{w}_1 \| = 1 \end{cases}$$

Pour déterminer $\mathbf{w}_1$, il suffit d'écrire l'expression du Lagrangien.

$$\begin{aligned} \mathrm{L}(\mathbf{w}_1, \lambda) &= \mathrm{Cov}(\mathbf{X}\mathbf{w}_1, \mathbf{Y}) - \lambda(\mathbf{w}_1'\mathbf{w}_1 - 1) \\ &= w_1^1 \mathrm{Cov}(\mathbf{X}^1, \mathbf{Y}) + \cdots + w_1^p \mathrm{Cov}(\mathbf{X}^p, \mathbf{Y}) - \lambda\big[(w_1^1)^2 + \cdots + (w_1^p)^2 - 1\big] \end{aligned}$$

où λ est le multiplicateur de Lagrange associé à la contrainte et $\mathbf{X}^p$ la p^{e}colonne de $\mathbf{X}$.

Les solutions à ce problème d'optimisation sont obtenues en dérivant $\mathrm{L}(\mathbf{w}_1, \lambda)$ par rapport à $w_1^1, \cdots, w_1^p, \lambda$. Les $p + 1$ équations aux dérivées partielles ou équations normales s'écrivent

$$\begin{cases} \mathrm{Cov}(\mathbf{X}^1, \mathbf{Y}) - 2\lambda w_1^1 &= 0 \\ \qquad\qquad\qquad\qquad\quad \vdots & \\ \mathrm{Cov}(\mathbf{X}^p, \mathbf{Y}) - 2\lambda w_p^1 &= 0 \\ (w_1^1)^2 + \cdots (w_1^p)^2 &= 1 \end{cases}$$

En remplaçant dans la dernière équation de ce système les composantes de $\mathbf{w}_1$ tirées dans les p premières équations, nous obtenons

$$\left[\text{Cov}(\mathbf{X}^1, \mathbf{Y})/2\lambda\right]^2 + \cdots + \left[\text{Cov}(\mathbf{X}^p, \mathbf{Y})/2\lambda\right]^2 = 1$$

D'où

$$\sum_{j=1}^{p} [\text{Cov}(\mathbf{X}^j, \mathbf{Y})]^2 = 4\lambda^2$$

Et

$$\lambda = \sqrt{\sum [\text{Cov}(\mathbf{X}^j, \mathbf{Y})]^2}/2$$

En reportant cette valeur de λ dans chacune des p premières équations normales, nous avons

$$w_1^j = \text{Cov}(\mathbf{X}^j, \mathbf{Y})/\sqrt{\sum [\text{Cov}(\mathbf{X}^j, \mathbf{Y})]^2}$$

Ainsi, la première composante $\mathbf{t}_1 = w_1^1\mathbf{X}^1 + \cdots + w_1^p\mathbf{X}^p$ est construite. Puis, il est effectué une régression simple de $\mathbf{Y}$ sur $\mathbf{t}_1$

$$\mathbf{Y} = c_1\mathbf{t}_1 + \mathbf{Y}_1$$

où c_1 est le coefficient de régression et $\mathbf{Y}_1$ le vecteur des résidus.

S'il reste encore de l'information, il est construit une deuxième variable latente $\mathbf{t}_2 \perp \mathbf{t}_1$. Cette deuxième variable latente est combinaison linéaire des colonnes de $\mathbf{X}_1$, résidu de la régression linéaire de $\mathbf{X}$ sur $\mathbf{t}_1$.

A l'étape 2, $\mathbf{w}_2 = (w_2^1 \cdots w_2^p)'$ est solution du problème d'optimisation

$$\begin{cases} \max \text{Cov}(\mathbf{X}_1\mathbf{w}_2, \mathbf{Y}_1) \\ \| \mathbf{w}_2 \| = 1 \end{cases}$$

La deuxième variable latente $\mathbf{t}_2$ construite, il est effectué une régression linéaire multiple de $\mathbf{Y}$ sur $\mathbf{t}_1$ et $\mathbf{t}_2$

$$\mathbf{Y} = c_1\mathbf{t}_1 + c_2\mathbf{t}_2 + \mathbf{Y}_2$$

Cette procédure itérative peut ainsi continuer en utilisant les résidus $\mathbf{Y}_2$, $\mathbf{X}_2$ des régressions de $\mathbf{Y}$, $\mathbf{X}$ sur $\mathbf{t}_1$ et $\mathbf{t}_2$.

Le nombre de composantes $\mathbf{t}_1, \cdots, \mathbf{t}_H$ à retenir avec $H \leqslant \text{rang}(\mathbf{X})$, peut être déterminé à l'aide de trois critères : l'ajustement de l'échantillon d'apprentissage ($\mathbf{X}$, $\mathbf{Y}$) par ($\widehat{\mathbf{X}}$, $\widehat{\mathbf{Y}}$), la prédiction sur un échantillon externe et la prédiction interne aux données d'apprentissage appelée validation croisée.

3.2 La méthode APLAT : linéarisation autour d'un témoin

Cette méthode a fait l'objet d'un article publié aux Comptes rendus de l'académie des sciences dont l'original se trouve en Annexe B.

I. Dieng, E. Gozé, R. Sabatier

Linéarisation autour d'un témoin pour prédire la réponse de cultures

C. R. Biologies 329 (2006) 148-155

3.2.1 Le modèle proposé

Si nous partons du modèle de simulation de cultures, le rendement d'un génotype i dans un environnement j est la somme :

1. d'un rendement potentiel prédit avec le modèle de simulation ;
2. d'un biais, espérance de l'écart du potentiel au réalisé ;
3. d'une erreur aléatoire.

$$Y_{ij} = f(\mathbf{Z}_j, \boldsymbol{\theta}_i) + \xi_j + u_{ij} \tag{3.2}$$

où $\mathbf{Z}_j$ est le vecteur des variables telles que la pluie, la température, etc. mesurées sur l'environnement j et $\boldsymbol{\theta}_i$ le vecteur de longueur P des paramètres du génotype i. Nous supposons que le biais ξ_j ne dépend que de l'environnement j : il est donc le même pour tous les génotypes d'un même environnement. L'erreur u_{ij} est supposée aléatoire avec $\mathbb{E}(u_{ij}) = 0$ et $\mathrm{Var}(u_{ij}) = \sigma_u^2$.

Comme dit précédemment, les paramètres des modèles de simulation de cultures ne sont généralement connus que pour un petit nombre de génotypes. Considérons un modèle de simulation de cultures et un génotype de référence dont les paramètres sont connus et appelons $\boldsymbol{\theta}_0$ le vecteur de ses paramètres. Alors, supposons f de classe C^1 dans un voisinage de $\boldsymbol{\theta}_0$ et f' dérivable sur ce voisinage. De plus supposons $\boldsymbol{\theta}_i$ au voisinage de $\boldsymbol{\theta}_0$. En pratique, les génotypes dont nous chercherons à estimer les paramètres seront choisis de telle sorte qu'ils ne soient pas trop éloignés du génotype de référence. Alors, un développement en série de Taylor à l'ordre 1 nous donne :

$$f(\mathbf{Z}_j, \boldsymbol{\theta}_i) = f(\mathbf{Z}_j, \boldsymbol{\theta}_0) + \sum_{p=1}^{P} \left[\frac{\partial f}{\partial \theta^{(p)}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \mathbf{Z}=\mathbf{Z}_j} (\theta_i^{(p)} - \theta_0^{(p)}) + \circ\left[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)\right] \tag{3.3}$$

avec $\theta_i^{(p)}$ et $\theta_0^{(p)}$ la p^{e} composante du vecteur de paramètres respectivement du génotype i et du génotype de référence.

Posons

$$X_j^{(p)} = \left[\frac{\partial f}{\partial \theta^{(p)}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \mathbf{Z}=\mathbf{Z}_j}$$

c'est une fonction de l'environnement j, et

$$\beta_i^{(p)} = \theta_i^{(p)} - \theta_0^{(p)}$$

une fonction du génotype i.

La fonction $X_j^{(p)}$ est la dérivée partielle de la sortie du modèle de simulation de cultures pour l'environnement j par rapport à la p^e composante du vecteur de paramètres de la variété de référence. Comme la fonction f n'est pas généralement connue analytiquement, ces sensibilités peuvent être obtenues par une méthode de dérivation numérique. Nous avons retenu tout simplement

$$X_j^{(p)} = \left[\frac{\partial f}{\partial \theta^{(p)}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \mathbf{Z}=\mathbf{Z}_j} \simeq \left[\frac{f\left(\theta_0^{(p)} + h_{\theta_0^{(p)}}\right) - f\left(\theta_0^{(p)} - h_{\theta_0^{(p)}}\right)}{2h_{\theta_0^{(p)}}}\right]_{\mathbf{Z}=\mathbf{Z}_j}$$

avec $h_{\theta_0^{(p)}}$ très petit, de l'ordre de $\theta_0^{(p)}.10^{-4}$ en pratique. D'autres méthodes existent, celle-ci étant la plus simple et la plus économe en calculs.

Avec ces notations et d'après les équations (3.2) et (3.3) qui permettent d'écrire

$$f(\mathbf{Z}_j, \boldsymbol{\theta}_0) = Y_{0j} - \xi_j - u_{0j}$$

nous pouvons poser, en négligeant $\circ[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)]$:

$$Y_{ij} - Y_{0j} = \sum_{p=1}^{P} X_j^{(p)} \cdot \beta_i^{(p)} + \epsilon_{ij} \tag{3.4}$$

où $\epsilon_{ij} = u_{ij} - u_{0j}$.

Ainsi, $\mathbb{E}(\epsilon_{ij}) = 0$, $\mathrm{Var}(\epsilon_{ij}) = 2\sigma_u^2$, $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$, mais $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j}) = \sigma_u^2$.

Si nous disposons de I génotypes et de J environnements, nous pouvons poser le modèle suivant :

$$\mathbf{Y} - (\mathbf{Y}_0 \otimes \mathbf{1}_I) = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.5}$$

Le vecteur $\mathbf{Y}$ représente le rendement de tous les génotypes dans tous les environnements, rangé par environnement et par génotype. Si tous les génotypes

ont été observés une fois dans chaque environnement, ce vecteur est de longueur IJ. Puis $\mathbf{Y}'_0 = (Y_{01} \cdots Y_{0J})$ et $\mathbf{1}_I$ est un vecteur formé de 1, de longueur I.

Le symbole $\otimes$ désigne le produit de Kronecker.

Le vecteur $\boldsymbol{\epsilon}$ est un vecteur d'erreur aléatoire. Sa matrice de covariance est de la forme $\sigma_u^2 \boldsymbol{\Omega}$ avec

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\omega}_1 & & & & \\ & \ddots & & & 0 \\ & & \boldsymbol{\omega}_j & & \\ & 0 & & \ddots & \\ & & & & \boldsymbol{\omega}_J \end{pmatrix}$$

où

$$\boldsymbol{\omega}_j = \begin{pmatrix} 2 & & 1 \\ & \ddots & \\ 1 & & 2 \end{pmatrix}$$

Les matrices $\boldsymbol{\Omega}$ et $\boldsymbol{\omega_j}$ sont carrées de nombre de lignes, respectivement le nombre d'observations de tous les environnements et le nombre d'observations de l'environnement j. La matrice $\boldsymbol{\Omega}$ est bloc diagonale tandis que $\boldsymbol{\omega}_j$ est une matrice formée de 1 partout sauf des 2 sur la diagonale. Dans le cas où tous les génotypes ont été vus une seule fois dans chaque environnement, $\boldsymbol{\Omega}$ est de dimension $IJ \times IJ$ et $\boldsymbol{\omega}_j$ de dimension $I \times I$.

Ensuite,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \otimes \mathbf{I}_I & \cdots & \mathbf{X}^{(p)} \otimes \mathbf{I}_I & \cdots & \mathbf{X}^{(P)} \otimes \mathbf{I}_I \end{bmatrix}$$

où

$$\mathbf{X}^{(p)\prime} = \left[\begin{array}{ccc} X_1^{(p)} & \cdots & X_J^{(p)} \end{array} \right]$$

de longueur J, est le vecteur des dérivées des sorties de notre modèle de simulation de cultures, SarraH, dans chacun des environnements par rapport au p^{e} paramètre du vecteur de paramètres du témoin.

La matrice $\mathbf{I}_I$ est la matrice identité d'ordre I, la matrice $\mathbf{X}^{(p)} \otimes \mathbf{I}_I$ s'écrit alors de cette façon :

$$\left(\begin{array}{c} X_1^{(p)} \\ \vdots \\ X_J^{(p)} \end{array} \right) \otimes \left(\begin{array}{ccc} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{array} \right) = \left(\begin{array}{ccc} X_1^{(p)} & & 0 \\ & \ddots & \\ 0 & & X_1^{(p)} \\ & \vdots & \\ X_J^{(p)} & & 0 \\ & \ddots & \\ 0 & & X_J^{(p)} \end{array} \right)$$

C'est une matrice à IJ lignes et I colonnes. Par conséquent, la matrice $\mathbf{X}$ est de dimension $IJ \times PI$.

Enfin

$$\boldsymbol{\beta} = \left[\begin{array}{ccc} \boldsymbol{\beta}^{(1)\prime} & \cdots & \boldsymbol{\beta}^{(P)\prime} \end{array} \right]'$$

avec

$$\boldsymbol{\beta}^{(p)\prime} = \left[\begin{array}{ccc} \beta_1^{(p)} & \cdots & \beta_I^{(p)} \end{array} \right]$$

Nous avons proposé d'appeler cette méthode par l'acronyme APLAT : Approximation Par Linéarisation Autour d'un Témoin. Elle consiste à approcher, localement, le rendement prédit par un modèle de simulation de cultures,

par série de Taylor à l'ordre 1 au voisinage du vecteur de paramètres d'un génotype de référence. Cette linéarisation permet, par régression linéaire, l'estimation des paramètres de ces génotypes. Par la suite, la prédiction de l'écart entre le rendement de ces génotypes et celui du génotype de référence dans des environnements nouveaux, c'est à dire où ils ne sont pas encore testés, pourra se faire si le climat de ces derniers est connu.

Il y a en général beaucoup de paramètres dans un modèle de simulation de cultures et peu d'environnements dans un essai multienvironnement, ce qui rend souvent PI grand par rapport à IJ.

Pour notre exemple, nous avons utilisé SarraH comme modèle de simulation de cultures. Ce modèle dispose de 61 paramètres fonction du génotype. Avec un tel nombre de prédicteurs, l'estimation de $\boldsymbol{\beta}$ s'est faite par régression PLS. Ceci permet de réduire l'espace des régresseurs de rang de $\mathbf{X}$ à k dimensions.

La régression PLS, voir section 3.1, s'effectue selon le principe de l'algorithme NIPALS, *Nonlinear estimation by Iterative Partial Least Squares*, (Tenenhaus, 2001) où un ensemble de régressions successives par moindres carrés ordinaires est effectué, en même temps que le calcul des composantes.

Ici, la matrice de covariance de $\boldsymbol{\epsilon}$ est égale à $\sigma_u^2\boldsymbol{\Omega}$ et non à $\sigma_u^2\mathbf{I}_{IJ}$. La solution serait d'effectuer toutes les régressions partielles par moindres carrés généralisés. Mais cette matrice de covariance est inconnue. Elle s'écrit tout de même à une constante multiplicative près en fonction de $\boldsymbol{\Omega}$ qui elle est connue. La matrice $\boldsymbol{\Omega}$ étant symétrique et semi-définie positive, par décomposition de Cholesky, il existe une matrice $\boldsymbol{\eta}$ tel que $\boldsymbol{\eta}'\boldsymbol{\eta} = \boldsymbol{\Omega}^{-1}$.

Si nous multiplions à gauche tous les termes du modèle 3.5 par $\boldsymbol{\eta}$, nous obtenons le modèle 3.6 dont les erreurs sont indépendantes.

$$\boldsymbol{\eta}\mathbf{Y} - \boldsymbol{\eta}(\mathbf{Y}_0 \otimes \mathbf{1}_I) = \boldsymbol{\eta}\mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\eta}\boldsymbol{\epsilon} \tag{3.6}$$

En effet, la variance de l'erreur $\boldsymbol{\eta\epsilon}$ s'écrit

$$\mathbb{E}(\boldsymbol{\eta\epsilon\epsilon'\eta'}) = \boldsymbol{\eta}\mathbb{E}(\boldsymbol{\epsilon\epsilon'})\boldsymbol{\eta'} = \boldsymbol{\sigma}_u^2\boldsymbol{\eta\Omega\eta'} = \boldsymbol{\sigma}_u^2\boldsymbol{\eta}(\boldsymbol{\eta'\eta})^{-1}\boldsymbol{\eta'} = \boldsymbol{\sigma}_u^2\boldsymbol{\eta\eta}^{-1}(\boldsymbol{\eta'})^{-1}\boldsymbol{\eta'} = \boldsymbol{\sigma}_u^2\mathbf{I}_{IJ}$$

$\boldsymbol{\beta}$ peut alors être estimé à l'aide des moindres carrés ordinaires, appelons $\tilde{\boldsymbol{\beta}}_{PLS}$ son estimateur.

Le nombre de composantes à retenir est déterminé par le PRESS, *Prediction Error Sum of Squares* (Tenenhaus, 2001).

3.2.2 Illustration avec les données de l'essai pluriannuel

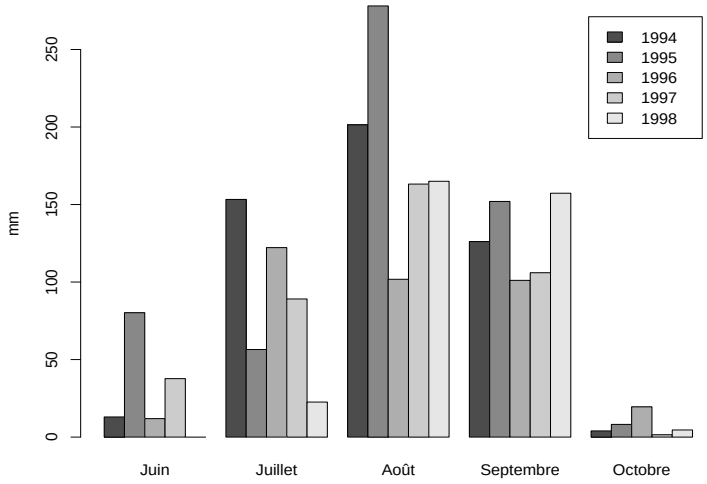
Les données

Les données sont les résultats de l'essai pluriannuel (tableau 1.1, page 13). Rappelons que le génotype de référence est le *55-437*, un génotype de 90 jours. C'est une variété largement cultivée au Sénégal, dans le bassin arachidier, à ce titre elle a été présente comme témoin dans tous les essais et c'est elle aussi qui a servi pour le paramétrage du modèle SarraH.

Dans ce milieu à forte variabilité des pluies dans l'espace et même dans le temps pour un même lieu, nous avons considéré chacune des 5 années d'expérimentation comme un environnement (figure 3.1).

SarraH a été utilisé pour calculer $\mathbf{X}$. Compte tenu du nombre de données disponibles, seuls deux paramètres ($P = 2$) ont été considérés parmi les 61 de SarraH. Le premier paramètre est en fait un coefficient multiplicateur qui agit sur 5 paramètres de SarraH : Coefficient moyen d'angle des feuilles, Coefficient de conversion en assimilat, Coefficient d'efficience d'assimilation des feuilles à la phase végétative juvénile, Coefficient d'efficience d'assimilation

FIG. 3.1 – Répartition des pluies sur la station de Bambey au Sénégal de 1994 à 1998.



des feuilles à la première phase de maturation - phase sensible de remplissage des grains - et Coefficient d'efficience d'assimilation des feuilles à la deuxième phase de maturation - phase non sensible -. Le deuxième paramètre est le poids moyen des gousses.

Validation croisée

Pour valider notre modèle, nous avons réservé successivement chacune des années et estimé les paramètres des génotypes sur les années restantes. Pour chacune des cinq années, nous avons identifié un modèle par la méthode APLAT et les rendements observés ont été comparés à ceux ainsi prédits. Les rendements sont exprimés en kilogrammes de gousses par hectare.

L'évaluation de la qualité de chaque modèle proposé est faite avec l'erreur quadratique moyenne de prédiction MSEP, *Mean Squared Error of Prediction*

(Wallach et Goffinet, 1987). La MSEP est utilisée comme critère pour comparer différents modèles dont le modèle moyen (Colson, Wallach, Bouniols, Denis et Jones, 1995) défini pour nos données par :

$$Y_{ij} = m + g_i + E_j + \delta_{ij} \tag{3.7}$$

où m est la moyenne de la population et g_i l'effet génotype. L'effet E_j de l'environnement j est supposé aléatoire, d'espérance nulle et de variance σ_E^2. Les erreurs δ_{ij} sont indépendantes, d'espérance nulle et de variance σ_δ^2. De plus, E_j et δ_{ij} sont supposés indépendants.

Le modèle moyen n'est rien d'autre que le modèle linéaire mixte (2.1) où l'interaction aléatoire G×E, imprévissible, est fusionnée avec le terme d'erreur qui porte les mêmes indices i pour le génotype et j pour l'environnement.

Nous avons calculé les intervalles de confiance des coefficients estimés par la méthode *bootstrap* (Efron, 1979). Cette technique permet d'estimer la loi inconnue d'un estimateur par une loi empirique obtenue à partir d'une procédure de rééchantillonnage fondée sur des tirages aléatoires avec remise des données. Les intervalles de confiance construits sont de type percentile-t (Aji, Tavolaro, Lantz, et Faraj, 2003).

Soit $z_{i,\mathrm{PLS}}^{(p)\star b}$ la variable aléatoire définie par :

$$z_{i,\mathrm{PLS}}^{(p)\star b} = \frac{\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b} - \tilde{\beta}_{i,\mathrm{PLS}}^{(p)}}{\tilde{s}^\star(\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b})} \tag{3.8}$$

où $\tilde{\beta}_{i,\mathrm{PLS}}^{(p)}$ est le $(p.i)^{\mathrm{e}}$ élément de $\widetilde{\boldsymbol{\beta}}_{\mathrm{PLS}}$, il s'agit du p^{e} paramètre de la i^{e} variété estimé par la méthode PLS. $\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b}$ est obtenu au b^{e} tirage avec $b = 1, \cdots, \mathrm{B}$.

$\tilde{s}^{\star}(\tilde{\beta}^{(p)\star b}_{i,\mathrm{PLS}})$, l'écart-type estimé de $\widetilde{\boldsymbol{\beta}}^{\star b}_{\mathrm{PLS}}$, est donné par le $(p.i)^{\mathrm{e}}$ élément diagonal de la matrice de variance de $\tilde{\beta}^{(p)\star b}_{i,\mathrm{PLS}}$.

Soit $\widehat{\mathrm{F}}_{\mathrm{B}}$ la fonction de répartition empirique des $z^{(p)\star b}_{i,\mathrm{PLS}}$. Le fractile d'ordre α, $\widehat{\mathrm{F}}^{-1}_{\mathrm{B}}(\alpha)$ est estimé par la valeur $\widehat{t}(\alpha)$ telle que

$$\tfrac{1}{\mathrm{B}}\sum_{b=1}^{\mathrm{B}} \mathbb{1}_{\left\{z^{(p)\star b}_{i,\mathrm{PLS}} \leqslant \widehat{t}(\alpha)\right\}} = \alpha$$

Alors un intervalle de confiance percentile-t pour le $(p.i)^{\mathrm{e}}$ élément de $\boldsymbol{\beta}$ peut s'écrire :

$$\left[\tilde{\beta}^{(p)}_{i,\mathrm{PLS}} - \tilde{s}(\tilde{\beta}^{(p)}_{i,\mathrm{PLS}}) \cdot \widehat{t}(1-\alpha)\ ,\ \tilde{\beta}^{(p)}_{i,\mathrm{PLS}} - \tilde{s}(\tilde{\beta}^{(p)}_{i,\mathrm{PLS}}) \cdot \widehat{t}(\alpha)\right] \tag{3.9}$$

Résultats

Pour les modèles sans les données respectivement de 1994, 1995 et 1997, le PRESS minimal est atteint avec 6 composantes. Pour les deux autres modèles, le PRESS est minimal avec 9 composantes, mais nous avons réduit leur espace à 5 dimensions car le PRESS n'y est pas trop différent de ses valeurs minimales (figure 3.2).

Les coefficients des régressions PLS et les intervalles de confiance qui leur sont associés, sont représentés à la figure 3.3.

Les MSEP estimées pour les modèles APLAT, sauf celle sans les données de l'année 1998, sont inférieures aux MSEP des modèles moyens correspondants (tableau 4.2). Ce qui signifie que pour ces modèles, prédire le rendement par la méthode APLAT est meilleur que par la moyenne des rendements du passé. Ainsi, 4 fois sur 5, la méthode APLAT s'est révélée meilleure que le modèle moyen. Toutefois, cette étude souffre de la faible taille de notre échantillon.

FIG. 3.2 – Evolution du PRESS en fonction du nombre de composantes. Le modèle (-1994) utilise les données sauf celles de l'année 1994 et ainsi de suite.

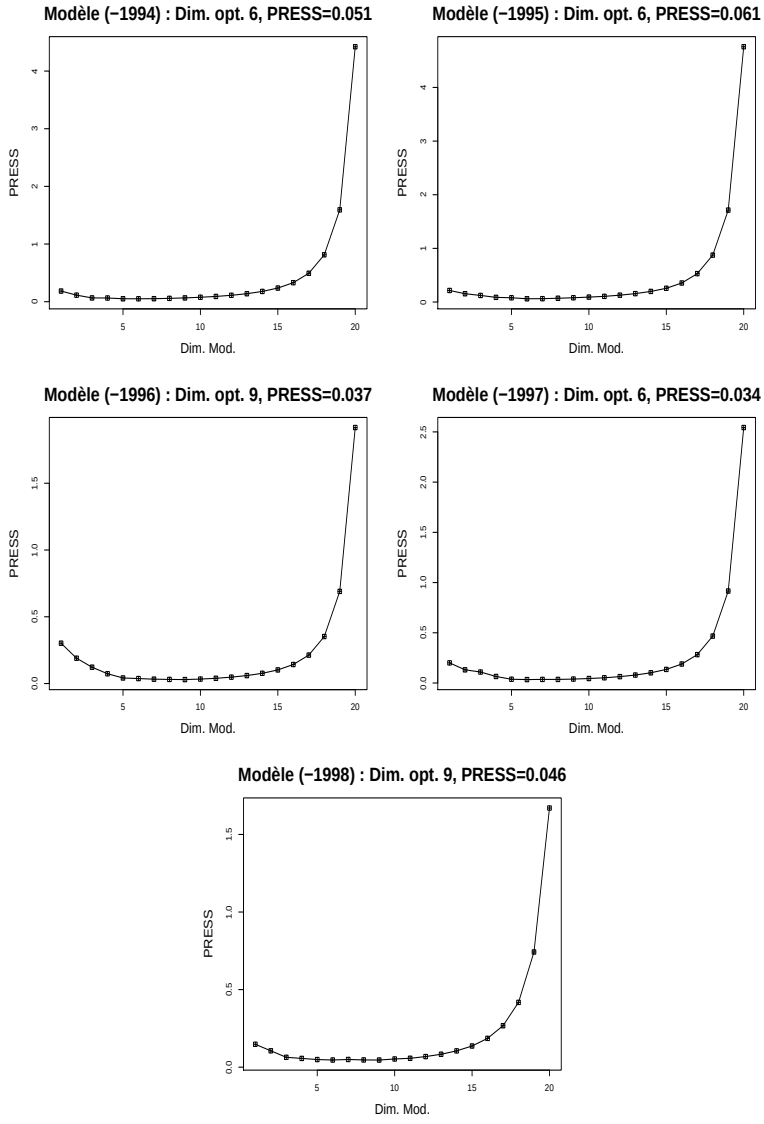
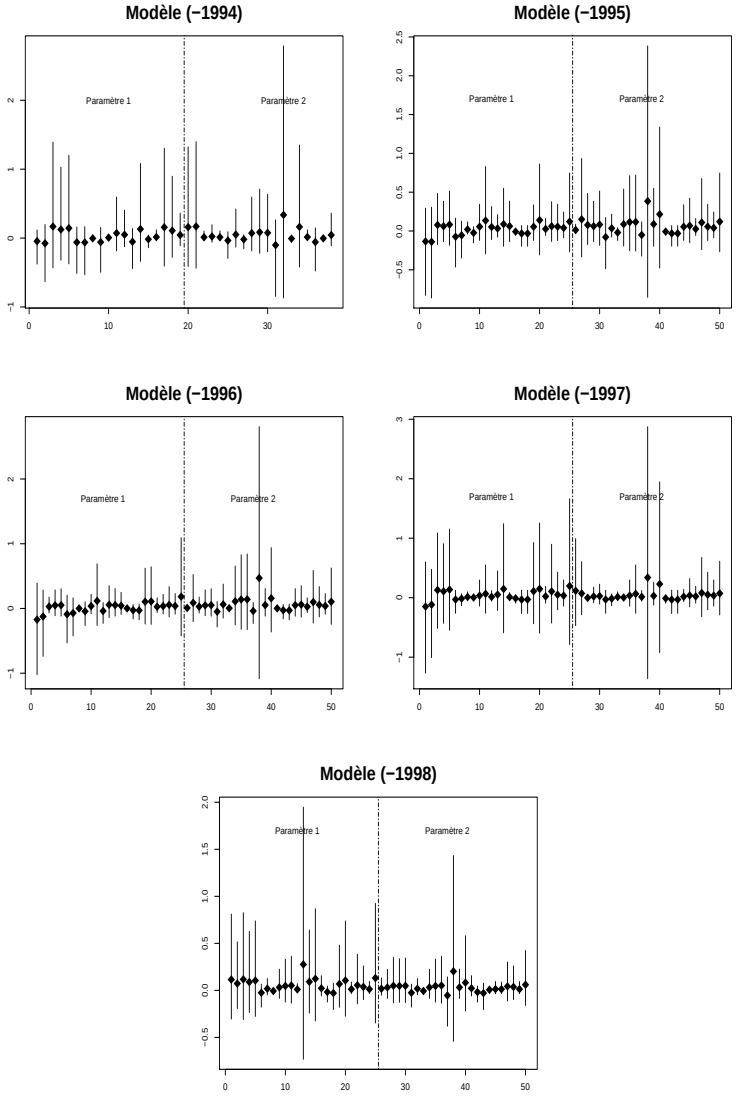


FIG. 3.3 – Intervalle de confiance percentile-t à 95% des coefficients estimés. Le modèle (-1994) est sans les données de 1994 et ainsi de suite. Sur l'axe des abscisses, les valeurs de chaque paramètre sont rangées par ordre alphabétique de génotype. Le symbole ♦ représente l'estimation ponctuelle des coefficients.



	APLAT	Modèle moyen
Modèle (-1994)	24 687,3	64 651,6
Modèle (-1995)	5 915,0	7 160,6
Modèle (-1996)	35 446,1	37 814,8
Modèle (-1997)	10 038,3	18 201,1
Modèle (-1998)	118 304,9	84 963,6

TAB. 3.1 – MSEP des différents modèles APLAT et modèles moyens correspondants de l'essai pluriannuel. Le modèle (-1994) est sans les données de 1994 et ainsi de suite.

3.2.3 Illustration avec les données de l'essai multilocal

Les données sont les résultats de l'essai multilocal (tableau 1.2, page 15). Chaque lieu est réservé et l'estimation des paramètres effectuée sur les lieux restants.

Le modèle SarraH a aussi été utilisé pour simuler le rendement de la variété de référence dans les différents lieux. Les deux paramètres à réestimer restent les mêmes qu'à la section précédente.

Les MSEP estimés pour les modèles APLAT sont inférieures aux MSEP des modèles moyens correspondants, sauf celle sans les données de Wenthiwy (tableau 3.2). Ainsi, pour cet essai multilocal, la méthode APLAT s'est montrée meilleure que le modèle moyen 10 fois sur 11 (c'était 4 fois sur 5 avec l'essai pluriannuel).

3.2.4 Conclusion

Au Sahel, l'interaction G×E est largement due aux aléas climatiques, dont la probabilité peut être estimée à l'aide de longues chroniques de relevés météorologiques au sol. Cependant, relier l'interaction G×E et la pluviométrie à l'aide d'un modèle de simulation de cultures n'est habituellement possible que pour des variétés dont on a estimé les paramètres, au prix d'une

	APLAT	Modèle moyen
Modèle (sans Bambey)	10 289,4	45 374,1
Modèle (sans Gatte)	6 457,1	102 784,7
Modèle (sans Keur Fary)	3 474,2	44 457,9
Modèle (sans Ndangalma)	7 445,7	74 477,6
Modèle (sans Ndiadiane)	12 824,9	99 194,2
Modèle (sans Nioro)	47 015,4	157 989,6
Modèle (sans Nioro Sud)	234 506,4	738 389,6
Modèle (sans Pasokoto)	76 987,8	559 545,3
Modèle (sans Keur Samseun)	1 677,8	120 068,3
Modèle (sans Sinthiou Thiabala)	11 041,1	20 998,8
Modèle (sans Wenthiwy)	80 887,3	48 687,1

TAB. 3.2 – MSEP des différents modèles APLAT et modèles moyens correspondants de l'essai multilocal.

expérimentation spécifique. Le modèle APLAT permet de prédire cette interaction avec les seules données d'une expérimentation multilocale classique, sans autre instrumentation que des stations météorologiques simples.

La méthode APLAT peut être vue ainsi comme un outil d'aide à la décision pour la sélection au Sahel. Dans l'exemple où un sélectionneur doit tester plusieurs génotypes dans un nouvel environnement, cette méthode lui permettra d'écarter d'emblée certains génotypes qui donneront une production faible. En lieu et place de longs essais multilocaux ou pluriannuels ou d'une tentative de paramétrisation d'un modèle de simulation de cultures qui implique un coût élevé. Son attention sera portée par la suite sur l'ensemble restreint des génotypes retenus avec APLAT où il pourra appliquer les schémas classiques de sélection.

Cette méthode évite ainsi de recourir soit à de longs essais multilocaux et pluriannuels, soit à la paramétrisation coûteuse d'un modèle de simulation de culture. De plus, elle permet d'assumer le choix des lieux comme devant expliquer largement la diversité du milieu plutôt que de la représenter à travers un sondage équiprobable. En effet, avec APLAT, seuls les résidus de

la modélisation de l'interaction G×E sont aléatoires, et non pas l'interaction toute entière.

Chapitre 4

La méthode APLAT-mixte

Dans ce chapitre, nous partirons de la méthode APLAT, qui nous le rappelons, consiste à linéariser autour du vecteur de paramètres d'un génotype de référence, la réponse de génotypes prédite par un modèle de simulation de cultures.

Notre hypothèse pour cette méthode, était qu'un modèle de simulation de cultures, qui est une fonction des paramètres des génotypes et des caractéristiques des environnements où nous voulons faire la prédiction, permet de capter la majeure partie de l'effet aléatoire de l'environnement. Ce qui autorise la réduction des interactions aléatoires G×E, et par là, facilite la sélection variétale.

Dans cette partie, nous revenons sur cette hypothèse faite en première approximation et considérons que certes, si l'aléa de l'environnement peut être pris en compte par un modèle de simulation de cultures, il ne l'est toutefois pas totalement. Ainsi, si nous pensons toujours pouvoir modéliser les variations de la réponse d'un génotype dans un environnement par le biais de tels modèles de simulation, il subsistera néanmoins de l'aléa environnemental,

responsable d'éventuelles interactions G×E, dont il conviendra d'estimer sa variance, le cas échéant.

Ce chapitre traite alors de notre deuxième méthode mise au point pour estimer les paramètres fixes issus de la linéarisation du rendement des génotypes par le modèle SarraH et les composantes de variance de l'aléa environnemental restant. Avec les nombreux paramètres de SarraH, ce qui est le cas pour la plupart des modèles de simulation de cultures, nous devons identifier un modèle où un nombre important de régresseurs et des composantes de variance coexistent. Nous avons à cet effet, proposé une méthode combinée de régression PLS et de modèle mixte pour l'estimation des paramètres inconnus. Il s'agit d'une extension de la méthode APLAT proposée au précédent chapitre où cette fois-ci, il est noté la présence d'effets aléatoires additionnels dont nous nous attacherons à estimer les variances. Cette méthode a été dénommée APLAT-Mixte.

Nous débutons ce chapitre par la Section 4.1 où nous faisons un retour sur le modèle mixte. Si les composantes de variance étaient connues, c'est-à-dire s'il y avait uniquement une contrainte à savoir le nombre important de régresseurs par rapport aux observations, une extension de la méthode PLS devrait suffire pour résoudre le problème. Une telle procédure est alors détaillée à la Section 4.2. Comme ces composantes sont est en fait souvent inconnues, une méthode combinée de PLS et d'algorithme EM est présentée à la Section 4.3 pour estimer les paramètres inconnus. Les développements de cette méthode appelée PLS-Mixte, sont fondés dans un premier temps, sur un modèle mixte où des effets aléatoires sont supposés simplement de variance $\sigma^2\mathbf{I}$. Pour l'éprouver, nous nous sommes éloignés de nos données qui ne s'y prêtent pas et avons eu recours à des données de NIRS. Ensuite, puisque nos données d'interaction G×E peuvent être appréhendées à travers un modèle mixte où les effets aléatoires sont supposés de variance $\sigma^2\mathbf{\Delta}$, nous avons adapté la méthode PLS-Mixte à ce type de modèle.

4.1 Le modèle mixte

Considérons un modèle linéaire mixte comme décrit par McCulloch et Searle (2001). Le vecteur des observations $\mathbf{Y}$ de dimension $n \times 1$ est supposé suivre le modèle suivant :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{4.1}$$

où $\boldsymbol{\beta}$ d'ordre p est le vecteur de paramètres des effets fixes, $\mathbf{u}$ d'ordre q le vecteur de paramètres des effets aléatoires, $\mathbf{X}$ et $\mathbf{Z}$ deux matrices d'incidence connues, et $\mathbf{e}$ le vecteur d'erreur aléatoire.

Dans ce modèle, si nous avons r effets aléatoires, $\mathbf{Zu}$ peut être décomposé comme suit :

$$\mathbf{Z}\mathbf{u} = \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_r \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_r \end{bmatrix} = \sum_{k=1}^{r} \mathbf{Z}_k \mathbf{u}_k$$

où $\mathbf{u}_k$ d'ordre q_k, est le vecteur des effets aléatoires pour le facteur k avec les suppositions $\mathbb{E}(\mathbf{u}_k) = \mathbf{0}$, $\mathrm{Var}(\mathbf{u}_k) = \sigma_k^2 \mathbf{I}_{q_k}$ $\forall$ k, et $\mathrm{Cov}(\mathbf{u}_k, \mathbf{u}'_{k'}) = \mathbf{0}$ pour $k \neq k'$,

$$q = \sum_{k=1}^{r} q_k$$

Aussi $\mathbb{E}(\mathbf{e}) = \mathbf{0}$, $\mathrm{Var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}_n$ et $\mathrm{Cov}(\mathbf{u}_k, \mathbf{e}') = \mathbf{0}$ $\forall$ k. La fonction $\mathbb{E}(\cdot)$ indique l'espérance.

En posant $\mathbf{u}_0 = \mathbf{e}$, $q_0 = n$ et $\mathbf{Z}_0 = \mathbf{I}_n$ comme dans la présentation de Rao et Kleffe (1988), l'équation (4.1) devient

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{r} \mathbf{Z}_k \mathbf{u}_k \tag{4.2}$$

et $\mathbf{V} = \sum_{k=0}^{r} \mathbf{Z}_k \mathbf{Z}_k' \sigma_k^2$

L'estimation des paramètres $\boldsymbol{\beta}$ et σ_k^2 peut se faire concomitamment au moyen des méthodes de vraisemblance ML, *Maximum likelihood* ou REML, *Restricted or residual maximum likelihood*. Pour chacune de ces méthodes, une fonction log-vraisemblance est maximisée par rapport aux paramètres inconnus. La fonction log-vraisemblance pour la méthode ML (REML étant une variation de ML) s'écrit,

$$l \quad = \quad (-1/2)\big[\log|\mathbf{V}| + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + n\log(2\pi)\big]$$

En dérivant la fonction l par rapport à $\boldsymbol{\beta}$ et à chacun des σ_k^2 et en annulant ces dérivées, nous obtenons $r+1$ équations pour σ_k^2 et une équation pour $\boldsymbol{\beta}$.

$$\begin{cases} \partial l/\partial \boldsymbol{\beta} & = 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} - 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \\ \partial l/\partial \sigma_k^2 & = -(1/2)\Big[\mathrm{tr}(\mathbf{V}^{-1}\mathbf{V}_k') - (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{V}_k'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\Big] \end{cases}$$

où $\mathbf{V}_k' = \partial\mathbf{V}/\partial\sigma_k^2$

Les solutions de ce système d'équations ne sont généralement pas obtenues de façon explicite. Pour résoudre ce problème d'optimisation, l'on a recours à des algorithmes itératifs tels que celui de Newton-Raphson ou l'algorithme EM, *Expectation maximization*.

Ces deux méthodes d'itérations requièrent des valeurs initiales pour les paramètres inconnus. L'algorithme EM permet de s'approcher de la région de

l'optimum plus rapidement mais la progression vers l'optimum est ensuite lente. Au contraire, celui de Newton-Raphson, malgré qu'il soit instable loin de l'optimum, permet une convergence vers celui-ci beaucoup plus rapidement une fois dans sa région.

L'algorithme de Newton-Raphson utilise un développement de premier ordre de la fonction score c'est-à-dire du gradient de la fonction log-vraisemblance autour de l'estimation du paramètre à la m^{e} itération pour fournir l'estimation à la $(m+1)^{\text{e}}$ itération. Chaque étape dans l'algorithme nécessite le calcul de la fonction score et de sa dérivée, la matrice hessienne de la log-vraisemblance.

L'algorithme EM (Meng and van Dyk, 1997) permet l'estimation de paramètres dans des modèles avec des données incomplètes. L'argumentaire de l'utilisation de cet algorithme dans le cadre du modèle mixte est fourni en détail par Searle, Cassella et McCulloch (1992, pp. 297-303). Ainsi, les effets aléatoires sont-ils vus comme des données non observées. Searle et al. considèrent alors que si ces effets aléatoires étaient connus, l'estimation des paramètres inconnus pourrait facilement se faire. En effet, il suffirait d'adopter une démarche à deux étapes.

D'abord, estimer la variance de chaque effet aléatoire par

$$\widehat{\sigma}_k^2 = (1/q_k) \sum_{q=1}^{q_k} (u_k^q)^2 = \mathbf{u}_k' \mathbf{u}_k / q_k$$

où $\mathbf{u}_k$ d'ordre q_k est supposé gaussien d'espérance nulle et de variance σ_k^2.

Ensuite, déduire ces effets aléatoires supposés connus du vecteur des données $\mathbf{Y}$ et appliquer une régression OLS, *Ordinary least squares*, sur le modèle suivant

$$\mathbf{Y} - \sum_{k=1}^{r} \mathbf{Z}_k \mathbf{u}_k \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_0^2 \mathbf{I}_n)$$

Mais comme ces effets aléatoires ne sont pas connus en réalité, l'algorithme EM permet de calculer les valeurs conditionnelles de $\mathbf{u}_k'\mathbf{u}_k$ à utiliser à la place de $\mathbf{u}_k'\mathbf{u}_k$ et les valeurs conditionnelles de $\mathbf{u}_k$ à la place de $\mathbf{u}_k$.

Selon la terminologie de l'algorithme EM, dans le cas du modèle mixte, les données observées $\mathbf{Y}$ sont appelées données incomplètes. Les données complètes comportent en plus les effets aléatoires non observés $\mathbf{u}_1, \ldots, \mathbf{u}_r$. Nous rappelons ci-dessous cet algorithme pour l'estimation des paramètres fondée sur une variation de ML publiée par Laird (1982). Les valeurs calculées $\sigma_k^{2(m)}$ de σ_k^2 sont obtenues après la m^{e} itération et sont utilisées pour la mise à jour de la variance $\mathbf{V}^{-1(m)}$.

Étape 0 — Poser $m = 0$ et choisir des valeurs initiales $\sigma_k^{2(0)}$

Étape 1 (Étape-E) — Calculer
$Q(\sigma^2; \sigma^{2(m)}) = \mathbb{E}_{\sigma^{2(m)}}(\mathbf{u}_k'\mathbf{u}_k \mid \mathbf{Y})$
$= q_k\sigma_k^{2(m)} + \sigma_k^{4(m)}\left[\mathbf{Y}'\mathbf{P}^{(m)}\mathbf{Z}_k\mathbf{Z}_k'\mathbf{P}^{(m)}\mathbf{Y} - \text{tr}(\mathbf{Z}_k'\mathbf{V}^{-1(m)}\mathbf{Z}_k)\right]$
où $\mathbf{P}^{(m)} = \mathbf{V}^{-1(m)} - \mathbf{V}^{-1(m)}\mathbf{X}\big(\mathbf{X}'\mathbf{V}^{-1(m)}\mathbf{X}\big)^{-}\mathbf{X}'\mathbf{V}^{-1(m)}$

Étape 2 (Étape-M) — Determiner $\sigma_k^{2(m+1)}$ qui maximise $Q(\sigma^2; \sigma^{2(m)})$ c'est-à-dire, tel que $Q(\sigma^{2(m+1)}; \sigma^{2(m)}) \geqslant Q(\sigma^2; \sigma^{2(m)})$. Alors,
$\sigma_k^{2(m+1)} = \mathbb{E}_{\sigma^{2(m)}}(\mathbf{u}_k'\mathbf{u}_k \mid \mathbf{Y})/q_k$ pour $k = 0, 1, \cdots, r$

Étape 3 — A la convergence, prendre $\widehat{\sigma}_k^2 = \sigma_k^{2(m+1)}$ et alors calculer
$\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1(m+1)}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m+1)}\mathbf{Y}$
sinon ajouter 1 à m et retourner à *l'Étape 1.*

4.2 La régression PLS sur un modèle de variance connue

La régression PLS est une méthode d'estimation particulière utilisée pour les modèles linéaires avec l'éventualité $n < p$. Pour ce type d'analyse, l'objectif est de prédire $\mathbf{Y}$ par des combinaisons linéaires des colonnes de $\mathbf{X}$ appelées variables latentes. Il est habituellement mis en oeuvre à l'aide de l'algorithme NIPALS, *Nonlinear estimation by iterative partial least squares* (Wold 1966 ; de Jong 1993) où le calcul des variables latentes est effectué simultanément avec un ensemble de régressions par OLS. Cependant, ces régressions sont adéquates seulement dans le cas des erreurs sur $\mathbf{Y}$ iid.

L'algorithme PLS appliqué à un modèle de variance connue est effectué en remplaçant les régressions OLS sur les variables latentes par des régressions GLS, *General least squares*, dont voici la description.

1. Centrer et éventuellement réduire $\mathbf{X}$ et $\mathbf{Y}$: $\mathbf{x}_0 = \mathbf{X}$, $\mathbf{y}_0 = \mathbf{Y}$
2. Pour $h = 1, \cdots, H$ avec $1 \leqslant H \leqslant \text{rang}(\mathbf{X})$
   (a) Calculer les p-vecteurs $\mathbf{w}_h = [w_h^1 \cdots w_h^p]'$
   où $w_h^p = \text{Cov}(\mathbf{x}_h^p, \mathbf{y}_h) / \sqrt{\sum_p \text{Cov}^2(\mathbf{x}_h^p, \mathbf{y}_h)}$ et $\mathbf{x}_h^p$ la p^{e} colonne de $\mathbf{x}_h$
   (b) Normer $\mathbf{w}_h$: $\mathbf{w}_h = \mathbf{w}_h / \parallel \mathbf{w}_h \parallel$
   (c) Calculer les variables latentes PLS $\mathbf{t}_h = \mathbf{x}_{h-1}\mathbf{w}_h$
   (d) Calculer c_h par régression GLS de $\mathbf{y}_{h-1}$ sur $\mathbf{t}_h$
   $\mathbf{y}_{h-1} = \mathbf{t}_h c_h + \mathbf{y}_h$ où $\text{Var}(\mathbf{y}_{h-1}) = \mathbf{V}$
   $c_h = \left(\mathbf{t}_h' \mathbf{V}^{-1} \mathbf{t}_h\right)^{-} \mathbf{t}_h' \mathbf{V}^{-1} \mathbf{y}_{h-1}$
   (e) Calculer $\mathbf{p}_h$ par régression de $\mathbf{x}_{h-1}$ sur $\mathbf{t}_h$
   $\mathbf{x}_{h-1} = \mathbf{t}_h \mathbf{p}_h' + x_h$ d'où $\mathbf{p}_h' = \left(\mathbf{t}_h' \mathbf{t}_h\right)^{-1} \mathbf{t}_h' \mathbf{x}_{h-1}$
   (f) Calculer les résidus $\mathbf{x}_h$ et $\mathbf{y}_h$

(g) Alors $\mathbf{Y} = \mathbf{t}_1 c_1 + \cdots + \mathbf{t}_h c_h + \mathbf{y}_h$

Ainsi, le seul changement par rapport à l'algorithme PLS classique est le remplacement de la régression OLS par la régression GLS au point 2.(d).

4.3 La méthode PLS-Mixte

Les méthodes de vraisemblances, ML ou REML, comme techniques pour estimer les paramètres fixes et les composantes de variance dans un modèle linéaire mixte, ne sont applicables que dans le cas classique où le nombre de régresseurs est faible devant le nombre d'observations, c'est-à-dire $n > p$. Dans ce cas, comme nous l'avons vu, un algorithme itératif tel que l'algorithme EM est nécessaire pour obtenir l'estimation des paramètres inconnus.

Pour traiter du cas $n < p$, nous proposons d'imbriquer une méthode de réduction de dimension telle que la régression PLS dans l'algorithme EM. Puisqu'il s'agira d'estimer des composantes de variance dans un contexte de réduction de dimension, nous avons appelé cette technique PLS-Mixte.

Avant cette méthode proposée, dans les modèles où il y avait plus de régresseurs que d'observations et plusieurs sources de variation, l'estimation des paramètres inconnus se faisait simplement par régression PLS, c'est-à-dire sans tenir compte précisément des sources de variation. Aussi, avons-nous comparé l'estimation faite par simple régression PLS à celle faite par notre méthode en utilisant le critère MSEP, *Mean square error of prediction*, dans les différentes applications de ce chapitre. La question de la convergence sera abordée plus loin.

4.3.1 La méthode PLS-Mixte sur un modèle à effets aléatoires indépendants de variances homogènes

Nous nous plaçons dans le cadre du modèle 4.2 et considérons que $n < p$. Nous proposons la méthode PLS-Mixte qui consiste donc à imbriquer une méthode de réduction de dimension telle que la régression PLS dans l'algorithme EM. Cette méthode d'estimation est fondée sur ML et ses variantes. L'estimation par ML est réalisée en maximisant la vraisemblance de $\mathbf{Y}$ par rapport aux paramètres inconnus.

Suivant l'algorithme EM, nous prenons des valeurs de départ pour les paramètres inconnus. Ces valeurs permettent de calculer les variables latentes obtenues à partir du modèle $(\mathbf{X}, \mathbf{Y})$ vu que la variance $\mathbf{V}^{(0)}$ est connue, les valeurs initiales $\sigma_k^{2(0)}$ étant choisies. Sur ces variables latentes, sont calculées les composantes de variance $\sigma_k^{2(1)}$ comme dans le cas classique de l'algorithme EM. Avec les valeurs $\sigma_k^{2(1)}$, ces étapes sont répétées avec les composantes $\sigma_k^{2(0)}$ remplacées par les estimations courantes $\sigma_k^{2(1)}$. Ce processus itératif est alors continué jusqu'à convergence. Nous décrivons ci-dessous l'algorithme de la méthode proposée avant de le détailler plus loin.

Étape 0 — Mettre $m = 0$ et choisir des valeurs de départ $\sigma_k^{2(0)}$

Étape 1 (Étape-E) — Centrer et réduire $\mathbf{X}$ et $\mathbf{Y}$

Étape 1.1 — Réduire la dimension de l'espace des régresseurs en déterminant les h variables latentes $\mathbf{T}^{(m)} = [\mathbf{t}_1^{(m)} \cdots \mathbf{t}_h^{(m)}]$ vu que la variance $\mathbf{V}^{(m)}$ est connue

Étape 1.2 — Calculer

$$
\begin{aligned}
&Q(\sigma^2; \sigma^{2(m)}) \\
&= \mathbb{E}_{\sigma^{2(m)}}(\mathbf{u}_i' \mathbf{u}_k \mid \mathbf{Y}) \\
&= q_k \sigma_k^{2(m)} + \sigma_k^{4(m)} \left[\mathbf{Y}' \mathbf{P}^{(m)} \mathbf{Z}_k \mathbf{Z}_k' \mathbf{P}^{(m)} \mathbf{Y} - \mathrm{tr}(\mathbf{Z}_k' \mathbf{V}^{-1(m)} \mathbf{Z}_k) \right] \\
&\text{où } \mathbf{P}^{(m)} = \mathbf{V}^{-1(m)} - \mathbf{V}^{-1(m)} \mathbf{T}^{(m)} \left(\mathbf{T}^{(m)\prime} \mathbf{V}^{-1(m)} \mathbf{T}^{(m)} \right)^{-} \mathbf{T}^{(m)\prime} \mathbf{V}^{-1(m)}
\end{aligned}
$$

Étape 2 (Étape-M) — Determiner $\sigma_k^{2(m+1)}$ qui maximise $\mathbb{E}_{\sigma^{2(m)}}(\mathbf{u}_i' \mathbf{u}_k \mid \mathbf{Y})$ càd, tel que $Q_{\mathbf{T}(m)}(\sigma^{2(m+1)}; \sigma^{2(m)}) \geqslant Q_{\mathbf{T}(m)}(\sigma^2; \sigma^{2(m)})$

$\sigma_k^{2(m+1)} = \mathbb{E}_{\sigma^{2(m)}}(\mathbf{u}_i' \mathbf{u}_k \mid \mathbf{Y}) / q_k$ pour $k = 0, 1, \cdots, r$

Étape 3 — Si convergence, prendre $\widehat{\sigma}_k^2 = \sigma_k^{2(m+1)}$ et alors calculer

$\mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1(m+1)}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m+1)}\mathbf{Y}$

sinon ajouter 1 à m et retourner à *Étape 1.*

Les variables latentes sont ainsi recalculées au début de chaque itération dans l'algorithme EM avec les composantes de variance mises à jour. Cet algorithme de la méthode PLS-Mixte fondée sur une variation de ML est détaillé de la façon suivante.

Étape 0 Mettre $m = 0$ et choisir des valeurs de départ $\sigma_k^{2(0)}$

Étape 1 Centrer et réduire $\mathbf{X}$ et $\mathbf{Y}$: $\mathbf{x}_0 = \mathbf{X}$, $\mathbf{y}_0 = \mathbf{Y}$

Étape 1.1 Pour $h = 1, 2, \ldots, \text{rang}(\mathbf{X})$

(a) Calculer les p-vector $\mathbf{w}_h = [w_h^1 \cdots w_h^p]'$
où $w_h^p = \text{Cov}(\mathbf{x}_h^p, \mathbf{y}_h) / \sqrt{\sum_p \text{Cov}^2(\mathbf{x}_h^p, \mathbf{y}_h)}$ et $\mathbf{x}_h^p$ la p^{e} colonne de $\mathbf{x}_h$

(b) Normer $\mathbf{w}_h$: $\mathbf{w}_h = \mathbf{w}_h / \parallel \mathbf{w}_h \parallel$

(c) Calculer les variables latentes $\mathbf{t}_h^{(m)} = \mathbf{x}_{h-1}\mathbf{w}_h$

(d) Calculer c_h par régression GLS de $\mathbf{y}_{h-1}$ sur $\mathbf{t}_h^{(m)}$
$\mathbf{y}_{h-1} = \mathbf{t}_h^{(m)} c_h + \mathbf{y}_h$ où $\text{Var}(\mathbf{y}_{h-1}) = \mathbf{V}^{(m)} = \sum_{k=0}^{r} \mathbf{Z}_k \mathbf{Z}_k' \sigma_k^{2(m)}$
$c_h = \left(\mathbf{t}_h^{(m)\prime} \mathbf{V}^{-1(m)} \mathbf{t}_h^{(m)}\right)^{-} \mathbf{t}_h^{(m)\prime} \mathbf{V}^{-1(m)} \mathbf{y}_{h-1}$

(e) Calculer $\mathbf{p}_h$ par régression de $\mathbf{x}_{h-1}$ sur $\mathbf{t}_h^{(m)}$
$\mathbf{x}_{h-1} = \mathbf{t}_h^{(m)} \mathbf{p}_h' + \mathbf{x}_h$ d'où $\mathbf{p}_h' = \left(\mathbf{t}_h^{(m)\prime} \mathbf{t}_h^{(m)}\right)^{-1} \mathbf{t}_h^{(m)\prime} \mathbf{x}_{h-1}$

(f) Calculer les résidus $\mathbf{x}_h$ and $\mathbf{y}_h$

(g) Finalement $\mathbf{Y} = \mathbf{T}^{(m)}\mathbf{C} + \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{u}_i$
où $\mathbf{T}^{(m)} = [\mathbf{t}_1^{(m)} \cdots \mathbf{t}_h^{(m)}]$ et $\mathbf{C} = [c_1 \cdots c_h]'$

Étape 1.2 Calculer
$\sigma_k^{2(m+1)} = \sigma_k^{2(m)} + (\sigma_k^{4(m)} / q_k) \left[\mathbf{Y}'\mathbf{P}^{(m)}\mathbf{Z}_k\mathbf{Z}_k'\mathbf{P}^{(m)}\mathbf{Y} - \text{tr}(\mathbf{Z}_k'\mathbf{V}^{-1(m)}\mathbf{Z}_k)\right]$
où $\mathbf{P}^{(m)} = \mathbf{V}^{-1(m)} - \mathbf{V}^{-1(m)}\mathbf{T}^{(m)}\left(\mathbf{T}^{(m)\prime}\mathbf{V}^{-1(m)}\mathbf{T}^{(m)}\right)^{-}\mathbf{T}^{(m)\prime}\mathbf{V}^{-1(m)}$

Étape 2 Si convergence, prendre $\widehat{\sigma}_k^2 = \sigma_k^{2(m+1)}$; sinon ajouter 1 à m et retourner à l'Étape 1.1

La procédure REML, quant à elle, maximise la vraisemblance de certaines combinaisons linéaires des éléments de $\mathbf{Y}$ par rapport aux paramètres inconnus (McCulloch and Searle, 2001). Notre méthode fondée sur cette procédure REML est effectuée en remplaçant l'Étape 1.2 par

$$\sigma_i^{2(m+1)} = \sigma_i^{2(m)} + (\sigma_i^{4(m)}/q_i)\left[\mathbf{Y}'\mathbf{P}^{(m)}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{Y} - \mathrm{tr}(\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{Z}_i)\right]$$

où $\mathbf{P}^{(m)}$ reste défini comme ci-dessus.

Convergence de l'algorithme de la méthode PLS-Mixte :

Les propriétés de convergence de l'algorithme EM restent valides avec cette méthode PLS-Mixte, en particulier sa convergence monotone. Par exemple, pour la monotonocité de l'algorithme EM, nous devons montrer que la fonction de vraisemblance, $L(\sigma^2 \mid \mathbf{Y})$ ne décroît pas après une itération, c'est-à-dire,

$$L_{\mathbf{T}^{(m)}}(\sigma_k^{2(m+1)} \mid \mathbf{Y}) \geqslant L_{\mathbf{T}^{(m-1)}}(\sigma_k^{2(m)} \mid \mathbf{Y}) \tag{4.3}$$

où la valeur de la vraisemblance $L_{\mathbf{T}^{(m)}}(\sigma_k^{2(m+1)} \mid \mathbf{Y})$ est calculée à la $(m+1)^{\mathrm{e}}$ itération en utilisant les valeurs actuelles de $\mathbf{T}^{(m)}$ comme régresseurs et la valeur de la vraisemblance $L_{\mathbf{T}^{(m-1)}}(\sigma_k^{2(m)} \mid \mathbf{Y})$ calculée à la m^{e} itération avec les valeurs de $\mathbf{T}^{(m-1)}$

Pour l'algorithme EM classique, le fait que la vraisemblance augmente à chaque itération est un résultat bien connu qui a été montré par Dempster, Laird et Rubin (1977) et discuté plus tard par Wu (1983) et par McLachlan et Krishnan (1997).

Nous nous sommes fondés sur ce résultat pour montrer l'inégalité (4.3). La fonction de densité conditionnelle des données complètes $[\mathbf{Y}, \mathbf{u}']$ sachant les données incomplètes $\mathbf{Y}$, où $\mathbf{u}' = [\mathbf{u}_1', \cdots, \mathbf{u}_r']$, est égale à

$$k([\mathbf{Y}, \mathbf{u}'] \mid \mathbf{Y}, \sigma^2) = \frac{f_{[\mathbf{Y},\mathbf{u}']}([\mathbf{Y}, \mathbf{u}'] \mid \sigma^2)}{f(\mathbf{Y} \mid \sigma^2)} \tag{4.4}$$

Alors la fonction log-vraisemblance des données incomplètes peut s'écrire comme suit :

$$\begin{aligned} l(\sigma^2 \mid \mathbf{Y}) &= \ln\big(f(\mathbf{Y} \mid \sigma^2)\big) \\ &= \ln\big(f_{[\mathbf{Y},\mathbf{u}']}([\mathbf{Y},\mathbf{u}'] \mid \sigma^2)/k([\mathbf{Y},\mathbf{u}'] \mid \mathbf{Y},\sigma^2)\big) \\ &= \ln\big(f_{[\mathbf{Y},\mathbf{u}']}([\mathbf{Y},\mathbf{u}'] \mid \sigma^2)\big) - \ln\big(k([\mathbf{Y},\mathbf{u}'] \mid \mathbf{Y},\sigma^2)\big) \\ &= l_{[\mathbf{Y},\mathbf{u}']}(\sigma^2 \mid [\mathbf{Y},\mathbf{u}']) - \ln\big(k([\mathbf{Y},\mathbf{u}'] \mid \mathbf{Y},\sigma^2)\big) \end{aligned}$$

L'espérance de cette équation est prise par rapport à la distribution conditionnelle des données complètes sachant les données incomplètes, et en utilisant $\sigma^{2(m)}$ à la place de σ^2. Ainsi,

$$\begin{aligned} \mathbb{E}_{\sigma^2}\big[l(\sigma^2 \mid \mathbf{Y}) \mid \mathbf{Y}\big] &= l_{\mathbf{T}^{(m-1)}}(\sigma^2 \mid \mathbf{Y}) \\ &= \mathbb{E}_{\sigma^{2(m)}}\big[l_{[\mathbf{Y},\mathbf{u}']}(\sigma^2 \mid [\mathbf{Y},\mathbf{u}']) \mid \mathbf{Y}\big] \\ &\quad -\mathbb{E}_{\sigma^{2(m)}}\big[\ln\big(k_{\mathbf{T}^{(m-1)}}([\mathbf{Y},\mathbf{u}'] \mid \mathbf{Y},\sigma^2)\big) \mid \mathbf{Y}\big] \end{aligned} \tag{4.5}$$

Nous avons, à l'Étape-E de l'algorithme du modèle PLS-Mixte avec une variance connue (section 2.2), la relation suivante :

$$\begin{aligned} \mathbb{E}_{\sigma^{2(m)}}\big[l_{[\mathbf{Y},\mathbf{u}']}(\sigma^2 \mid [\mathbf{Y},\mathbf{u}']) \mid \mathbf{Y}\big] &\propto \mathbb{E}_{\sigma^2}(\mathbf{u}'\mathbf{u} \mid \mathbf{Y}) \\ &= Q_{\mathbf{T}^{(m-1)}}(\sigma^2;\sigma^{2(m)}) \end{aligned}$$

et en posant

$$G_{\mathbf{T}^{(m-1)}}(\sigma^2;\sigma^{2(m)}) = \mathbb{E}_{\sigma^{2(m)}}\big[\ln\big(k_{\mathbf{T}^{(m-1)}}([\mathbf{Y},\mathbf{u}'] \mid \mathbf{Y},\sigma^2)\big) \mid \mathbf{Y}\big]$$

l'équation (4.5) devient

$$l_{\mathbf{T}^{(m-1)}}(\sigma^2 \mid \mathbf{Y}) = Q_{\mathbf{T}^{(m-1)}}(\sigma^2;\sigma^{2(m)}) - G_{\mathbf{T}^{(m-1)}}(\sigma^2;\sigma^{2(m)})$$

Nous pouvons maintenant calculer

$$\begin{aligned} l_{\mathbf{T}^{(m)}}(\sigma^{2(m+1)} \mid \mathbf{Y}) - l_{\mathbf{T}^{(m-1)}}(\sigma^{2(m)} \mid \mathbf{Y}) &= Q_{\mathbf{T}^{(m)}}(\sigma^{2(m+1)}; \sigma^{2(m)}) \\ &\quad -G_{\mathbf{T}^{(m)}}(\sigma^{2(m+1)}; \sigma^{2(m)}) \\ &\quad -Q_{\mathbf{T}^{(m-1)}}(\sigma^{2(m)}; \sigma^{2(m)}) \\ &\quad +G_{\mathbf{T}^{(m-1)}}(\sigma^{2(m)}; \sigma^{2(m)}) \end{aligned}$$

La quantité $Q_{\mathbf{T}^{(m)}}(\sigma^{2(m+1)}; \sigma^{2(m)}) - Q_{\mathbf{T}^{(m-1)}}(\sigma^{2(m)}; \sigma^{2(m)})$ est positive car $\sigma^{2(m+1)}$ est choisie tel que

$$Q_{\mathbf{T}^{(m)}}(\sigma^{2(m+1)}; \sigma^{2(m)}) \geq Q_{\mathbf{T}^{(m-1)}}(\sigma^{2}; \sigma^{2(m)}) \ \forall \ \ \sigma^2$$

Et, $\forall\ \sigma^2$

$$\begin{aligned} G_{\mathbf{T}^{(m-1)}}(\sigma^{2}; \sigma^{2(m)}) - G_{\mathbf{T}^{(m-1)}}(\sigma^{2(m)}; \sigma^{2(m)}) &= \mathbb{E}_{\sigma^{2(m)}}\left[\ln\left(k_{\mathbf{T}^{(m-1)}}([\mathbf{Y}, \mathbf{u}'] \mid \mathbf{Y}, \sigma^2)\right) \mid \mathbf{Y}\right] \\ &\quad -\mathbb{E}_{\sigma^{2(m)}}\left[\ln\left(k_{\mathbf{T}^{(m-1)}}([\mathbf{Y}, \mathbf{u}'] \mid \mathbf{Y}, \sigma^{2(m)})\right) \mid \mathbf{Y}\right] \\ &\leq \ln(1) \\ &= 0 \end{aligned}$$

Ce résultat est montré par McLachlan et Krishnan (1997) pour l'algorithme EM classique où les régresseurs sont considérés comme fixes. Ici, la quantité

$$G_{\mathbf{T}^{(m-1)}}(\sigma^{2}; \sigma^{2(m)}) - G_{\mathbf{T}^{(m-1)}}(\sigma^{2(m)}; \sigma^{2(m)})$$

est prise, $\forall\ \sigma^2$, par rapport aux mêmes régresseurs $\mathbf{T}^{(m-1)}$. Ce qui ne change donc pas le résultat.

Nous avons alors

$$l_{\mathbf{T}^{(m)}}(\sigma^{2(m+1)} \mid \mathbf{Y}) - l_{\mathbf{T}^{(m-1)}}(\sigma^{2(m)} \mid \mathbf{Y}) \geq 0$$

La méthode PLS-Mixte, utilise entre autres, la technique de réduction de dimension. Et dans ce sens, il faudra déterminer la dimension du modèle retenue. Pour cela, nous choisissons un nombre maximum h ($h < \text{rang}(X)$) de variables latentes au départ. La méthode itérative décrite plus haut permettra de calculer et d'actualiser tour à tour ces h variables latentes et les composantes de variance. Avec ces composantes de variance estimées au final sur le modèle à h variables latentes, les PRESS, *Prediction error sum of squares* (Stone 1974), des h sous-modèles avec respectivement 1, 2, $\cdots$, h régresseurs sont calculés. La dimension retenue est celle du sous-modèle à plus faible PRESS.

Illustration avec des données de NIRS

Les algorithmes présentés ci-dessus sont testés et appliqués à titre illustratif sur un jeu de données de Rami, Dufour, Trouche, Fliedel, Mestres, Davrieux, Blanchard et Hamon (1998).

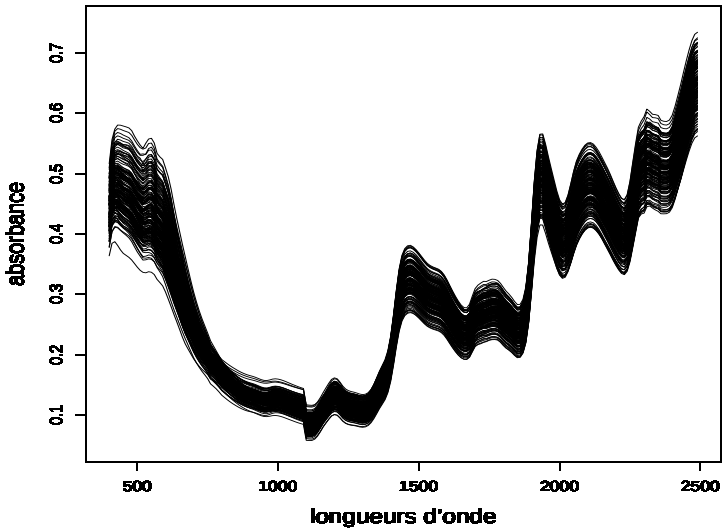
Il s'agit d'une population de lignées recombinantes de sorgho étudiée à la station expérimentale de l'INERA au Burkina Faso. Cette population est obtenue par une méthode "modified single-seed descent" avec le génotype IS 2807 (collection de l'ICRISAT) considéré comme femelle croisée avec le génotype 249 (collection du CIRAD) considéré comme mâle. Cette population de 90 individus fut récoltée en 1995 à la génération F_7, à partir d'un dispositif expérimental en lattice 9×10 avec trois répétitions.

Nous avons seulement retenu la teneur en protéine parmi un ensemble de mesures biochimiques effectuées sur les lignées, pour nous conformer aux conditions de notre méthode ; c'est-à-dire qu'il n'est considéré qu'une seule variable à expliquer.

Pour cette teneur en protéine, il a été utilisé une méthode traditionnelle de mesure, fiable mais longue à effectuer ; les valeurs mesurées sur toutes les lignées constituent le vecteur $\mathbf{Y}$.

Parallèlement à cette méthode de mesure, la technique NIRS, *Near infrared reflectance spectroscopy*, beaucoup plus rapide a été effectuée ; les spectres d'absorption obtenus pour 1050 longueurs d'onde constituent la matrice $\mathbf{X}$. La série de longueurs d'onde considérée est constituée d'une séquence de 400 à 1098 par 2 (figure 4.1).

FIG. 4.1 – Spectres d'absorption proches de l'infrarouge pour une population de lignées de sorgho récoltées en 1995 à la Station expérimentale de l'INERA au Burkina Faso.



Au delà de notre souci d'utiliser ces données pour tester et éprouver notre méthode, l'intérêt de cette étude chez le praticien, pourrait être de calibrer la technique NIRS sur le sorgho pour la teneur en protéine dans le but d'obtenir une méthode de mesure rapide en tenant compte de la structure particulière de la covariance des erreurs induite par le dispositif expérimental.

Pour la mise en oeuvre de la méthode PLS-Mixte sur ces données, il existe cependant des données manquantes à savoir la troisième répétition et trois lignées des deux autres répétitions. Eu égard au nombre important de longueurs d'onde (1050) et la corrélation massive habituellement constatée sur les données de NIRS, nous avons décidé de conserver systématiquement une longueur d'onde toutes les cinq.

Le modèle considéré pour ces données s'écrit comme suit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e} \tag{4.6}$$

où $\mathbf{u}_1$ d'ordre 2 est l'effet aléatoire de la répétition et $\mathbf{u}_2$ d'ordre 20 l'effet aléatoire du bloc hiérarchisé dans la répétition.

Finalement, la dimension de $\mathbf{Y}$ est 174×1 et la dimension de $\mathbf{X}$ est 174×210. Aussi, $\mathbf{Z}_1$ est une matrice 174×2 et $\mathbf{u}_1$ un vecteur 2×1 tandis que $\mathbf{Z}_2$ est une matrice 174×20 et $\mathbf{u}_2$ un vecteur 20×1.

Les hypothèses pour ce modèle sont par conséquent

$$\begin{cases} \mathbf{u}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I}_2) \\ \mathbf{u}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_2^2\mathbf{I}_{20}) \\ \mathbf{u}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I}_{174}) \qquad \text{avec} \;\; \mathbf{u}_0 = \mathbf{e} \end{cases}$$

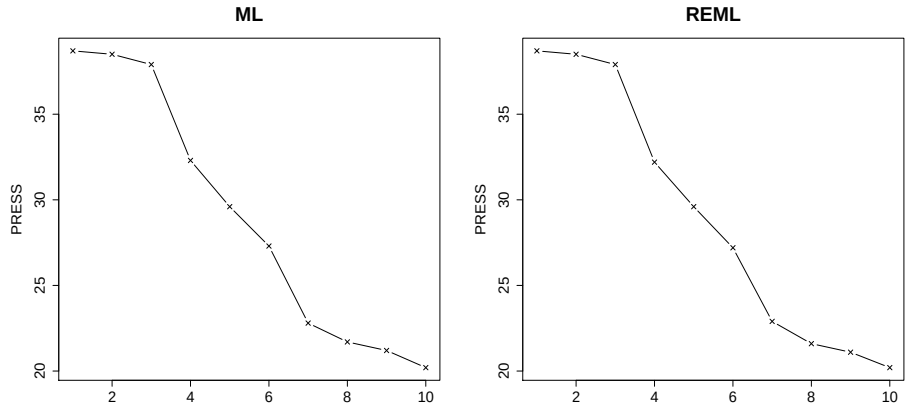
Nous avons pour ce modèle, plus de régresseurs que d'observations (210 contre 174) et trois composantes de variance à estimer (σ_1^2 pour l'effet aléatoire de la répétition, σ_2^2 pour l'effet aléatoire du bloc et σ_0^2 pour l'erreur résiduelle aléatoire). Nous allons donc avoir recours à la méthode PLS-Mixte fondée sur ML et sur REML pour estimer les quatres paramètres $\boldsymbol{\beta}$, σ_1^2, σ_2^2 et σ_0^2.

Pour le calcul du PRESS, nous avons, comme annoncé plus haut, choisi un nombre maximum de 10 variables latentes. Avec un modèle de dimension 10,

nous avons utilisé la méthode PLS-Mixte pour estimer les trois composantes de variance. Par la suite, pour chacun des sous-modèles de dimension 1 à 10, nous avons enlevé successivement chaque lignée (dans les deux répétitions) dont nous avons prédit sa teneur en protéine par les lignées restantes en utilisant une régression GLS avec les composantes de variance estimées sur le modèle de dimension 10.

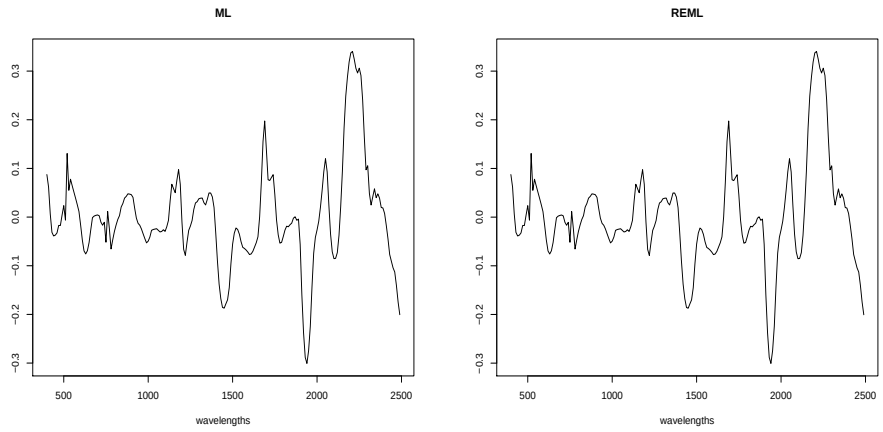
Le PRESS décroît continuellement avec le nombre de variables latentes pour les deux modèles fondés sur ML et sur REML quand le nombre de composantes maximum avait initialement été fixé à 10. Cependant, nous avons décidé pour ces modèles de réduire leur espace à 7 dimensions étant donné que le PRESS n'y est pas trop différent de ses valeurs minimales (figure 4.2).

FIG. 4.2 – Évolution du PRESS selon le nombre de variables latentes quand le nombre maximal de variables latentes avait initialement été fixé à dix pour les données de NIRS.



Par rapport à l'objectif de cette étude, il serait question de sélectionner les régresseurs influents. Nous pouvons alors pour cela, considérer les coefficients les plus grands en valeur absolue des régresseurs pour le modèle sélectionné. Nous notons ainsi que les bandes d'absorption localisées autour de 2210, 1940, 1690 et 1450 ont les plus grands coefficients (figure 4.3).

FIG. 4.3 – Évolution des coefficients $\widehat{\boldsymbol{\beta}}$ (modèle (4.6)) selon les différentes longueurs d'onde pour les modèles à 7 variables latentes fondés sur ML et REML des données de NIRS.



Les estimations de σ_k^2 pour ML et pour REML sont montrées au tableau 4.1. Nous notons que la variance de l'effet bloc hiérarchisé dans la répétition est deux fois plus faible que celle de l'effet répétition.

	ML	REML
σ_1^2	0,11273	0,11276
σ_2^2	0,06248	0,06309
σ_0^2	0,46839	0,48917

TAB. 4.1 – Estimation des composantes de variance pour les modèles à 7 variables latentes fondés sur ML et REML des données de NIRS.

Au tableau 4.2, nous avons le MSEP du modèle 4.6 analysé par régression PLS et le MSEP du même modèle analysé par la méthode PLS-Mixte fondée sur ML et sur REML. La prédiction par la méthode PLS-Mixte, comme nous nous y attendons, de la teneur en protéine du sorgho par les bandes d'absorption, quand il y a plusieurs sources de variation, est meilleure que celle faite par simple régression PLS.

		MSEP
PLS-Mixte	ML	0,64637
	REML	0,64639
PLS		0,91479

TAB. 4.2 – MSEP des modèles de la méthode PLS-Mixte et de la régression PLS pour les données de NIRS.

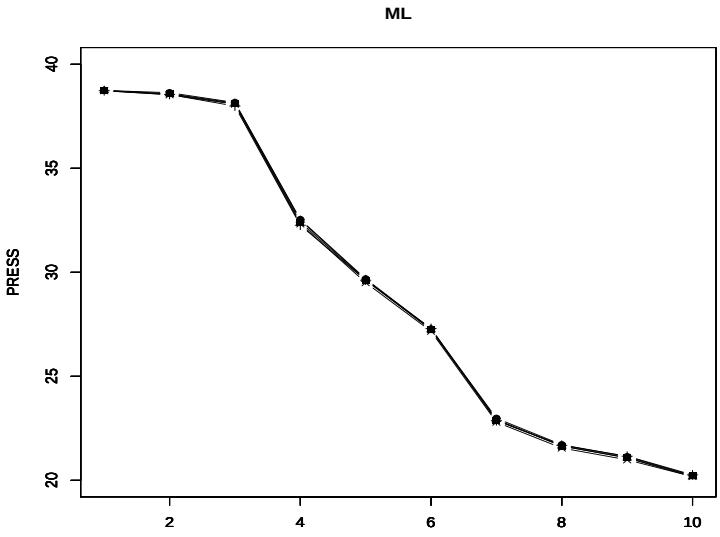
Toutefois, la détermination de la dimension optimale à travers le calcul du PRESS est ici liée à h, le nombre maximum de variables latentes choisi au départ. En effet, les composantes de variance sont estimées avec les h variables latentes et utilisées par la suite respectivement pour chaque sous-modèle à 1, 2, $\cdots$, h variables latentes pour estimer $\boldsymbol{\beta}$ et finalement pour calculer les PRESS. Cependant, le choix de h semble ne pas avoir un impact important sur le calcul des PRESS dans notre example. A la figure 4.4, nous avons présenté les valeurs des PRESS calculées pour un nombre maximum de variables latentes choisi initialement pour être égal respectivement à 10, 20, 30, 40 et 50. L'évolution du PRESS jusqu'à la dixième variable latente n'est quasiment pas affectée par le choix initial du nombre maximum de variables latentes.

4.3.2 La méthode PLS-Mixte sur un modèle à effets aléatoires corrélés de variances hétérogènes

La méthode PLS-Mixte, présentée ci-dessus et appliquée aux données de NIRS, a été écrite avec les hypothèses classiques de normalité des effets aléatoires et surtout de la forme particulière des matrices de variance de ces effets aléatoires. En effet, il a fallu supposer que cette variance pour chaque effet aléatoire $\mathbf{u}_k$ était de la forme $\sigma_k^2\mathbf{I}_{q_k}$.

Ici, nous allons légèrement relâcher cette hypothèse et considérer que la variance pour chaque effet aléatoire pourrait être de la forme $\sigma_k^2\boldsymbol{\Delta}_k$ où $\boldsymbol{\Delta}_k$ est

FIG. 4.4 – Variation du PRESS selon les dix premières variables latentes quand le nombre maximum de variables a initialement été fixé de façon successive à 10, 20, 30, 40 et 50 pour le modèle fondé sur ML pour les données de NIRS. Le modèle fondé sur REML donne des résultats voisins de ceux présentés ici.



une matrice quelconque symétrique connue. Nous allons alors voir comment la méthode PLS-Mixte s'écrira avec ces nouvelles données du problème.

Nous nous plaçons toujours dans le cadre du modèle (4.2) que nous allons rappeler.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{r} \mathbf{Z}_k \mathbf{u}_k \tag{4.7}$$

Nous supposons cette fois-ci donc que $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \boldsymbol{\Delta}_k)$, d'où $\mathbf{V} = \sum_{k=0}^{r} \mathbf{Z}_k \boldsymbol{\Delta}_k \mathbf{Z}_k' \sigma_k^2$.

Avant de présenter notre méthode avec un modèle où $n < p$ et la variance $\mathbf{V}$ des observations écrite comme ci-dessus, nous traiterons dans un premier

temps, du cas plus simple $n > p$ où nous verrons comment s'effectuent les estimations des paramètres inconnus $\boldsymbol{\beta}$ et σ_k^2 à l'aide de l'algorithme EM fondé sur ML.

4.3.2.1 Le modèle mixte à effets aléatoires corrélés

Dans cette partie, nous sommes intéressés à trouver les estimations des paramètres du modèle mixte où les effets aléatoires sont corrélés. Pour cela, comme dans le cas classique du modèle mixte à effets aléatoires homogènes et indépendants, il est calculé la log-vraisemblance de la loi jointe de $\mathbf{Y}$ et $\mathbf{u} = [\ \mathbf{u}_1' \quad \mathbf{u}_2' \quad \cdots \quad \mathbf{u}_r'\ ]'$, qui est maximisée par rapport aux paramètres inconnus. En annulant les dérivées de cette fonction log-vraisemblance par rapport à chacun des paramètres, il est obtenu un système d'équations dont les solutions ne sont généralement pas obtenues de façon explicite. L'algorithme EM permet alors de trouver ces estimations.

Searle et al. (1992) ont montré en détail les calculs pour l'estimation des paramètres inconnus du modèle mixte par maximum de vraisemblance avec l'algorithme EM. Ces calculs qui ne concernent que le cas des effets aléatoires indépendants à variances homogènes, nous allons les reprendre ici en considérant ces effets aléatoires corrélés.

4.3.2.1.1 Calcul de la vraisemblance des données complètes

Avec les suppositions du modèle (4.7), la loi jointe de $\mathbf{Y}$ et $\mathbf{u}$ est gaussienne de la forme $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ où

$$\boldsymbol{\mu} = \begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \qquad \text{et} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{V} & \text{Cov}(\mathbf{Y}, \mathbf{u}') \\ \text{Cov}(\mathbf{u}, \mathbf{Y}') & \text{Var}(\mathbf{u}) \end{bmatrix} \tag{4.8}$$

Nous avons

$$\text{Cov}(\mathbf{Y}, \mathbf{u}'_{k'}) = \text{Cov}(\mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{r} \mathbf{Z}_k \mathbf{u}_k, \mathbf{u}'_{k'}) = \mathbf{Z}_{k'} \text{Cov}(\mathbf{u}_{k'}, \mathbf{u}'_{k'}) = \sigma^2_{k'} \mathbf{Z}_{k'} \boldsymbol{\Delta}_{k'} \tag{4.9}$$

car $\text{Var}(\mathbf{u}_k) = \sigma^2_k \boldsymbol{\Delta}_k \ \forall \ k$, et $\text{Cov}(\mathbf{u}_k, \mathbf{u}'_{k'}) = \mathbf{0}$ pour $k \neq k'$.

donc $\text{Cov}(\mathbf{Y}, \mathbf{u}') = \left[\ \sigma^2_1 \mathbf{Z}_1 \boldsymbol{\Delta}_1 \quad \cdots \quad \sigma^2_r \mathbf{Z}_r \boldsymbol{\Delta}_r \ \right]$

$$\text{Cov}(\mathbf{u}, \mathbf{Y}') = \begin{bmatrix} \sigma^2_1 \boldsymbol{\Delta}_1 \mathbf{Z}'_1 \\ \vdots \\ \sigma^2_r \boldsymbol{\Delta}_r \mathbf{Z}'_r \end{bmatrix}$$

$$\text{et } \text{Var}(\mathbf{u}) = \begin{bmatrix} \sigma^2_1 \boldsymbol{\Delta}_1 & & \\ 0 & \ddots & 0 \\ & & \sigma^2_r \boldsymbol{\Delta}_r \end{bmatrix}$$

Ce qui implique la fonction de densité suivante

$$f_{\mathbf{Y}, \mathbf{u}_1, \cdots, \mathbf{u}_r}(\mathbf{Y}, \mathbf{u}_1, \cdots, \mathbf{u}_r) = (2\pi)^{-\frac{1}{2}\sum_{k=0}^{r} q_k} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp(-\frac{1}{2}Q) \tag{4.10}$$

où

$$Q = \begin{bmatrix} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' & \mathbf{u}'_1 & \cdots & \mathbf{u}'_r \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_r \end{bmatrix} \tag{4.11}$$

Dans les développements qui vont suivre, nous allons trouver les valeurs de $|\boldsymbol{\Sigma}|$ et de $\boldsymbol{\Sigma}^{-1}$ dans le but de simplifier la densité (4.10). Il s'agira par la suite, de dériver cette densité, pour avoir les estimations des paramètres.

Pour calculer $|\boldsymbol{\Sigma}|$, nous faisons appel au résultat déjà établi (Searle et al. p 453, 1992) : Si $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ et $\mathbf{D}$ sont quatre matrices avec les bonnes dimensions, alors

$$\left| \begin{array}{cc} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{array} \right| = |\mathbf{D}||\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}|$$

Ainsi,

$$|\mathbf{\Sigma}| = \left| \begin{array}{cc} \mathbf{V} & \mathrm{Cov}(\mathbf{Y}, \mathbf{u}') \\ \mathrm{Cov}(\mathbf{u}, \mathbf{Y}') & \mathrm{Var}(\mathbf{u}) \end{array} \right| = |\mathrm{Var}(\mathbf{u})||\mathbf{V} - \mathrm{Cov}(\mathbf{Y}, \mathbf{u}')(\mathrm{Var}(\mathbf{u}))^{-1}\mathrm{Cov}(\mathbf{u}, \mathbf{Y}')|$$

Or,

$$|\mathrm{Var}(\mathbf{u})| = \left| \begin{array}{ccc} \sigma_1^2\mathbf{\Delta}_1 & & \\ 0 & \ddots & 0 \\ & & \sigma_r^2\mathbf{\Delta}_r \end{array} \right| = |\sigma_1^2\mathbf{\Delta}_1| \times \cdots \times |\sigma_r^2\mathbf{\Delta}_r| = \prod_{k=1}^{r} (\sigma_k^2)^{q_k}|\mathbf{\Delta}_k|$$

car chaque matrice $\mathbf{\Delta}_k$ est carrée de nombre de lignes q_k

D'autre part,

$$\begin{aligned} &\mathrm{Cov}(\mathbf{Y}, \mathbf{u}')(\mathrm{Var}(\mathbf{u}))^{-1}\mathrm{Cov}(\mathbf{u}, \mathbf{Y}') \\ &= [\ \sigma_1^2\mathbf{Z}_1\mathbf{\Delta}_1 \quad \cdots \quad \sigma_r^2\mathbf{Z}_r\mathbf{\Delta}_r\ ] \left[\begin{array}{ccc} \sigma_1^{-2}\mathbf{\Delta}_1^{-1} & & \\ 0 & \ddots & 0 \\ & & \sigma_r^{-2}\mathbf{\Delta}_r^{-1} \end{array} \right] \left[\begin{array}{c} \sigma_1^2\mathbf{\Delta}_1\mathbf{Z}_1' \\ \vdots \\ \sigma_r^2\mathbf{\Delta}_r\mathbf{Z}_r' \end{array} \right] \\ &= \sum_{k=1}^{r} \sigma_k^2\mathbf{Z}_k\mathbf{\Delta}_k\mathbf{Z}_k' \end{aligned}$$

D'où

$$\begin{aligned} |\mathbf{\Sigma}| &= \prod_{k=1}^{r} (\sigma_k^2)^{q_k}|\mathbf{\Delta}_k||\mathbf{V} - \sum_{k=1}^{r} \sigma_k^2\mathbf{Z}_k\mathbf{\Delta}_k\mathbf{Z}_k'| \qquad (4.12) \\ &= \prod_{k=1}^{r} (\sigma_k^2)^{q_k}|\mathbf{\Delta}_k||\sigma_0^2\mathbf{\Delta_0}| \\ &= \prod_{k=0}^{r} (\sigma_k^2)^{q_k}|\mathbf{\Delta}_k| \end{aligned}$$

Pour calculer $\mathbf{\Sigma}^{-1}$, nous utilisons les résultats des inverses généralisées (Searle et al. p 450, 1992) et établissons que

$$\begin{aligned}\mathbf{\Sigma}^{-1} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathrm{Var}(\mathbf{u}))^{-1} \end{bmatrix} \\ &\quad + \begin{bmatrix} \mathbf{I} \\ -(\mathrm{Var}(\mathbf{u}))^{-1}\mathrm{Cov}(\mathbf{u},\mathbf{Y}') \end{bmatrix} \sigma_0^{-2}\mathbf{\Delta}_0^{-1} \begin{bmatrix} \mathbf{I} & -\mathrm{Cov}(\mathbf{Y},\mathbf{u}')(\mathrm{Var}(\mathbf{u}))^{-1} \end{bmatrix}\end{aligned}$$

Or,

$$(\mathrm{Var}(\mathbf{u}))^{-1}\mathrm{Cov}(\mathbf{u},\mathbf{Y}') = \begin{bmatrix} \sigma_1^{-2}\mathbf{\Delta}_1^{-1} & & \\ 0 & \ddots & 0 \\ & & \sigma_r^{-2}\mathbf{\Delta}_r^{-1} \end{bmatrix} \begin{bmatrix} \sigma_1^2\mathbf{\Delta}_1\mathbf{Z}_1' \\ \vdots \\ \sigma_r^2\mathbf{\Delta}_r\mathbf{Z}_r' \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_r' \end{bmatrix}$$

Par ailleurs,

$$\mathrm{Cov}(\mathbf{Y},\mathbf{u}')(\mathrm{Var}(\mathbf{u}))^{-1} = \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_r \end{bmatrix}$$

Par conséquent,

$$\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1}$$

où $\mathbf{\Sigma}_1^{-1} = \begin{bmatrix} \mathbf{0} & & \mathbf{0} & \\ & \begin{bmatrix} \sigma_1^{-2}\mathbf{\Delta}_1^{-1} & & \\ \mathbf{0} & \ddots & \mathbf{0} \\ & & \sigma_r^{-2}\mathbf{\Delta}_r^{-1} \end{bmatrix} & & \end{bmatrix}$

et $\mathbf{\Sigma}_2^{-1} = \begin{bmatrix} \mathbf{I} \\ -\begin{bmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_r' \end{bmatrix} \end{bmatrix} \sigma_0^{-2}\mathbf{\Delta}_0^{-1} \begin{bmatrix} \mathbf{I} & -\begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_r \end{bmatrix} \end{bmatrix}$

Nous allons donc remplacer $\mathbf{\Sigma}^{-1}$ par $\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1}$ dans (4.11) pour avoir la valeur de Q.

Ainsi,

$$Q = Q_1 + Q_2$$

$$\text{où } Q_1 = \begin{bmatrix} (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})' & \mathbf{u}_1' & \cdots & \mathbf{u}_r' \end{bmatrix} \mathbf{\Sigma}_1^{-1} \begin{bmatrix} \mathbf{Y}-\mathbf{X}\boldsymbol{\beta} \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_r \end{bmatrix}$$

$$\text{et } Q_2 = \begin{bmatrix} (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})' & \mathbf{u}_1' & \cdots & \mathbf{u}_r' \end{bmatrix} \mathbf{\Sigma}_2^{-1} \begin{bmatrix} \mathbf{Y}-\mathbf{X}\boldsymbol{\beta} \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_r \end{bmatrix}$$

$$\begin{aligned} Q_1 &= \begin{bmatrix} (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})' & \mathbf{u}_1' & \cdots & \mathbf{u}_r' \end{bmatrix} \begin{bmatrix} \mathbf{0} & & \mathbf{0} & \\ & \begin{bmatrix} \sigma_1^{-2}\mathbf{\Delta}_1^{-1} & & \\ \mathbf{0} & \ddots & \mathbf{0} \\ & & \sigma_r^{-2}\mathbf{\Delta}_r^{-1} \end{bmatrix} & \end{bmatrix} \begin{bmatrix} \mathbf{Y}-\mathbf{X}\boldsymbol{\beta} \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_r \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \begin{bmatrix} \mathbf{u}_1' & \cdots & \mathbf{u}_r' \end{bmatrix} \begin{bmatrix} \sigma_1^{-2}\mathbf{\Delta}_1^{-1} & & \\ \mathbf{0} & \ddots & \mathbf{0} \\ & & \sigma_r^{-2}\mathbf{\Delta}_r^{-1} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \mathbf{Y}-\mathbf{X}\boldsymbol{\beta} \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_r \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{u}_1' & \cdots & \mathbf{u}_r' \end{bmatrix} \begin{bmatrix} \sigma_1^{-2}\mathbf{\Delta}_1^{-1} & & \\ \mathbf{0} & \ddots & \mathbf{0} \\ & & \sigma_r^{-2}\mathbf{\Delta}_r^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_r \end{bmatrix} \\ &= \sum_{k=1}^{r} \frac{\mathbf{u}_k'\mathbf{\Delta}_k^{-1}\mathbf{u}_k}{\sigma_k^2} \end{aligned}$$

Pour calculer Q_2, trouvons d'abord une formule simplifiée de $\mathbf{\Sigma}_2^{-1}$

$$\begin{aligned}
\boldsymbol{\Sigma}_2^{-1} &= \begin{bmatrix} \mathbf{I} \\ -\begin{bmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_r' \end{bmatrix} \end{bmatrix} \sigma_0^{-2} \boldsymbol{\Delta}_0^{-1} \begin{bmatrix} \mathbf{I} & -\begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_r \end{bmatrix} \end{bmatrix} \\
&= \sigma_0^{-2} \begin{bmatrix} \boldsymbol{\Delta}_0^{-1} & -\boldsymbol{\Delta}_0^{-1} \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_r \end{bmatrix} \\ -\begin{bmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_r' \end{bmatrix} \boldsymbol{\Delta}_0^{-1} & \begin{bmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_r' \end{bmatrix} \boldsymbol{\Delta}_0^{-1} \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_r \end{bmatrix} \end{bmatrix}
\end{aligned}$$

Ainsi $Q_2 = \sigma_0^{-2} \begin{bmatrix} Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_r \end{bmatrix}$

avec $Q_{21} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Delta}_0^{-1} - \begin{bmatrix} \mathbf{u}_1' & \cdots & \mathbf{u}_r' \end{bmatrix} \begin{bmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_r' \end{bmatrix} \boldsymbol{\Delta}_0^{-1}$

et $Q_{22} = -(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Delta}_0^{-1} \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_r \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1' & \cdots & \mathbf{u}_r' \end{bmatrix} \begin{bmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_r' \end{bmatrix} \boldsymbol{\Delta}_0^{-1} \begin{bmatrix} \mathbf{Z}_1 & \cdots & \mathbf{Z}_r \end{bmatrix}$

D'où,

$$
\begin{aligned}
Q_2 &= \sigma_0^{-2}\left((\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Delta}_0^{-1} - \begin{bmatrix}\mathbf{u}_1' & \cdots & \mathbf{u}_r'\end{bmatrix}\begin{bmatrix}\mathbf{Z}_1'\\ \vdots \\ \mathbf{Z}_r'\end{bmatrix}\boldsymbol{\Delta}_0^{-1}\right)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad -\sigma_0^{-2}\left((\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Delta}_0^{-1}\begin{bmatrix}\mathbf{Z}_1 & \cdots & \mathbf{Z}_r\end{bmatrix} - \begin{bmatrix}\mathbf{u}_1' & \cdots & \mathbf{u}_r'\end{bmatrix}\begin{bmatrix}\mathbf{Z}_1'\\ \vdots \\ \mathbf{Z}_r'\end{bmatrix}\boldsymbol{\Delta}_0^{-1}\begin{bmatrix}\mathbf{Z}_1 & \cdots & \mathbf{Z}_r\end{bmatrix}\right) \\
&\quad \times \begin{bmatrix}\mathbf{u}_1\\ \vdots \\ \mathbf{u}_r\end{bmatrix} \\
&= \sigma_0^{-2}\left((\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Delta}_0^{-1} - \sum_{k=1}^{r}\mathbf{u}_k'\mathbf{Z}_k'\boldsymbol{\Delta}_0^{-1}\right)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad -\sigma_0^{-2}\left((\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Delta}_0^{-1}\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k - \sum_{k=1}^{r}\mathbf{u}_k'\mathbf{Z}_k'\boldsymbol{\Delta}_0^{-1}\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right) \\
&= \sigma_0^{-2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Delta}_0^{-1}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right) - \sigma_0^{-2}\sum_{k=1}^{r}\mathbf{u}_k'\mathbf{Z}_k'\boldsymbol{\Delta}_0^{-1}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right) \\
&= \sigma_0^{-2}\left((\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Delta}_0^{-1} - \sum_{k=1}^{r}\mathbf{u}_k'\mathbf{Z}_k'\boldsymbol{\Delta}_0^{-1}\right)\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right) \\
&= \sigma_0^{-2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right)'\boldsymbol{\Delta}_0^{-1}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right)
\end{aligned}
$$

En ajoutant Q_1 et Q_2, nous obtenons

$$
Q = \sum_{k=1}^{r}\frac{\mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k}{\sigma_k^2} + \sigma_0^{-2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right)'\boldsymbol{\Delta}_0^{-1}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right) \quad (4.13)
$$

A présent, comme nous avons les valeurs de $|\boldsymbol{\Sigma}|$ et de $\boldsymbol{\Sigma}^{-1}$ et donc de Q, en remplaçant (4.12) et (4.13) dans (4.10) et en passant au logarithme, nous avons

$$
\begin{aligned}
l &= -\frac{1}{2}\left(\sum_{k=0}^{r} q_k\right)\log 2\pi - \frac{1}{2}\sum_{k=0}^{r}(q_k \log\sigma_k^2 + \log|\boldsymbol{\Delta}_k|) - \frac{1}{2}\sum_{k=1}^{r}\frac{\mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k}{\sigma_k^2} \\
&\quad -\frac{\sigma_0^{-2}}{2}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right)' \boldsymbol{\Delta}_0^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k\right) \\
&= -\frac{1}{2}\left(\sum_{k=0}^{r} q_k\right)\log 2\pi - \frac{1}{2}\sum_{k=0}^{r}(q_k \log\sigma_k^2 + \log|\boldsymbol{\Delta}_k|) - \frac{1}{2}\sum_{k=0}^{r}\frac{\mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k}{\sigma_k^2}
\end{aligned}
$$

car $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k = \mathbf{e} = \mathbf{u}_0$

De ce fait, il suffira de dériver la fonction l ci-dessus par rapport à chacun des paramètres inconnus du modèle (4.7) pour trouver les estimateurs suivants

$$
\widehat{\sigma}_k^2 = \mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k/q_k, \qquad k = 0, 1, \cdots r \tag{4.14}
$$

et

$$
\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Delta}_0^{-1}\mathbf{X})^{-}\mathbf{X}'\boldsymbol{\Delta}_0^{-1}\left(\mathbf{Y} - \sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k)\right) \tag{4.15}
$$

4.3.2.1.2 Calcul des espérances conditionnelles

Mais, comme les $\mathbf{u}_k$ ne sont pas connus, nous avons besoin des espérances conditionnelles, sachant les données incomplètes $\mathbf{Y}$, de

$$
\mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k
$$

et de

$$
\mathbf{Y} - \sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k
$$

c'est-à-dire $\mathbb{E}(\mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k|\mathbf{Y})$ et $\mathbb{E}(\mathbf{u}_k|\mathbf{Y})$ et par conséquent de $\mathbb{E}(\mathbf{Y} - \sum_{k=1}^{r}\mathbf{Z}_k\mathbf{u}_k|\mathbf{Y})$

Pour cela, nous pouvons utiliser le résultat suivant (Searle p ,) : si

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \right)$$

alors la loi conditionnelle de $\mathbf{x}_1$ sachant $\mathbf{x}_2$ est

$$\mathbf{x}_1|\mathbf{x}_2 \sim \mathcal{N}[\boldsymbol{\mu}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}]$$

Avec $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k^2\boldsymbol{\Delta}_k)$ et $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, nous avons

$$\begin{bmatrix} \mathbf{u}_k \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \sigma_k^2\boldsymbol{\Delta}_k & \mathrm{Cov}(\mathbf{u}_k, \mathbf{Y}') \\ \mathrm{Cov}(\mathbf{Y}, \mathbf{u}_k') & \mathbf{V} \end{bmatrix} \right)$$

D'après (4.9), il vient

$$\mathbf{u}_k|\mathbf{Y} \sim \mathcal{N}\left(\sigma_k^2\boldsymbol{\Delta}_k\mathbf{Z}_k'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \sigma_k^2\boldsymbol{\Delta}_k - \sigma_k^2\boldsymbol{\Delta}_k\mathbf{Z}_k'\mathbf{V}^{-1}\sigma_k^2\mathbf{Z}_k\boldsymbol{\Delta}_k\right)$$

D'où

$$\mathbb{E}(\mathbf{u}_k|\mathbf{Y}) = \sigma_k^2\boldsymbol{\Delta}_k\mathbf{Z}_k'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.16}$$

Ensuite, en utilisant le théorème suivant :

Théorème 1 *(Searle p 231, 1987)*

$$\text{si } \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V}) \text{ alors } \mathbb{E}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \ \forall \ \mathbf{A}$$

et d'après ce qui précède, nous avons

$$\begin{aligned}\mathbb{E}(\mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k|\mathbf{Y}) &= \mathrm{tr}(\sigma_k^2\boldsymbol{\Delta}_k^{-1}\boldsymbol{\Delta}_k - \sigma_k^4\boldsymbol{\Delta}_k^{-1}\boldsymbol{\Delta}_k\mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{Z}_k\boldsymbol{\Delta}_k)\\ &\quad +\sigma_k^4(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}_k\boldsymbol{\Delta}_k'\boldsymbol{\Delta}_k^{-1}\boldsymbol{\Delta}_k\mathbf{Z}_k'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\end{aligned}$$

D'où

$$\begin{aligned}\mathbb{E}(\mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k|\mathbf{Y}) &= \mathrm{tr}(\sigma_k^2\mathbf{I}_{q_k} - \sigma_k^4\mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{Z}_k\boldsymbol{\Delta}_k)\\ &\quad +\sigma_k^4(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}_k\boldsymbol{\Delta}_k'\mathbf{Z}_k'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \qquad (4.17)\end{aligned}$$

Nous pouvons alors établir l'algorithme EM dans le cas du modèle linéaire mixte où chaque effet aléatoire est de variance $\sigma_k^2\boldsymbol{\Delta}_k$. L'estimation, fondée sur ML, s'effectue en plusieurs itérations. Des valeurs initiales $\sigma_k^{2(0)}$ et $\boldsymbol{\beta}^{(0)}$ sont choisies au départ. A la m^{e} itération de l'Étape-E, sont calculées les espérances conditionnelles suivantes, à partir de (4.17),

$$\begin{aligned}\widehat{s}_k^{(m)} &= \mathbb{E}(\mathbf{u}_k'\boldsymbol{\Delta}_k^{-1}\mathbf{u}_k|\mathbf{Y})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}\ et\ \sigma_k^2=\sigma_k^{2(m)}} \qquad (4.18)\\ &= \mathrm{tr}(\sigma_k^{2(m)}\mathbf{I}_{q_k}) - \sigma_k^{4(m)}\mathrm{tr}(\mathbf{Z}_k'\mathbf{V}^{-1(m)}\mathbf{Z}_k\boldsymbol{\Delta}_k)\\ &\quad +\sigma_k^{4(m)}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^{(m)})'\mathbf{V}^{-1(m)}\mathbf{Z}_k\boldsymbol{\Delta}_k'\mathbf{Z}_k'\mathbf{V}^{-1(m)}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^{(m)})\\ &= q_k\sigma_k^{2(m)} + (\sigma_k^{4(m)})\\ &\quad \times\Big[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^{(m)})'\mathbf{V}^{-1(m)}\mathbf{Z}_k\boldsymbol{\Delta}_k'\mathbf{Z}_k'\mathbf{V}^{-1(m)}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^{(m)}) - \mathrm{tr}(\mathbf{Z}_k'\mathbf{V}^{-1(m)}\mathbf{Z}_k\boldsymbol{\Delta}_k)\Big]\end{aligned}$$

Et, à partir de (4.16)

$$
\begin{aligned}
\widehat{b}^{(m)} &= \mathbb{E}(\mathbf{Y} - \sum_{k=1}^{r} \mathbf{Z}_k \mathbf{u}_k | \mathbf{Y})_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)} \ et \ \sigma_k^2=\sigma_k^{2(m)}} \qquad (4.19)\\
&= \mathbf{Y} - \sum_{k=1}^{r} \mathbf{Z}_k \boldsymbol{\Delta}_k \mathbf{Z}_k' \sigma_k^{2(m)} \mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)})\\
&= \mathbf{Y} - (\mathbf{V}^{(m)} - \sigma_0^{2(m)} \boldsymbol{\Delta}_0) \mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)})\\
&= \mathbf{Y} - \mathbf{V}^{(m)} \mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)}) + \sigma_0^{2(m)} \boldsymbol{\Delta}_0 \mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)})\\
&= \mathbf{X}\boldsymbol{\beta}^{(m)} + \sigma_0^{2(m)} \boldsymbol{\Delta}_0 \mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)})
\end{aligned}
$$

A l'Étape-M, pour trouver les estimateurs, il suffit de prendre

$$
\begin{aligned}
\sigma_k^{2(m+1)} &= \widehat{s}_k^{(m)} / q_k \qquad (4.20)\\
&= \sigma_k^{2(m)} + (\sigma_k^{4(m)} / q_k)\\
&\quad \times \left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)})' \mathbf{V}^{-1(m)} \mathbf{Z}_k \boldsymbol{\Delta}_k' \mathbf{Z}_k' \mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)}) - \mathrm{tr}(\mathbf{Z}_k' \mathbf{V}^{-1(m)} \mathbf{Z}_k \boldsymbol{\Delta}_k) \right]
\end{aligned}
$$

Et, à partir de (4.15)

$$
\begin{aligned}
\mathbf{X}\widehat{\boldsymbol{\beta}}^{(m+1)} &= \mathbf{X}(\mathbf{X}'\boldsymbol{\Delta}_0^{-1}\mathbf{X})^{-}\mathbf{X}'\boldsymbol{\Delta}_0^{-1} \left(\mathbf{X}\boldsymbol{\beta}^{(m)} + \sigma_0^{2(m)} \boldsymbol{\Delta}_0 \mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)}) \right) \qquad (4.21)\\
&= \mathbf{X}(\mathbf{X}'\boldsymbol{\Delta}_0^{-1}\mathbf{X})^{-}\mathbf{X}'\boldsymbol{\Delta}_0^{-1}\mathbf{X}\boldsymbol{\beta}^{(m)} + \sigma_0^{2(m)} \mathbf{X}(\mathbf{X}'\boldsymbol{\Delta}_0^{-1}\mathbf{X})^{-}\mathbf{X}'\boldsymbol{\Delta}_0^{-1} \boldsymbol{\Delta}_0 \mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)})\\
&= \mathbf{X}\boldsymbol{\beta}^{(m)} + \sigma_0^{2(m)} \mathbf{X}(\mathbf{X}'\boldsymbol{\Delta}_0^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m)} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)})
\end{aligned}
$$

Ce qui donne, de façon itérative, les estimations du vecteur de paramètres des effets fixes (4.21) et de celui des effets aléatoires (4.20) du modèle mixte à effets aléatoires corrélés.

Cependant, pour simplifier les calculs, à la place d'itérer pour obtenir à la fois les valeurs de $\boldsymbol{\beta}$ et de σ_k^2, Laird (1982) suggère de trouver les valeurs de σ_k^2 et seulement, à la fin des itérations, de calculer la valeur de $\boldsymbol{\beta}$. En effet,

à la place de calculer $\mathbf{V}^{-1(m)}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)})$ dans l'équation (4.20), il est calculé $\mathbf{P}^{(m)}\mathbf{Y}$ où

$$\mathbf{P}^{(m)} = \mathbf{V}^{-1(m)} - \mathbf{V}^{-1(m)}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1(m)}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m)}$$

ne dépend pas de $\boldsymbol{\beta}$.

En effet, si $\mathbf{X}\boldsymbol{\beta}^{(m)}$ était l'estimation de $\mathbf{X}\boldsymbol{\beta}$, c'est-à-dire

$$\mathbf{X}\boldsymbol{\beta}^{(m)} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1(m)}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m)}\mathbf{Y}$$

alors

$$\begin{aligned}
\mathbf{V}^{-1(m)}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(m)}) &= \mathbf{V}^{-1(m)}\mathbf{Y} - \mathbf{V}^{-1(m)}\mathbf{X}\boldsymbol{\beta}^{(m)} \\
&= \mathbf{V}^{-1(m)}\mathbf{Y} - \mathbf{V}^{-1(m)}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1(m)}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m)}\mathbf{Y} \\
&= \left(\mathbf{V}^{-1(m)} - \mathbf{V}^{-1(m)}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1(m)}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1(m)}\right)\mathbf{Y} \\
&= \mathbf{P}^{(m)}\mathbf{Y}
\end{aligned}$$

De ce fait, il ne sera pas nécessaire de disposer de $\boldsymbol{\beta}^{(m)}$ dans les itérations et à l'Étape-E. Il ne sera alors plus calculé qu'une seule espérance conditionnelle

$$\begin{aligned}
\sigma_k^{2(m+1)} &= \widehat{s}_k^{(m)}/q_k \qquad (4.22) \\
&= \sigma_k^{2(m)} + (\sigma_k^{4(m)}/q_k)\left[\mathbf{Y}'\mathbf{P}^{(m)}\mathbf{Z}_k\boldsymbol{\Delta}_k'\mathbf{Z}_k'\mathbf{P}^{(m)}\mathbf{Y} - \mathrm{tr}(\mathbf{Z}_k'\mathbf{V}^{-1(m)}\mathbf{Z}_k\boldsymbol{\Delta}_k)\right]
\end{aligned}$$

Nous présentons ci-dessous l'algorithme EM fondé sur ML appliqué aux effets aléatoires corrélés du modèle mixte.

Étape 0 — Mettre $m = 0$ et choisir des valeurs initiales $\sigma_k^{2(0)}$

Étape 1 (Étape-E) — Calculer
$Q(\sigma^2 \mid \sigma^{2(m)})$
$= \mathbb{E}_{\sigma^{2(m)}}(\mathbf{u}_k' \boldsymbol{\Delta}_k^{-1} \mathbf{u}_k \mid \mathbf{Y})$
$= q_k \sigma_k^{2(m)} + \sigma_k^{4(m)} \left[\mathbf{Y}' \mathbf{P}^{(m)} \mathbf{Z}_k \boldsymbol{\Delta}_k' \mathbf{Z}_k' \mathbf{P}^{(m)} \mathbf{Y} - \mathrm{tr}(\mathbf{Z}_k' \mathbf{V}^{-1(m)} \mathbf{Z}_k \boldsymbol{\Delta}_k)\right]$
où $\mathbf{P}^{(m)} = \mathbf{V}^{-1(m)} - \mathbf{V}^{-1(m)} \mathbf{X} \big(\mathbf{X}' \mathbf{V}^{-1(m)} \mathbf{X}\big)^{-} \mathbf{X}' \mathbf{V}^{-1(m)}$

Étape 2 (Étape-M) — Determiner $\sigma_k^{2(m+1)}$ qui maximise $Q(\sigma^2 \mid \sigma^{2(m)})$ c'est-à-dire, tel que $Q(\sigma^{2(m+1)} \mid \sigma^{2(m)}) \geqslant Q(\sigma^2 \mid \sigma^{2(m)})$. Alors,
$\sigma_k^{2(m+1)} = \mathbb{E}_{\sigma^{2(m)}}(\mathbf{u}_k' \boldsymbol{\Delta}_k^{-1} \mathbf{u}_k \mid \mathbf{Y})/q_k$ for $k = 0, 1, \cdots, r$

Étape 3 — A la convergence, prendre $\widehat{\sigma}_k^2 = \sigma_k^{2(m+1)}$ et alors calculer
$\mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}' \mathbf{V}^{-1(m+1)} \mathbf{X})^{-} \mathbf{X}' \mathbf{V}^{-1(m+1)} \mathbf{Y}$
sinon ajouter 1 à m et retourner à *l'Étape 1.*

4.3.2.1.3 Simulations pour la convergence de l'algorithme EM appliqué aux effets aléatoires corrélés du modèle mixte

Pour tester la convergence de cet algorithme et vérifier la qualité des estimations, nous avons effectué des simulations numériques. Nous avons considéré à cet effet le modèle simple suivant :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1 \mathbf{u}_1 + \mathbf{Z}_2 \mathbf{u}_2 + \mathbf{Z}_0 \mathbf{u}_0 \tag{4.23}$$

où la dimension du vecteur des réponses $\mathbf{Y}$ est 120×1, celle de la matrice des observations $\mathbf{X}$ est 120×5 et le vecteur $\boldsymbol{\beta}$ des paramètres fixes est de longueur 5. Aussi, avons-nous supposé

$$\begin{cases} \mathbf{u}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_1^2 \boldsymbol{\Delta}_1) \\ \mathbf{u}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \boldsymbol{\Delta}_2) \\ \mathbf{u}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Delta}_0) \end{cases}$$

où $\boldsymbol{\Delta}_1 = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$, $\boldsymbol{\Delta}_2 = \begin{pmatrix} 3 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 1 & 1 & 2 \end{pmatrix}$ et $\boldsymbol{\Delta}_0 = \mathbf{I}_{120}$.

Nous avons choisi un dispositif équilibré avec le facteur 2 (6 niveaux) hiérarchisé dans le facteur 1 (5 niveaux). Les matrices $\mathbf{Z}_1$ et $\mathbf{Z}_2$ sont donc les matrices d'incidence correspondantes à ces deux facteurs tandis que la matrice $\mathbf{Z}_0$ est l'identité d'ordre 120.

Nous présentons au Tableau 4.3 les résultats des simulations obtenues avec le modèle 4.23 sur 300 jeux de données.

	Valeurs simulées	Valeurs estimées	
		Moyenne	Écart-type
	0,5402	0,5380	0,0937
	0,8948	0,8974	0,0931
$\boldsymbol{\beta}$	0,6476	0,6423	0,0912
	0,9513	0,9526	0,0921
	0,0772	0,0748	0,0933
σ_1^2	0,4653	0,4241	0,4416
σ_2^2	0,2458	0,2471	0,3324
σ_0^2	0,9226	0,8874	0,1233

TAB. 4.3 – Résultats des simulations pour un modèle mixte à effets aléatoires corrélés effectuées sur 300 jeux de données.

Au vu de ces résultats, les valeurs estimées pour les paramètres ne sont pas, en moyenne, éloignées des valeurs simulées.

4.3.2.2 Le modèle PLS-Mixte à effets aléatoires corrélés

Nous venons ainsi de voir, à la section précédente, l'estimation des paramètres fixes et des composantes de variance dans un modèle linéaire mixte, à l'aide de l'algorithme EM, où chaque effet aléatoire est de variance $\sigma_k^2 \mathbf{\Delta}_k$ et où $n > p$.

Le cas $n < p$ avec cette même structure de variance des effets aléatoires, nécessite, comme au chapitre précédent, d'imbriquer la régression PLS, en tant que méthode de réduction de dimension, à l'algorithme EM. Nous proposons l'algorithme PLS-Mixte suivant, fondé sur la variation de ML proposée par Laird

Étape 0 Mettre $m = 0$ et choisir des valeurs de départ $\sigma_k^{2(0)}$

Étape 1 Centrer et réduire $\mathbf{X}$ et $\mathbf{Y}$: $x_0 = \mathbf{X}$, $\mathbf{y}_0 = \mathbf{Y}$

Étape 1.1 For $h = 1, 2, \ldots, \text{rang}(\mathbf{X})$

(a) Calculer les p-vector $\mathbf{w}_h = [w_h^1 \cdots w_h^p]'$
où $w_h^p = \text{Cov}(\mathbf{x}_h^p, \mathbf{y}_h) / \sqrt{\sum_p \text{Cov}^2(\mathbf{x}_h^p, \mathbf{y}_h)}$ et $\mathbf{x}_h^p$ la p^{e} colonne de x_h

(b) Normer $\mathbf{w}_h$: $\mathbf{w}_h = \mathbf{w}_h / \parallel \mathbf{w}_h \parallel$

(c) Calculer les variables latentes $\mathbf{t}_h^{(m)} = \mathbf{x}_{h-1}\mathbf{w}_h$

(d) Calculer c_h par régression GLS de $\mathbf{y}_{h-1}$ sur $\mathbf{t}_h^{(m)}$
$\mathbf{y}_{h-1} = \mathbf{t}_h^{(m)} c_h + \mathbf{y}_h$ où $\text{Var}(\mathbf{y}_{h-1}) = \mathbf{V}^{(m)} = \sum_{k=0}^{r} \mathbf{Z}_k \mathbf{\Delta}_k \mathbf{Z}_k' \sigma_k^{2(m)}$
$c_h = \left(\mathbf{t}_h^{(m)\prime} \mathbf{V}^{-1(m)} \mathbf{t}_h^{(m)}\right)^{-} \mathbf{t}_h^{(m)\prime} \mathbf{V}^{-1(m)} \mathbf{y}_{h-1}$

(e) Calculer $\mathbf{p}_h$ par régression de $\mathbf{x}_{h-1}$ sur $\mathbf{t}_h^{(m)}$
$\mathbf{x}_{h-1} = \mathbf{t}_h^{(m)} \mathbf{p}_h' + x_h$ d'où $\mathbf{p}_h' = \left(\mathbf{t}_h^{(m)\prime} \mathbf{t}_h^{(m)}\right)^{-1} \mathbf{t}_h^{(m)\prime} \mathbf{x}_{h-1}$

(f) Calculer les résidus x_h and $\mathbf{y}_h$

(g) Finalement $\mathbf{Y} = \mathbf{T}^{(m)} \mathbf{C} + \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{u}_i$
où $\mathbf{T}^{(m)} = [\mathbf{t}_1^{(m)} \cdots \mathbf{t}_h^{(m)}]$ et $\mathbf{C} = [c_1 \cdots c_h]'$

Étape 1.2 Calculer
$\sigma_k^{2(m+1)} = \sigma_k^{2(m)} + (\sigma_k^{4(m)} / q_k) \left[\mathbf{Y}' \mathbf{P}^{(m)} \mathbf{Z}_k \mathbf{\Delta}_k' \mathbf{Z}_k' \mathbf{P}^{(m)} \mathbf{Y} - \text{tr}(\mathbf{Z}_k' \mathbf{V}^{-1(m)} \mathbf{Z}_k \mathbf{\Delta}_k)\right]$
où $\mathbf{P}^{(m)} = \mathbf{V}^{-1(m)} - \mathbf{V}^{-1(m)} \mathbf{T}^{(m)} \left(\mathbf{T}^{(m)\prime} \mathbf{V}^{-1(m)} \mathbf{T}^{(m)}\right)^{-} \mathbf{T}^{(m)\prime} \mathbf{V}^{-1(m)}$

Étape 2 Si convergence, prendre $\widehat{\sigma}_k^2 = \sigma_k^{2(m+1)}$; sinon ajouter 1 à m et retourner à *Étape 1.1*

Bien évidemment, le changement par rapport à l'algorithme de la section précédente concerne l'Étape 1.1 (d) et l'Étape 1.2 où l'estimation des σ_k^2 tient compte de la variance $\sigma_k^2 \mathbf{\Delta}_k$ des effets aléatoires $\mathbf{u}_k$.

En résumé, les estimations dans le modèle PLS-Mixte à effets aléatoires corrélés ont été écrites de manière explicite. Néanmoins, pour l'algorithme

itératif proposé à cet effet, la convergence n'a pas été établie et est restée locale.

Chapitre 5

Conclusion générale

Nous nous étions fixés comme objectif principal de modéliser les interactions G×E en fonction de covariables notamment climatiques observées sur le milieu, avec comme application sa prédiction dans une expérimentation multienvironnements sahélienne.

Nous avons commencé par exposer les principales méthodes existantes dans la littérature et montrer leurs limites. Ces méthodes ne sont adaptées dans le contexte sahélien que par la prise en compte de la grande variabilité climatique du milieu d'une année sur l'autre. Or, la plupart de ces méthodes ne considèrent pas justement les caractéristiques du milieu pour y prédire la performance des génotypes. La seule méthode à notre connaissance qui y tient compte - la régression factorielle - ne permet pas de gérer les nombreuses variables climatiques ordinairement mesurées sur ces lieux.

Les modèles de simulation de culture ont aussi été présentés comme une méthode alternative pour évaluer le comportement des génotypes selon les environnements. Cependant, ces modèles tout comme le modèle SarraH que nous avons utilisé pour cette étude, sont rarement paramétrés pour plus d'une variété. Or, pour pouvoir rendre compte des différences de production

des génotypes, il est essentiel que les paramètres de ces modèles puissent changer selon les génotypes.

Nous proposons en conséquence la méthode APLAT, Approximation par linéarisation autour d'un témoin. Cette méthode consiste à linéariser la réponse prédite par un modèle de simulation de culture de tout génotype d'un environnement donné autour du vecteur de paramètres connu d'un génotype de référence. Après cela, certains des paramètres du modèle de simulation de culture, qui peuvent être choisis aux dires d'experts, peuvent être réestimés à l'aide de cette linéarisation locale. L'avantage de cette méthode est de ne pas nécessiter pour tout génotype, l'expérimentation spécifique indispensable à la paramétrisation du témoin. A l'ensemble de ces expérimentations instrumentées requises pour toute nouvelle variété d'intérêt, est substitué un essai multilocal classique qu'on aura pris soin de munir de stations météorologiques simples.

La méthode APLAT a été validée avec les données d'arachide d'un essai pluriannuel mené à la station expérimentale du CERAAS au Sénégal de 1994 à 1998. Pour ces données, chacune des années a été successivement réservé et le rendement des génotypes prédit à l'aide des années restantes, d'abord à l'aide de la méthode APLAT, ensuite à l'aide du modèle moyen. Le modèle moyen est utilisé pour estimer la performance des génotypes pour une année donnée par la moyenne des autres années. Un modèle qui prédit moins bien en moyenne que le modèle moyen n'est pas acceptable. Nous avons trouvé que quatre fois sur cinq, la méthode APLAT s'était montrée meilleure que le modèle moyen.

Ensuite sur les données de 11 sites d'un essai multilocal mené durant la saison des pluies de 2005 au Sénégal, la méthode APLAT a fourni une meilleure prédiction moyenne du rendement des génotypes que le modèle moyen, 10 fois sur 11.

Par la suite, la méthode APLAT a été étendue au cas d'essais à plusieurs composantes de variance, qui devient APLAT-Mixte. Nous avons adjoint au modèle SarraH que nous avions utilisé pour rendre compte du comportement différencié des génotypes dans cette situation de variabilité environnementale accrue, un effet aléatoire de l'environnement responsable d'interactions G×E dont il a fallu estimer la variance. Cette méthode s'appuie sur la technique PLS-Mixte que nous avons proposé pour l'estimation des composantes de variance dans le cas où il y a plus de régresseurs que d'observations. Cette technique PLS-Mixte consiste à imbriquer la régression PLS dans l'algorithme EM et constitue un algorithme itératif d'estimation des paramètres inconnus dans le cas du modèle mixte dans un contexte de réduction de dimension.

Cette technique PLS-Mixte a été illustrée avec des données de NIRS obtenues à partir d'un dispositif expérimental en lattice 9×10 où l'effet du bloc et l'effet de la répétition ont été supposés aléatoires. Sur ces données, nous avons estimé les paramètres fixes et les composantes de variance du modèle avec cette méthode PLS-Mixte fondée sur ML et sur REML. Nous avons comparé ce type d'estimation à celui de la régression PLS qui était habituellement utilisée dans ce genre de situation c'est-à-dire, quand il y avait plus de régresseurs que d'observations. Il a alors été trouvé que les MSEP de la technique PLS-Mixte fondée sur ML et sur REML étaient plus faibles que le MSEP de la régression PLS faite simplement sans tenir compte des différentes sources de variation.

Nous avons voulu adapté la technique PLS-Mixte au cas d'effets aléatoires corrélés et hétérogènes. Pour cela, il a d'abord fallu se placer uniquement dans le cadre du modèle mixte avec de tels effets aléatoires et surseoir au cas de réduction de dimension. Ainsi, les estimations des composantes de variance pour ces effets aléatoires et celles des paramètres ont été écrits. Sur les simulations effectuées pour tester de la convergence et de la qualité des ces estimations, nous avons retrouvé en moyenne des valeurs très proches de

ceux des paramètres inconnus simulés. Cependant, l'algorithme écrit pour le cas PLS-Mixte avec des effets aléatoires corrélés et hétérogènes s'est heurté à des difficultés de convergence.

Il nous parait nécessaire d'attirer l'attention sur les limites de notre travail qui sont multiples. Tout d'abord, la méthode APLAT s'appuie essentiellement sur un modèle de simulation de culture. Et à ce titre, l'efficacité de cette méthode à bien gérer le comportement variétal selon les environnements est intrinsèquement lié à la capacité du modèle de simulation utilisé à bien exprimer les interactions G×E.

Ensuite, dans ce travail, il est proposé de réestimer les paramètres des génotypes utilisés dans le cadre d'un modèle de simulation de culture pour pouvoir mieux prédire leurs rendements. Or, de tels modèles de simulation disposent généralement d'un ensemble important de paramètres qui ont été estimés seulement pour un témoin, au moyen d'expérimentations spécifiques. Pour apprécier la différence de productivité des variétés selon les environnements, il serait nécessaire de réestimer tous les paramètres pour tout nouveau génotype avant l'utilisation de ces modèles. Ce qui peut être complexe et fastidieux. Mais comme tous les paramètres n'agissent pas de la même façon sur la performance des variétés, nous avons proposé d'en réestimer seulement quelques uns. Pour cela, nous avons effectué ce tri sélectif par dires d'experts pour gagner du temps et aussi parce qu'en général les promoteurs d'un modèle de simulation de culture ont une assez bonne vue du comportement général et individuel des paramètres variétaux. Néanmoins, cette prise en compte des connaissances *a priori* sur les paramètres peut être couplée à une analyse de sensibilité qui permet d'avoir l'impact quantitatif de ces paramètres sur la production des génotypes. Une autre voie est d'attacher une loi de probabilité aux paramètres, qui permet d'intégrer les connaissances *a priori* des experts sur les variétés. Cette méthode, non loin de celle que nous avons utilisé dans le cadre de ce travail, se place dans le cadre bayésien. Les dires d'experts,

pour chaque paramètre, peuvent ainsi être considérés comme une réalisation d'une variable aléatoire dont on connaît la loi *a priori*. Il s'agira par la suite de prendre comme estimation de chaque paramètre variétal l'espérance de la loi *a posteriori* obtenue avec les données.

Enfin, nous avons utilisé dans le cadre de cette étude des données d'essais multienvironnements. La courte série de cinq années de l'essai pluriannuel et le nombre important d'abandons de génotypes au cours de cet essai n'autorisent pas de conclure directement quant à l'efficience de la méthode APLAT. A l'essai multilocal, la série était plus longue, 11 lieux, mais ce qui gêne en réalité au Sahel, c'est moins la variabilité climatique entre les lieux que celle d'une année sur l'autre pour un même lieu. Pour cela, ce qu'il aurait fallu faire si le temps l'avait permis, c'était de répéter l'essai multilocal sur deux à cinq ans tout en diminuant le nombre de lieux. Ce qui est préférable à mener le même nombre d'essais sur plusieurs lieux en une seule année (Talbot, 1997).

En somme, nous avons proposé de prédire les interactions G×E par linéarisation d'un modèle de simulation de culture. Dès lors qu'il est possible, pour tout génotype, de réestimer les paramètres nécessaires à un modèle de simulation de culture, cette méthode permet la comparaison de variétés dans des environnements où elles n'ont pas été observées. Bien entendu, pour cela faudrait-il disposer de longues séries de données climatiques pour les lieux. Ces données climatiques pourront servir alors comme entrées d'un modèle de simulation des pluies journalières à l'exemple du modèle stochastique des chroniques pluviométriques développé par Gozé (1990).

Permettant de mesurer l'impact du changement climatique, cet outil est destiné en premier lieu aux sélectionneurs du Sahel qui pourront dans des programmes de sélection de courte durée (deux à cinq ans) réduire l'incertitude de la prédiction des performances des nouvelles variétés. Adapté en outre à l'échelle de la région et couplé à une méthode d'estimation des surfaces

cultivées, il peut être utilisé dans le cadre de la prévention des crises alimentaires en Afrique de lOuest, et particulièrement au Sahel. Ainsi, pour chaque année, le rapprochement entre les productions et les besoins prévisionnels peut alimenter le dispositif de veille déjà mis en place par le CILSS avec l'ensemble de ses partenaires techniques dans le cadre de l'alerte précoce, afin de faciliter la prise de décision dans l'élaboration des stratégies alimentaires au Sahel.

Références citées

[1] AASTVEIT A.H. et MARTENS H. ANOVA interactions interpreted by partial least squares regression. Biometrics, 1986, 42, 829-844.

[2] AFFHOLDER F. Empirically modelling the interaction between intensification and climatic risk in semi arid regions. Field Crops Research ; 1997, 52, 79-93.

[3] AJI S., TAVOLARO S., LANTZ F., FARAJ A. Apport du bootstrap à la régression PLS : application à la prédiction de la qualité des gazoles, Oil & Gas Science andTechnology-Rev. IFP, 2003, 58, 599-608.

[4] ANNEROSE D. et DIAGNE M. Caractérisation de la sécheresse agronomique en zone semi-aride. Présentation d'un modèle simple d'évaluation appliqué au cas de l'arachide cultivée au Sénégal. Oléagineux, 1990, 45, 12, 547-554.

[5] ANNEROSE D. et DIAGNE M. Les modèles de cultures : des outils de la recherche et du développement agricole. Arachide Infos, 1994, 5, 5-11.

[6] ANNICCHIARICO, P. Genotype x environment interaction - challenge and opportunities for plant breeding and cultivar recommendations. FAO, Rome, 150 p.

[7] BAKER R.J. Tests for crossover genotype-environmental interactions. Canadian Journal of Plant Science, 1988, 68, 405-410.

[8] BARIL, C.P. Etude de la stabilité du rendement chez le blé tendre d'hiver.- 170 p. Th : Agronomie : Paris-Sud centre d'Orsay : 1992.

[9] BARON C. Modèle de bilan hydrique et de croissance des plantes cereals : Mil, Sorgho et Arachide. CIRAD, 2002.

[10] BATTERBURY S., WARREN A. The African Sahel 25 years after the great drought : assessing progress and moving towards new agendas and approaches. Global Environmental Change, 2001, 11, 1-8.

[11] BECKER H.C. Correlations among some statistical measures of phenotypic stability. Euphytica, 1981, 30, 835-840.

[12] BECKER H.C. et LÉON J. Stability analysis in plant breeding. Plant Breeding, 1988, 101, 1-23.

[13] BELHASSEN E., THIS D. et MONNEVEUX P. L'adaptation génétique face aux contraintes de sécheresse. Cahiers agricultures, 1995, 4, 251-261.

[14] BRADU D. et GABRIEL K.R. The biplot as a diagnostic tool for models of two-way tables. Technometrics, 1978, 20, 47-68.

[15] BRANCOURT-HULMEL M., BIARNES-DUMOULIN V. et DENIS J.B. Points de repères dans l'analyse de la stabilité et de l'interaction génotype-milieu en amélioration des plantes. Agronomie, 1997, 17, 219-246.

[16] CALINSKI T., CZAJKA S. et KACZMAREK, Z. A model for the analysis of a series of experiments repeated at several places over a period of years. I. Theory. Biuletyn Oceny Odmian, 1987, 17/18, 7-33.

[17] CILLS. CSSA, Cadre stratégique de sécurité alimentaire durable dans une perspective de lutte contre la pauvreté au sahel. CILSS, Ougadougou, 2000, 88 p.

[18] COLSON J., WALLACH D., BOUNIOLS A., DENIS J., JONES J. Mean Squared Error of Yield Prediction by SOYGRO. Agronomy Journal, 1995, 87 , 397-407.

[19] CORNELIUS P.L. Functionc approximating Mandel's tables for the means and standard deviations of the first three roots of a Wishart matrix. Technometrics, 1980, 22, 613-616.

[20] CORNELIUS P.L. Statistical tests and retention of terms in the additive main effects and multiplicative interaction model for cultivar trials. Crop Science, 1993, 33, 1186-1193.

[21] CORNELIUS P.L., SEYEDSADR M.S. et CROSSA J. Using the shifted multiplicative model to search for "separability" in crop cultivar trials. Theoretical and Applied Genetics, 1992, 84, 161-172.

[22] CROSSA J. Statistical analyses of multilocation trials. Advances in agronomy, 1990, 44, 55-85.

[23] CROSSA J., CORNELIUS P.L., SAYRE K. et ORTIZ-MONASTERIO R.J. A shifted multiplicative model fusion method for grouping environments without cultivar rank change. Crop Science, 1995, 35, 54-62.

[24] CROSSA J., CORNELIUS P.L., SEYEDSADR M.S. et BYRNE, P. A shifted multiplicative model cluster analysis for grouping environments without genotypic rank range. Theoretical and Applied Genetics, 1993, 85, 577-586.

[25] de JONG S. SIMPLS : An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 1993, 18, 251-263.

[26] DEMPSTER A.P., LAIRD N.M. et RUBIN, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society B, 1977, 39, 1-38.

[27] DENIS J.B. et VINCOURT P. Panorama des méthodes statistiques d'analyse des intéractions génotype x milieu. Agronomie, 1982, 2, 3, 219-230

[28] DENIS J.B. L'analyse de régression factorielle. Biométrie Praximétrie, 1980, 10, 1-34.

[29] DENIS J.B. Two-way analysis using covariates. Statistics, 1988, 19, 123-132.

[30] DENIS J.B., PIEPHO H.P. et van EEUWIJK F.A. Modelling expectation and variance for genotype by environment data. Heredity, 1997, 79, 162-171.

[31] DIENG I. Prédiction de l'interaction génotype x environnement à partir d'indices variétaux de sensibilité à la sécheresse et de bilan hydrique à l'aide d'un modèle de régression factorielle dans des essais d'arachide au Sénégal.- 60 p. Mémoire de DEA : Biostatistique : Montpellier II : 2003.

[32] DIENG I. ; GOZE E., SABATIER R. Linéarisation autour d'un témoin pour prédire la réponse des cultures. Comptes Rendus de l'Académie des Sciences Biologies, 2006, 329, 148-155.

[33] DOORENBOS J. et JASSAM A. H. Yield response to water. In Irrigation and drainage Paper 33. FAO, Rome, 1979, pp 193.

[34] EAGLEMAN J.R. An experimentally derived model for actual evapotranspiration. Agricultural meteorology, 1971, 8, 4-3, 385-94.

[35] EBDON J.S. et GAUCH, H.G. Jr. Additive Main Effect and Multiplicative Interaction Analysis of National Turfgrass Performance Trials : I. Interpretation of Genotype x Environment Interaction. Crop Science, 2002a, 42, 489-496.

[36] EBDON J.S. et GAUCH H.G. Jr. Additive Main Effect and Multiplicative Interaction Analysis of National Turfgrass Performance Trials : II. Cultivar Recommendations. Crop Science, 2002b, 42, 497-506.

[37] EBERHART S.A. et RUSSELL W.A. Stability parameters for comparing varieties. Crop Science, 1966, 6, 36-40.

[38] EFRON B. Bootstrap Methods : Another Look at the Jacknife. Annals of statistics, 1979, 7, 1-26.

[39] FAO. IRSIS, Irrigation scheduling information system. FAO, Rome, 1987.

[40] FINLAY K.W. et WILKINSON G.N. The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research, 1963, 14, 742-754.

[41] FOREST F. et CLOPES A., 1991. Contribution à l'explication de la variabilité des rendements d'une culture de maïs plus ou moins intensifiée à l'aide d'un modèle de bila hydrique amélioré. In Bilan hydrique agricole et sécheresse en Afrique tropicale. France : John Libbey Eurotext, 1991.- pp 3-15.

[42] FOREST F. et CORTIER B. Le diagnostic hydrique des cultures et la prévision du rendement régional du mil cultivé dans les pays du CILSS. IRAT/CIRAD/AGRHYMET. IAHS, 1990, 544-557.

[43] FRANQUIN P. et FOREST F. Des programmes pour l'évaluation et l'analyse frequentielle des termes du bilan hydrique. Agronomie Tropicale, Janvier-Mars 1977, XXXII-1, 7-11.

[44] FREEMAN G. H. et PERKINS J. M. Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measures of these environments. Heredity, 1971, 27, 15-23.

[45] GAUCH H.G. et ZOBEL R.W. 1996. AMMI analysis of yield trials. In Genotype-by-environment interaction. Boca Raton, Floride : M.S. Kang & H.G. Gauch, 1996.- pp 85-122.

[46] GAUCH H.G. et ZOBEL R.W. Predictive and postdictive success of statistical analyses of yield trials. Theoretical and Applied Genetics, 1988, 76, 1-10.

[47] GAUCH H.G. Jr. Statistical Analysis of Yield Trials by AMMI and GGE. Crop Science, 2006, 46 :1488-1500.

[48] GAUCH H.G. Statistical analysis of regional yield trials : AMMI analysis of factorial designs. Amsterdam : Elsevier, 1992.

[49] GAUCH, H.G. Full and reduced models for yield trials. Theoretical and Applied Genetics, 1990, 80, 153-160.

[50] GOLLOB H.F. A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika, 1968, 33, 73-115.

[51] GOZE E. Modèle stochastique de la pluviométrie au Sahel - Application à l'agronomie- Th : Université Montpellier II : 1990.

[52] KEMPTON R.A. The use of biplots in interpreting variety by environment interactions. Journal of Agricultural Science, 1984, 103, 123-135.

[53] LAIRD N. M. Computation of variance components using the EM algorithm. Journal of Statistical Computation and Simulation, 1982, 14, 295-303.

[54] LIN C.S., BINNS M.R. et LEFKOVITCH L.P. Stability Analysis : Where Do We Stand ? Crop Science, 1986, 26, 894-900.

[55] MANDEL J. A new analysis of variance model for non-additive data. Technometrics, 1971, 13, 1-18.

[56] MANDEL J. Non-additivity in two-way analysis of variance. Journal of American Statistical Association, 1961, 56, 878-888.

[57] MCCULLOCH C. E. et SEARLE S. R. Generalized, Linear, and Mixed Models. New York : John Wiley & Sons, 2001.

[58] MCLACHLAN G.J. et KRISHNAN T. The EM Algorithm and Extensions. New Jersey : Wiley, 1997.

[59] MENG X.L. et van DYK D.A. Fast EM-type implementations for mixed effects models. Journal of the Royal Statistical Society B, 1998, 60, 559-578.

[60] MORENO-GONZALEZ J., CROSSA J. et CORNELIUS P.L. Additive Main Effects and Multiplicative Interaction Model : I. Theory on Variance Components for Predicting Cell Means. Crop Science, 2003a, 43, 1967-1975.

[61] MORENO-GONZALEZ J., CROSSA J. et CORNELIUS P.L. Additive Main Effects and Multiplicative Interaction Model : II. Theory on Shrinkage Factors for Predicting Cell Means. Crop Science, 2003b, 43 :1976-1982

[62] PERKINS J.M. et JINKS J.L. Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. Heredity, 1968, 23, 339-356.

[63] PIEPHO H.P. Best Linear Unbiased Prediction (BLUP) for regional yield trials : a comparison to additive main effects and multiplicative interaction (AMMI) analysis. Theoretical and Applied Genetics, 1989, 89, 647-654.

[64] PIEPHO H.P. Methods for Comparing the Yield Stability of cropping Systems - A Review. Journal of Agronomy and Crop Science, 1998, 180, 193-213.

[65] PIEPHO H.P., Denis J.B. et van Eeuwijk F.A. Predicting cultivar differences using covariates. Journal of Agricultural, Biological, and Environmental Statistics, 1998, 3, 151-162.

[66] PIEPHO, H.P. Robustness of statistical tests for multiplicative terms in the additive main effects and multiplicative interaction model for cultivar trials. Theoretical and Applied Genetics, 1995, 90, 438-443.

[67] RAMI J. F., DUFOUR P., TROUCHE G., FLIEDEL G., MESTRES C., DAVRIEUX F., BLANCHARD P., et HAMON P. Quantitative trait loci for grain quality, productivity, morphological and agronomical traits in Sorghum (*Sorghum bicolor* L. Moench). Theorical Applied genetics, 1998, 97, 605-616.

[68] RAO C. R. et KLEFFE J. Estimation of variance components and applications. Amsterdam : North Holland series in statistics and probability, Elsevier, 1988.

[69] SEARLE S. R. Linear models for unbalanced data. New York : John Wiley & Sons, 1987.

[70] SEARLE S. R., CASELLA G. et MCCULLOCH C. E. Variance Components. New York : John Wiley & Sons, 1992.

[71] SEIF E., EVANS J.C. et BALAAM L.N. A multivariate procedure for classifying environments according to their interaction with genotypes. Australian Journal of Agricultural Research, 1979, 30, 1021-1026.

[72] STONE, M. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society B, 1974, 36, 111-147.

[73] TALBOT M. et WHEELWRIGHT A.V. The analysis of genotype x environment interactions by partial least squares regression. Biuletyn Oceny Odmian, 1989, 21/22, 19-25.

[74] TALBOT M. Resource allocation for selection systems. In : Statistical methods for plant variety evaluation. London : KEMPTON RA, FOX PN, Chapman et Hall, 1997.- pp 162-174.

[75] TENENHAUS M. La Régression PLS : théorie et pratique. Parsi : Technip, 1998.- 254 p.

[76] THEOBALD C.M., TALBOT M. et NABUGOOMU F. A Bayesian approach to regional and local-area prediction from crop variety trials. Journal of Agricultural, Biological, and Environmental Statistics, 2002, 7, 12-28.

[77] TUCKER C.J., DREGNE H.E., NEWCOMB W.W. Expansion and contraction of the Sahara Desert from 1980 to 1990. Science, 1991, 253, 299-301.

[78] TUKEY J.W. One degree of freedom for non additivity. Biometrics, 1949, 5, 232-242.

[79] TURNER N. C., WRIGHT G. C. et SIDDIQUE K. H. M. Adaptation of grain legumes (pulses) to water-limited environments. Advances in agronomy, 2002, 71, 193-231.

[80] van EEUWIJK F.A. Interpreting genotype-by-environment interaction using redundancy analysis. Theoretical and Applied Genetics, 1992, 85, 89-100.

[81] van EEUWIJK F.A. Linear and bilinear models for the analysis of multi-environment trials. I. An inventory of models. Euphytica, 1995, 84, 1-7.

[82] van EEUWIJK F.A., DENIS J.B. et KANG M.S. 1996. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In Genotype-by-environment interaction. Boca Raton, Floride : M.S. Kang & H.G. Gauch, 1996.- pp 15-49.

[83] VARGAS M., CROSSA J., SAYRE K., REYNOLDS M., RAMÌREZ M.E. et TALBOT M. Interpreting genotype x environment interaction in wheat by partial least squares regression. Crop Science, 1998, 38, 679-689.

[84] VARGAS M., CROSSA J., van EEUWIJK F.A., RAMÌREZ M.E. et SAYRE K. Using partial least squares regression, factorial regression, and AMMI models for interpreting genotype x environment interaction. Crop Science, 1999, 39, 955-967.

[85] WALLACH D., GOFFINET B. Mean Squared Error of Prediction in Models for Studying Ecological and Agronomic Systems. Biometrics, 1987, 43, 561-573.

[86] WILLIAMS E.J. The interpretation of interactions in factorial experiments. Biometrika, 1952, 39, 65-81.

[87] WOLD H. Estimation of principal components and related models by iterative least squares. In Multivariate Analysis. New York : Krishnaiah P.R., Academic Press, 1966, pp 391-420.

[88] WOLD H. Modelling in Complex Situations with Soft Information. Third World Congress of Econometric Society. Toronto, Canada, 1975 August 21-26.

[89] WOLD S., ALBANO C., DUNN W.J., ESBENSEN K., HELLBERG S., JOHANSSON E., SJOSTROM M. Pattern recognition : Finding and

using patterns in Multivariate Data. In : Food Research and Data Analysis. London : Applied Science : Martens H, Russwurm H Jr, 1983 ;147-188.

[90] WOOD J. T. The use of environmental variables in the interpretation of genotype-environment interaction. Heredity, 1976, 37, 1-7.

[91] WU C. F. On the convergence properties of the EM algorithm. The Annals of Statistics, 1983, 11, 95-103.

[92] YAN W. et HUNT W.A. Interpretation of Genotype x Environment Interaction for Winter Wheat Yield in Ontario. Crop Science, 2001, 41, 19-25.

[93] YATES F. et COCHRAN W.G. The analysis of groups of experiments. Journal of Agricultural Science, 1938, 28, 556-580.

[94] ZOBEL R.W., WRIGHT M.J. et GAUCH H.G. Statistical analysis of a yield trial. Agronomy Journal, 1988, 80, 388-393.

Annexe A : Modèle de Régression factorielle

On dispose de trois matrices : $\mathbf{X}$ de dimension $I \times p$ est la matrice des covariables associées au premier facteur (facteur variété par exemple), $\mathbf{Z}$ de dimension $J \times q$ est la matrice des covariables associées au deuxième facteur (facteur lieu) et $\mathbf{Y}$ de dimension $I \times J$ est la matrice des observations. Les matrices $\mathbf{X}$, $\mathbf{Y}$ et $\mathbf{Z}$ s'écrivent ainsi :

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{I1} & x_{I2} & \cdots & x_{Ip} \end{pmatrix} \qquad \mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1q} \\ z_{21} & z_{22} & \cdots & z_{2q} \\ \vdots & \vdots & & \vdots \\ z_{J1} & z_{J2} & \cdots & z_{Jq} \end{pmatrix}$$

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1J} \\ y_{21} & y_{22} & \cdots & y_{2J} \\ \vdots & \vdots & & \vdots \\ y_{I1} & y_{I2} & \cdots & y_{IJ} \end{pmatrix}$$

L'idée est d'utiliser les deux matrices de covariables $\mathbf{X}$ et $\mathbf{Z}$ attachées aux deux facteurs variété et environnement pour expliquer la matrice $\mathbf{Y}$.

Pour cela, une première régression de $\mathbf{Y}$ sur $\mathbf{X}$ est faite :

$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} \;\Rightarrow\; \widehat{\mathbf{Y}}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

et $\quad \mathbf{R_X} = \mathbf{Y} - \hat{\mathbf{Y}}_\mathbf{X} = (\mathbf{I}_I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$

avec $\mathbf{I}_I$ représentant la matrice identité d'ordre I et $\mathbf{R_X}$ la matrice des résidus de la régression de $\mathbf{Y}$ sur $\mathbf{X}$ qui est de dimension $I \times J$.

Une deuxième régression de $\hat{\mathbf{Y}}'_\mathbf{X}$ sur $\mathbf{Z}$ est faite.

$\hat{\mathbf{Y}}'_\mathbf{X} = \mathbf{Z}\boldsymbol{\beta} \;\Rightarrow\; \hat{\mathbf{Y}}'_{\mathbf{XZ}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{Y}}'_\mathbf{X}$

Ainsi, $\hat{\mathbf{Y}}_{\mathbf{XZ}} = \hat{\mathbf{Y}}_\mathbf{X}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$

D'où $\mathbf{R}_{\hat{\mathbf{Y}}_\mathbf{X}} = \hat{\mathbf{Y}}_\mathbf{X} - \hat{\mathbf{Y}}_{\mathbf{XZ}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}$

$\Rightarrow \quad \mathbf{R}_{\hat{\mathbf{Y}}_\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{I}_J - \mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z})$

Enfin, une dernière régression de $\mathbf{R_X}$ sur $\mathbf{Z}$ est faite.

$\mathbf{R}'_\mathbf{X} = \mathbf{Z}\boldsymbol{\gamma} \;\Rightarrow\; \hat{\mathbf{R}}'_{\mathbf{XZ}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}'_\mathbf{X}$

$\hat{\mathbf{R}}_{\mathbf{XZ}} = \mathbf{R_X}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ d'où $\hat{\mathbf{R}}_{\mathbf{XZ}} = (\mathbf{I}_J - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$

Ainsi, $\mathbf{Y} = \hat{\mathbf{Y}}_{\mathbf{XZ}} + \mathbf{R}_{\hat{\mathbf{Y}}_\mathbf{X}} + \hat{\mathbf{R}}_{\mathbf{XZ}} + \mathbf{R}\textit{ésidus}$,

Posons

$$\hat{\mathbf{Y}}_{\mathbf{XZ}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{X}\mathbf{M}\mathbf{Z}' \qquad \text{(B.1)}$$

$$\mathbf{R}_{\hat{\mathbf{Y}}_\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(I_J - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') = \mathbf{X}\boldsymbol{\beta}' \qquad \text{(B.2)}$$

$$\hat{\mathbf{Y}}_{\mathbf{XZ}} = (I_J - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \boldsymbol{\alpha}\mathbf{Z}' \qquad \text{(B.3)}$$

$\mathbf{X}$ et $\mathbf{Z}$ peuvent être mis sous la forme $\mathbf{X} = (\ \mathbf{1}_I \quad \dot{\mathbf{X}}\ ) \quad \mathbf{Z} = (\ \mathbf{1}_J \quad \dot{\mathbf{Z}}\ )$

$\Rightarrow \quad \mathbf{X}' = \begin{pmatrix} \mathbf{1}'_I \\ \dot{\mathbf{X}}' \end{pmatrix}$ *et* $\mathbf{Z}' = \begin{pmatrix} \mathbf{1}'_I \\ \dot{\mathbf{Z}}' \end{pmatrix}$ avec $\dot{\mathbf{X}}$ se déduisant de la matrice $\mathbf{X}$ par centrage des colonnes (de même pour $\dot{\mathbf{Z}}$)

$\dot{\mathbf{X}}'\mathbf{X} = \begin{pmatrix} \mathbf{1}'_I \\ \dot{\mathbf{X}}' \end{pmatrix} (\ \mathbf{1}_I \quad \dot{\mathbf{X}}\ ) = \begin{pmatrix} I & \mathbf{1}'_I\dot{\mathbf{X}} \\ \dot{\mathbf{X}}'\mathbf{1}_I & \dot{\mathbf{X}}'\dot{\mathbf{X}} \end{pmatrix}$

Or $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}' & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{pmatrix}$

avec $\mathbf{E} = \mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}$ et $\mathbf{F} = \mathbf{A}^{-1}\mathbf{B}$

Ici $\mathbf{A} = I$, $\mathbf{B} = \mathbf{1}_I'\dot{\mathbf{X}} \Rightarrow \mathbf{B}' = \dot{\mathbf{X}}'1_I$ et $\mathbf{D} = \dot{\mathbf{X}}'\dot{\mathbf{X}}$

D'où $\mathbf{A}^{-1} = \frac{1}{I}$, $F = \frac{1}{I}\mathbf{1}_I'\dot{\mathbf{X}} = 0$ car $\dot{\mathbf{X}}$ centré

et $\mathbf{E} = \dot{\mathbf{X}}'\dot{\mathbf{X}} \Rightarrow \mathbf{E}^{-1} = \left(\dot{\mathbf{X}}'\dot{\mathbf{X}}\right)^{-1}$.

Ainsi $(\dot{\mathbf{X}}'\mathbf{X})^{-1} = \begin{pmatrix} I & \mathbf{1}_I'\dot{\mathbf{X}} \\ \dot{\mathbf{X}}'1_I & \dot{\mathbf{X}}'\dot{\mathbf{X}} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{I} & 0 \\ 0 & (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1} \end{pmatrix}$

De même $(\dot{\mathbf{Z}}'\mathbf{Z})^{-1} = \begin{pmatrix} \frac{1}{J} & 0 \\ 0 & (\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix}$

D'après l'équation (B.1) $\mathbf{M} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$

$$\mathbf{M} = \begin{pmatrix} \frac{1}{I} & 0 \\ 0 & (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{1}_I' \\ \dot{\mathbf{X}}' \end{pmatrix} \mathbf{Y} \begin{pmatrix} \mathbf{1}_J & \dot{\mathbf{Z}} \end{pmatrix} \begin{pmatrix} \frac{1}{J} & 0 \\ 0 & (\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} \frac{1}{I}\mathbf{1}_I'\mathbf{Y} \\ (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y} \end{pmatrix} \begin{pmatrix} \mathbf{1}_J & \dot{\mathbf{Z}} \end{pmatrix} \begin{pmatrix} \frac{1}{J} & 0 \\ 0 & (\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} \frac{1}{I}\mathbf{1}_I'\mathbf{Y}\mathbf{1}_J & \frac{1}{I}\mathbf{1}_I'\mathbf{Y}\dot{\mathbf{Z}} \\ (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y}1_J & (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y}\dot{\mathbf{Z}} \end{pmatrix} \begin{pmatrix} \frac{1}{J} & 0 \\ 0 & (\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix}$$

D'où

$$\boxed{M = \begin{pmatrix} \frac{1}{IJ}\mathbf{1}_I'\mathbf{Y}\mathbf{1}_J & \frac{1}{I}\mathbf{1}_I'\mathbf{Y}\dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \\ \frac{1}{J}(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y}\mathbf{1}_J & (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y}\dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix} = \begin{pmatrix} \mathbf{m} & \mathbf{m}_2' \\ \mathbf{m}_1 & \dot{\mathbf{M}} \end{pmatrix}}$$

D'après l'équation (B.2) $\boldsymbol{\beta}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{I}_J - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')$

$$(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y} = \begin{pmatrix} \frac{1}{I}\mathbf{1}_I'\mathbf{Y} \\ (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y} \end{pmatrix}$$

$$\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{pmatrix} \mathbf{1}_J & \dot{\mathbf{Z}} \end{pmatrix} \begin{pmatrix} \frac{1}{J} & 0 \\ 0 & (\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix} = \begin{pmatrix} \frac{1}{J}\mathbf{1}_J & \dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix}$$

Donc $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \begin{pmatrix} \frac{1}{J}\mathbf{1}_J & \dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{1}'_J \\ \dot{\mathbf{Z}}' \end{pmatrix} = \frac{1}{J}\mathbf{1}_J\mathbf{1}'_J + \dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1}\dot{\mathbf{Z}}'$

D'où $\boldsymbol{\beta}' = \begin{pmatrix} \frac{1}{I}\mathbf{1}'_I\mathbf{Y} \\ (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y} \end{pmatrix} (\mathbf{I}_J - \frac{1}{J}\mathbf{1}_J\mathbf{1}'_J - \dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1}\dot{\mathbf{Z}}')$

$$\beta' = \begin{pmatrix} \frac{1}{I}\mathbf{1}'_I\mathbf{Y}(\mathbf{I}_J - \frac{1}{J}\mathbf{1}_J\mathbf{1}'_J - \dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1}\dot{\mathbf{Z}}') \\ (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'\mathbf{Y}(\mathbf{I}_J - \frac{1}{J}\mathbf{1}_J\mathbf{1}'_J - \dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1}\dot{\mathbf{Z}}') \end{pmatrix} = \begin{pmatrix} \mathbf{b}' \\ \mathbf{B}' \end{pmatrix}$$

D'après l'équation (B.3) $\alpha = (\mathbf{I}_I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}$

$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \begin{pmatrix} \mathbf{1}_I & \dot{\mathbf{X}} \end{pmatrix} \begin{pmatrix} \frac{1}{I}\mathbf{1}'_I \\ (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}' \end{pmatrix} = \frac{1}{I}\mathbf{1}_I\mathbf{1}'_I + \dot{\mathbf{X}}(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}'$

D'où $\boldsymbol{\alpha} = (\mathbf{I}_I - \frac{1}{I}\mathbf{1}_I\mathbf{1}'_I - \dot{\mathbf{X}}(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}')\begin{pmatrix} \frac{1}{J}\mathbf{Y}\mathbf{1}_J & \mathbf{Y}\dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix}$

$$\boldsymbol{\alpha} = \begin{pmatrix} (\mathbf{I}_I - \frac{1}{I}\mathbf{1}_I\mathbf{1}'_I - \dot{\mathbf{X}}(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}')\frac{1}{J}\mathbf{Y}\mathbf{1}_J & (\mathbf{I}_I - \frac{1}{I}\mathbf{1}_I\mathbf{1}'_I - \dot{\mathbf{X}}(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}')\mathbf{Y}\dot{\mathbf{Z}}(\dot{\mathbf{Z}}'\dot{\mathbf{Z}})^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{a} & \mathbf{A} \end{pmatrix}$$

Ainsi $\mathbf{Y} = \hat{\mathbf{Y}}_{\mathbf{XZ}} + \mathbf{R}_{\hat{\mathbf{Y}}_{\mathbf{X}}} + \hat{\mathbf{R}}_{\mathbf{XZ}} + \mathbf{R}\acute{e}sidus = \mathbf{X}\mathbf{M}\mathbf{Z}' + \mathbf{X}\boldsymbol{\beta}' + \boldsymbol{\alpha}\mathbf{Z}' + \mathbf{R}\acute{e}sidus$

D'où $\hat{\mathbf{Y}} = \begin{pmatrix} \mathbf{1}_I & \dot{\mathbf{X}} \end{pmatrix} \begin{pmatrix} \mathbf{m} & \mathbf{m}'_2 \\ \mathbf{m}_1 & \dot{\mathbf{M}} \end{pmatrix} \begin{pmatrix} \mathbf{1}'_J \\ \dot{\mathbf{Z}} \end{pmatrix} + \begin{pmatrix} \mathbf{1}_I & \dot{\mathbf{X}} \end{pmatrix} \begin{pmatrix} \mathbf{b}' \\ \mathbf{B}' \end{pmatrix} + \begin{pmatrix} \mathbf{a} & \mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{1}'_J \\ \dot{\mathbf{Z}} \end{pmatrix}$

$\hat{\mathbf{Y}} = \begin{pmatrix} \mathbf{1}_I\mathbf{m} + \dot{\mathbf{X}}\mathbf{m}_1 & \mathbf{1}_I\mathbf{m}'_2 + \dot{\mathbf{X}}\dot{\mathbf{M}} \end{pmatrix} \begin{pmatrix} \mathbf{1}'_J \\ \dot{\mathbf{Z}}' \end{pmatrix} + \mathbf{1}_I\mathbf{b}' + \dot{\mathbf{X}}\mathbf{B}' + \mathbf{a}\mathbf{1}'_J + \mathbf{A}\dot{\mathbf{Z}}'$

D'où

$$\hat{\mathbf{Y}} = \mathbf{1}_I\mathbf{m}\mathbf{1}'_J + \dot{\mathbf{X}}\mathbf{m}_1\mathbf{1}'_J + \mathbf{1}_I\mathbf{m}'_2\dot{\mathbf{Z}}' + \dot{\mathbf{X}}\dot{\mathbf{M}}\dot{\mathbf{Z}}' + \mathbf{1}_I\mathbf{b}' + \dot{\mathbf{X}}\mathbf{B}' + \mathbf{a}\mathbf{1}'_J + \mathbf{A}\dot{\mathbf{Z}}'$$

Annexe B : Article au C.R. Biologie

Available online at www.sciencedirect.com
SCIENCE DIRECT
C. R. Biologies 329 (2006) 148–155
COMPTES RENDUS BIOLOGIES
http://france.elsevier.com/direct/CRASS3/

Biomodélisation / Biological modelling

Linéarisation autour d'un témoin pour prédire la réponse de cultures

Ibnou Dieng [a,*], Éric Gozé [b], Robert Sabatier [c]

[a] *Centre d'étude régional pour l'amélioration de l'adaptation à la sécheresse, BP 3320, Thiès-Escale, Thiès, Sénégal*
[b] *Centre de coopération internationale en recherche agronomique pour le développement, TA 70/09, avenue d'Agropolis, 34398 Montpellier cedex 5, France*
[c] *Laboratoire de physique moléculaire et structurale, faculté de pharmacie, 15, avenue Charles-Flahault, 34060 Montpellier, France*

Reçu le 18 avril 2005 ; accepté le 17 janvier 2006

Disponible sur Internet le 9 février 2006

Présenté par Michel Thellier

Résumé

Une nouvelle méthode pour modéliser les interactions génotype × environnement : APLAT. Le rendement de génotypes prédit par un modèle de simulation de cultures est développé en série de Taylor à l'ordre 1 au voisinage du vecteur de paramètres d'un génotype de référence. À l'aide de cette linéarisation locale, l'estimation des paramètres de ces génotypes se fait par régression linéaire des rendements observés sur la sensibilité des sorties du modèle de simulation de cultures par rapport aux paramètres. ***Pour citer cet article : I. Dieng et al., C. R. Biologies 329 (2006).***

Abstract

Prediction of crop response by linearisation about control approximation. A new method for modelling genotype × environment interaction: APLAT. The yield predicted by a crop-simulation model is developed as a Taylor series in the neighbourhood of a parameter vector of a control genotype. With this local linearisation, these genotype parameters can be estimated by a linear regression of the observed yield on the derivatives of the crop-simulation model predictions with respect to its parameters. ***To cite this article: I. Dieng et al., C. R. Biologies 329 (2006).***

Mots-clés : Linéarisation ; Prédiction de la réponse de cultures ; Témoin ; Interaction genotype × environnement

Keywords: Linearization; Predict responses culture; Control; Genotype × environment interaction

Abridged English version

In Sahel, genotype × environment interactions are often large: this is the justification behind multilocation and pluriannual trials. Because of these sizeable environment effects and interactions, the prediction of an expected yield with a linear mixed model is generally imprecise.

Improving this prediction can be achieved by modelling the environment effect. It is then partly shifted from the random part to the fixed part of a mixed model, by the use of a crop-simulation model like DHC, IRSIS, SarraH... This could not be possible with the empirical

* Auteur correspondant.
Adresses e-mail : ibnou.dieng@ceraas.org, dieng_ibnou@yahoo.fr (I. Dieng), eric.goze@cirad.fr (É. Gozé), sabatier@univ-montp1.fr (R. Sabatier).

1631-0691/$ – see front matter
doi:10.1016/j.crvi.2006.01.005

genotype × environment interactions analysis methods like AMMI and joint regression, which do not make use of environmental variables. The factorial regression method does make use of environmental variables; however, it requires their effect on the production to be linear, which might not be the case.

Unfortunately, most crop-simulation models bear a number of parameters, the estimation of which requires a specific and costly experiment. As a consequence, these parameters are usually known, but for a small set of reference genotypes. It would not be sensible to invest in a parameter estimation experiment for every new genotype that is proposed for selection.

To overcome this problem, one can notice that multisite experiments usually share a control variety for which parameters have already been estimated. In this paper, we propose to develop as a Taylor series the modelled response about the parameters of this control genotype. The other genotypes' parameters can then be estimated by a linear regression of the observed yields on the sensitivity to parameters, i.e., on the derivatives of the response with respect to the parameters. With this estimation, one can predict the new genotype responses in environments where they have not been tested. In a given location, this estimation can benefit from the available historic climatic records to estimate a distribution of probable yields.

Let $f(\mathbf{Z}_j, \boldsymbol{\theta}_i)$ denote the yield of a genotype i predicted by a crop simulation in an environment j and Y_{ij} the observed yield. We can write:

$$Y_{ij} = f(\mathbf{Z}_j, \boldsymbol{\theta}_i) + \xi_j + u_{ij}$$

where $\mathbf{Z}_j$ is the vector of the jth environment regressors and $\boldsymbol{\theta}_i$ the P-vector of the ith genotype parameters. The bias ξ_j is that of the crop-simulation model. We suppose that it depends only on environment and is the same for all genotypes in a same environment. The error term u_{ij} is supposed random with expectation 0 and variance σ_u^2.

Let us consider a control genotype, i.e., whose parameters are known or at least already estimated. Let $\boldsymbol{\theta}_0$ be the vector of parameters of this control genotype and let us suppose that f is a C^1 class function in a neighbourhood of $\boldsymbol{\theta}_0$ and f' derivable in this neighbourhood. Moreover, let us suppose $\boldsymbol{\theta}_i$ in the neighbourhood of $\boldsymbol{\theta}_0$. Then, a Taylor series expansion yields:

$$f(\mathbf{Z}_j, \boldsymbol{\theta}_i) = f(\mathbf{Z}_j, \boldsymbol{\theta}_0) + \sum_{p=1}^{P} \left[\frac{\partial f}{\partial \theta^{(p)}} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \mathbf{Z}=\mathbf{Z}_j} \right] \times \left(\theta_i^{(p)} - \theta_0^{(p)}\right) + \circ\left[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)\right]$$

with $\theta_i^{(p)}$ the pth component of the parameters vector of the ith genotype, $\theta_0^{(p)}$ that of the control genotype.

Let $X_j^{(p)} = [\frac{\partial f}{\partial \theta^{(p)}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \mathbf{Z}=\mathbf{Z}_j}]$ and $\beta_i^{(p)} = \theta_i^{(p)} - \theta_0^{(p)}$ for $p = 1, \ldots, P$. As f is not known in closed form, one has to estimate its derivatives by numerical approximation. The function $X_j^{(p)}$ is a function of environment j, while $\beta_i^{(p)}$ is a function of genotype i. Then, the local linearization yields:

$$Y_{ij} - Y_{0j} = \sum_{p=1}^{P} X_j^{(p)} \cdot \beta_i^{(p)} + \epsilon_{ij}$$

where Y_{0j} is the control response in the environment j and $\epsilon_{ij} = u_{ij} - u_{0j}$. So, $\mathbb{E}(\epsilon_{ij}) = 0$, $\mathbb{V}(\epsilon_{ij}) = 2\sigma_u^2$, $\mathbb{C}\mathrm{ov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$, but $\mathbb{C}\mathrm{ov}(\epsilon_{ij}, \epsilon_{i'j}) = \sigma_u^2$.

This equation can be put in the form of a linear model with correlated errors:

$$\mathbf{Y} - (\mathbf{Y}_0 \otimes \mathbf{1}_I) = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

In this equation, $\mathbf{Y}$ is the vector of responses of the I genotypes in the J environments, $\mathbf{Y}_0' = (Y_{01} \cdots Y_{0J})$, $\mathbf{1}_I$ is unit vector of size $I \times 1$. The symbol $\otimes$ indicates the Kronecker product and $\boldsymbol{\epsilon}$ is a random error vector. Its covariance matrix is $\sigma_u^2 \boldsymbol{\Omega}$ where:

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\omega}_1 & & & 0 \\ & \ddots & & \\ & & \boldsymbol{\omega}_j & \\ & & & \ddots \\ 0 & & & \boldsymbol{\omega}_J \end{pmatrix} \quad \text{and}$$

$$\boldsymbol{\omega}_j = \begin{pmatrix} 2 & & 1 \\ & \ddots & \\ 1 & & 2 \end{pmatrix}$$

The number of columns of the square matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\omega}_j$ are respectively the number of observations for all the environments and the number of observations for environment j.

Also, $\mathbf{X} = [\mathbf{X}^{(1)} \otimes \mathbf{I}_I \cdots \mathbf{X}^{(P)} \otimes \mathbf{I}_I]$ where $\mathbf{X}^{(p)\prime} = [X_1^{(p)} \cdots X_J^{(p)}]$ is a $J \times 1$ vector and $\mathbf{I}_I$ is the $I \times I$ unit matrix. The dimension of $\mathbf{X}$ is then $IJ \times PI$.

Finally, $\boldsymbol{\beta}' = [\boldsymbol{\beta}^{(1)\prime} \cdots \boldsymbol{\beta}^{(P)\prime}]$ where $\boldsymbol{\beta}^{(p)\prime} = [\beta_1^{(p)} \cdots \beta_I^{(p)}]$.

We call this method APLAT for *Approximation Par Linéarisation Autour d'un Témoin*.

Because of the large number of columns of $\mathbf{X}$, some dimension reduction method like Partial Least Squares regression is necessary. The dimension of the space spanned by the regressors is then reduced from rank of $\mathbf{X}$ to k. The PLS regression is usually carried out

with the NIPALS (Nonlinear estimation by Iterative Partial Least Squares) algorithm, where the calculation of the components is performed simultaneously with a set of regressions by ordinary least squares. Here, the error covariance matrix is $\sigma_u^2\boldsymbol{\Omega}$, not $\sigma_u^2\mathbf{I}_{IJ}$, generalized least squares should be used instead. As $\boldsymbol{\Omega}$ is symmetric and positive semi-definite, a work around consists in factorizing its inverse, finding a matrix $\boldsymbol{\eta}$ such that $\boldsymbol{\eta}'\boldsymbol{\eta} = \boldsymbol{\Omega}^{-1}$.

Then, estimating $\boldsymbol{\beta}$ by PLS with regressions by generalized least squares is equivalent to consider the model:

$$\boldsymbol{\eta}\mathbf{Y} - \boldsymbol{\eta}(\mathbf{Y}_0 \otimes \mathbf{1}_I) = \boldsymbol{\eta}\mathbf{X}\cdot\boldsymbol{\beta} + \boldsymbol{\eta}\boldsymbol{\epsilon}$$

where $\tilde{\boldsymbol{\beta}}_{\mathrm{PLS}}$ is the estimation with regressions made by ordinary least squares.

The number of components is chosen to minimize the PRESS (Prediction Error Sum of Squares) criterion.

To calculate the confidence interval of the coefficients, we used a bootstrap technique. Let $z_{i,\mathrm{PLS}}^{(p)\star b}$ be the random variable defined by:

$$z_{i,\mathrm{PLS}}^{(p)\star b} = \frac{\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b} - \tilde{\beta}_{i,\mathrm{PLS}}^{(p)}}{\tilde{s}^{\star}(\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b})}$$

where $\tilde{\beta}_{i,\mathrm{PLS}}^{(p)}$ is the $(p\cdot i)$th element of $\tilde{\boldsymbol{\beta}}_{\mathrm{PLS}}$, $\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b}$ is obtained at the bth draw with $b = 1,\ldots,\mathrm{B}$ and $\tilde{s}^{\star}(\hat{\beta}_{i,\mathrm{PLS}}^{(p)\star b})$ is the standard error of $\tilde{\boldsymbol{\beta}}_{\mathrm{PLS}}^{\star b}$. Let $\widehat{\mathrm{F}}_{\mathrm{B}}$ be the empirical distribution function of $z_{i,\mathrm{PLS}}^{(p)\star b}$. The fractile $\widehat{\mathrm{F}}_{\mathrm{B}}^{-1}(\alpha)$ is estimated by $\hat{t}(\alpha)$ such that $\#\{z_{i,\mathrm{PLS}}^{(p)\star b} \leqslant \hat{t}(\alpha)\} = \alpha\mathrm{B}$.

A percentile-t confidence interval for the $(p\cdot i)$th element of $\boldsymbol{\beta}$ is in the following form:

$$\big[\tilde{\beta}_{i,\mathrm{PLS}}^{(p)} - \tilde{s}\big(\tilde{\beta}_{i,\mathrm{PLS}}^{(p)}\big)\cdot\hat{t}(1-\alpha),\ \tilde{\beta}_{i,\mathrm{PLS}}^{(p)} - \tilde{s}\big(\tilde{\beta}_{i,\mathrm{PLS}}^{(p)}\big)\cdot\hat{t}(\alpha)\big]$$

To evaluate the quality of the new model, we compared its MSEP (Mean Squared Error of Prediction) with that of the average model defined for our data as follows:

$$Y_{ij} = m + g_i + E_j + \delta_{ij}$$

where m is the population mean and g_i the genotype effect. The term E_j is the year effect and it is assumed random with expectation 0 and variance σ_E^2. Errors δ_{ij} are distributed independently with expectation 0 and variance σ_δ^2. The terms E_j and δ_{ij} are assumed to be mutually independent.

The data set consists of plant yields of 26 groundnut genotypes. The experiments have been carried out at Bambey (14°42N and 16°28W) in Senegal, over a period of five years from 1994 to 1998. The data of each year were kept in turn as a test sample. Yields are expressed in kilograms of pods per hectare.

We used SarraH, a crop simulation model developed by CIRAD in collaboration with CERAAS, to calculate $\mathbf{X}$. Taking into account the available number of data, we estimated two of its varietal parameters.

The PRESS is minimal with six components for models adjusted without the data of 1994, 1995 and 1997. For each of the others, the PRESS is minimal with nine components. However, we decided to keep only five components, as the PRESS was not very different from its minimum value.

The APLAT MSEPs are lower than the average model MSEP, except for prediction of 1998 data. Then the prediction of yield for these models by APLAT was better than that made with the average model four times out of five.

With the APLAT method, the prediction of a genotype in a new environment comes at a relatively low price, using mostly available data, except for the environmental data, which has to be recorded for every site of the experiment, according to the crop-simulation model needs. This method seems promising, but requires additional studies with more numerous data.

1. Introduction

Au Sahel, les interactions genotype × environnement constatées lors des essais multilocaux et pluriannuels sont généralement importantes. Sur les réponses moyennes par variété et par environnement, le modèle linéaire généralement adopté s'écrit :

$$Y_{ij} = m + g_i + E_j + (gE)_{ij} + e_{ij} \tag{1}$$

où Y_{ij} est la réponse du génotype i de l'environnement j, m la moyenne générale et g_i l'effet fixe du génotype i. L'effet E_j de l'environnement j et l'interaction $(gE)_{ij}$ peuvent être fixes ou aléatoires. Pour l'objectif de prédiction des réponses de génotypes dans l'ensemble des environnements potentiels auxquels ils sont destinés, l'optique aléatoire est plus pertinente. Ainsi, supposons ces deux effets et le terme d'erreur e_{ij} aléatoires, iid et indépendants les uns des autres avec $\mathbb{E}(E_j) = \mathbb{E}[(gE)_{ij}] = \mathbb{E}(e_{ij}) = 0$ et $\mathbb{V}(E_j) = \sigma_E^2$, $\mathbb{V}[(gE)_{ij}] = \sigma_{gE}^2$ et $\mathbb{V}(e_{ij}) = \sigma_e^2$ où $\mathbb{E}(\cdot)$ et $\mathbb{V}(\cdot)$ désignent l'espérance et la variance.

Choisir un génotype i dans un environnement j suppose d'estimer l'espérance de sa performance dans j. La précision de cette estimation est fonction de σ_E^2, σ_{gE}^2 et de σ_e^2. Dans cette zone du Sahel, l'environnement est variable, c'est-à-dire que σ_E^2 et σ_{gE}^2 sont grands, ce qui

dégrade cette précision. Pour l'améliorer, une solution est de modéliser les variations de Y_{ij} en fonction de l'environnement par l'utilisation de modèles de simulation de cultures tels que DHC [1], IRSIS [2], SarraH [3], etc. De ce fait, une partie de l'effet aléatoire de l'environnement est reportée dans la partie fixe du modèle. Cette approche n'est pas possible avec les modèles classiques de l'interaction génotype × environnement. En effet, la méthode AMMI, *Additive Main effects and Multiplicative Interactions* [4] ainsi que la régression conjointe [5,6] ne tiennent pas compte des nouveaux environnements pour y prédire les réponses des génotypes. La régression factorielle [4,5] en tient compte, mais suppose que l'action des variables des environnements sur la production est linéaire, ce qui n'est pas certain.

Cependant, les paramètres des modèles de simulation de cultures ne sont pour la plupart connus que pour un petit nombre de génotypes, car leur évaluation demande une expérimentation spécifique et des mesures coûteuses.

L'objectif de cette étude se pose alors en ces termes : comment prédire le comportement de génotypes dans de nouveaux environnements en tenant compte de ces derniers, sans coût excessif ?

2. Le modèle proposé

Si nous partons du modèle de simulation de cultures, chacune des sorties de ce modèle, le rendement potentiel par exemple, peut s'interpréter comme la réponse d'un génotype i dans un environnement j :

$$Y_{ij} = f(\mathbf{Z}_j, \boldsymbol{\theta}_i) + \xi_j + u_{ij} \tag{2}$$

où $\mathbf{Z}_j$ est le vecteur des variables telles que la pluie, la température, etc., mesurées sur l'environnement j et $\boldsymbol{\theta}_i$ le vecteur de longueur P des paramètres du génotype i. L'erreur ξ_j est le biais du modèle de simulation de cultures ; nous supposons qu'elle ne dépend que de l'environnement j : elle est donc la même pour tous les génotypes d'un même environnement. Le terme u_{ij} est pris aléatoire, avec $\mathbb{E}(u_{ij}) = 0$ et $\mathbb{V}(u_{ij}) = \sigma_u^2$.

Comme on l'a dit précédemment, les paramètres des modèles de simulation de cultures ne sont généralement connus que pour un petit nombre de génotypes. Considérons un modèle de simulation de cultures et un génotype de référence dont les paramètres sont connus et appelons $\boldsymbol{\theta}_0$ le vecteur de ses paramètres. Alors, supposons f de classe C^1 dans un voisinage de $\boldsymbol{\theta}_0$ et f' dérivable sur ce voisinage. De plus supposons $\boldsymbol{\theta}_i$ au voisinage de $\boldsymbol{\theta}_0$. En pratique, les génotypes dont nous chercherons à estimer leurs paramètres seront choisis de telle sorte qu'ils ne soient pas trop éloignés du génotype de référence. Alors, un développement en série de Taylor à l'ordre 1 nous donne :

$$f(\mathbf{Z}_j, \boldsymbol{\theta}_i) = f(\mathbf{Z}_j, \boldsymbol{\theta}_0) + \sum_{p=1}^{P} \left[\frac{\partial f}{\partial \theta^{(p)}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \mathbf{Z}=\mathbf{Z}_j} \times (\theta_i^{(p)} - \theta_0^{(p)}) + \circ\left[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)\right] \tag{3}$$

avec $\theta_i^{(p)}$ et $\theta_0^{(p)}$ la p^{e} composante du vecteur de paramètres respectivement du génotype i et du génotype de référence.

Posons $X_j^{(p)} = [\frac{\partial f}{\partial \theta^{(p)}}]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \mathbf{Z}=\mathbf{Z}_j}$: c'est une fonction de l'environnement j et $\beta_i^{(p)} = \theta_i^{(p)} - \theta_0^{(p)}$ une fonction du génotype i. La fonction $X_j^{(p)}$ est la dérivée partielle de la sortie du modèle de simulation de cultures pour l'environnement j par rapport à la p^{e} composante du vecteur de paramètres de la variété de référence. Comme la fonction f n'est pas généralement connue analytiquement, ces sensibilités peuvent être obtenues par une méthode de dérivation numérique. Nous avons retenu tout simplement :

$$X_j^{(p)} = \left[\frac{\partial f}{\partial \theta^{(p)}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0, \mathbf{Z}=\mathbf{Z}_j} \simeq \left[\frac{f(\theta_0^{(p)} + h_{\theta_0^{(p)}}) - f(\theta_0^{(p)} - h_{\theta_0^{(p)}})}{2h_{\theta_0^{(p)}}}\right]_{\mathbf{Z}=\mathbf{Z}_j}$$

avec $h_{\theta_0^{(p)}}$ très petit, de l'ordre de $\theta_0^{(p)} \times 10^{-4}$ en pratique. D'autres méthodes existent, celle-ci étant la plus simple et économe en calculs.

Avec ces notations et d'après l'Éq. (2), qui permet d'écrire $f(\mathbf{Z}_j, \boldsymbol{\theta}_0) = Y_{0j} - \xi_j - u_{0j}$, nous pouvons écrire, en négligeant $\circ[(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)'(\boldsymbol{\theta}_i - \boldsymbol{\theta}_0)]$:

$$Y_{ij} - Y_{0j} = \sum_{p=1}^{P} X_j^{(p)} \cdot \beta_i^{(p)} + \epsilon_{ij} \tag{4}$$

où $\epsilon_{ij} = u_{ij} - u_{0j}$. Ainsi, $\mathbb{E}(\epsilon_{ij}) = 0$, $\mathbb{V}(\epsilon_{ij}) = 2\sigma_u^2$, $\mathbb{C}\text{ov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$, mais $\mathbb{C}\text{ov}(\epsilon_{ij}, \epsilon_{i'j}) = \sigma_u^2$.

Si nous disposons de I génotypes et de J environnements, nous pouvons poser le modèle suivant :

$$\mathbf{Y} - (\mathbf{Y}_0 \otimes \mathbf{1}_I) = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5}$$

Le vecteur $\mathbf{Y}$ représente le rendement de tous les génotypes dans tous les environnements ; il est de longueur IJ, $\mathbf{Y}_0' = (Y_{01} \cdots Y_{0J})$ et $\mathbf{1}_I$ est un vecteur formé de 1, de longueur I. Le symbole $\otimes$ désigne le produit de Kronecker. Le vecteur $\boldsymbol{\epsilon}$ est un vecteur d'erreur aléa-

toire. Sa matrice de covariance est de la forme $\sigma_u^2 \boldsymbol{\Omega}$, avec :

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\omega}_1 & & & & \\ & \ddots & & & 0 \\ & & \boldsymbol{\omega}_j & & \\ & 0 & & \ddots & \\ & & & & \boldsymbol{\omega}_J \end{pmatrix} \quad \text{où}$$

$$\boldsymbol{\omega}_j = \begin{pmatrix} 2 & & 1 \\ & \ddots & \\ 1 & & 2 \end{pmatrix}$$

Les matrices $\boldsymbol{\Omega}$ et $\boldsymbol{\omega}_j$ sont carrées de nombre de lignes, respectivement le nombre d'observations de tous les environnements et le nombre d'observations de l'environnement j.

Ensuite, $\mathbf{X} = [\mathbf{X}^{(1)} \otimes \mathbf{I}_I \cdots \mathbf{X}^{(P)} \otimes \mathbf{I}_I]$ où $\mathbf{X}^{(p)\prime} = [X_1^{(p)} \cdots X_J^{(p)}]$ est de longueur J et $\mathbf{I}_I$ est la matrice identité d'ordre I. La matrice $\mathbf{X}$ est donc de dimension $IJ \times PI$.

Enfin, $\boldsymbol{\beta}' = [\boldsymbol{\beta}^{(1)\prime} \cdots \boldsymbol{\beta}^{(P)\prime}]$ avec $\boldsymbol{\beta}^{(p)\prime} = [\beta_1^{(p)} \cdots \beta_I^{(p)}]$.

Nous proposons d'appeler cette méthode par l'acronyme APLAT : Approximation Par Linéarisation Autour d'un Témoin. Elle consiste à approcher, localement, le rendement prédit par un modèle de simulation de cultures, par série de Taylor à l'ordre 1 au voisinage du vecteur de paramètres d'un génotype de référence. Cette linéarisation permet, par régression linéaire, l'estimation des paramètres de ces génotypes. Par la suite, la prédiction de l'écart entre le rendement de ces génotypes et celui du génotype de référence dans des environnements nouveaux, c'est-à-dire où ils ne sont pas encore testés, pourra se faire si le climat de ces derniers est connu.

3. Estimation des paramètres et validation du modèle

Il y a en général beaucoup de paramètres dans un modèle de simulation de cultures et peu d'environnements dans un essai multienvironnement, ce qui rend souvent PI grand par rapport à IJ. Pour notre exemple, nous avons utilisé SarraH comme modèle de simulation de cultures. Ce modèle dispose de 61 paramètres, qui sont fonction du génotype. Avec un tel nombre de prédicteurs, l'estimation de $\boldsymbol{\beta}$ s'est faite par régression PLS, *Partial Least Squares* [7]. Il s'agit donc pour nous d'écrire un modèle linéaire de prédiction des rendements des génotypes pour de nouveaux environnements par les sensibilités par rapport aux paramètres des génotypes des sorties d'un modèle de simulation de cultures, fondé sur la construction de composantes orthogonales dans l'image de $\mathbf{X}$. Ceci permet de réduire l'espace des régresseurs de rang de $\mathbf{X}$ à k dimensions. La régression PLS s'effectue selon le principe de l'algorithme NIPALS, *Nonlinear estimation by Iterative Partial Least Squares* [7], où un ensemble de régressions partielles par moindres carrés ordinaires est effectué, en même temps que le calcul des composantes. Ici, la matrice de covariance de $\boldsymbol{\epsilon}$ est égale à $\sigma_u^2 \boldsymbol{\Omega}$ et non à $\sigma_u^2 \mathbf{I}_{IJ}$. La solution serait d'effectuer toutes les régressions partielles par moindres carrés généralisés. Mais cette matrice de covariance est inconnue. Elle s'écrit tout de même, à une constante multiplicative près, en fonction de $\boldsymbol{\Omega}$, qui elle est connue. La matrice $\boldsymbol{\Omega}$ étant symétrique et semi-définie positive, par décomposition de Cholesky, il existe une matrice $\boldsymbol{\eta}$ tel que $\boldsymbol{\eta}'\boldsymbol{\eta} = \boldsymbol{\Omega}^{-1}$.

Ainsi, estimer $\boldsymbol{\beta}$ par PLS avec les régressions partielles par moindres carrés généralisés consiste à poser le modèle suivant :

$$\boldsymbol{\eta}\mathbf{Y} - \boldsymbol{\eta}(\mathbf{Y}_0 \otimes \mathbf{1}_I) = \boldsymbol{\eta}\mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\eta}\boldsymbol{\epsilon} \tag{6}$$

où $\tilde{\boldsymbol{\beta}}_{\mathrm{PLS}}$ est l'estimation avec les régressions partielles effectuées par moindres carrés ordinaires.

Dans ce cas, la variance de l'erreur $\boldsymbol{\eta}\boldsymbol{\epsilon}$ s'écrit :

$$\begin{aligned} \mathbb{E}(\boldsymbol{\eta}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{\eta}') &= \boldsymbol{\eta}\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\boldsymbol{\eta}' = \sigma_u^2 \boldsymbol{\eta}\boldsymbol{\Omega}\boldsymbol{\eta}' = \sigma_u^2 \boldsymbol{\eta}(\boldsymbol{\eta}'\boldsymbol{\eta})^{-1}\boldsymbol{\eta}' \\ &= \sigma_u^2 \boldsymbol{\eta}\boldsymbol{\eta}^{-1}(\boldsymbol{\eta}')^{-1}\boldsymbol{\eta}' = \sigma_u^2 \mathbf{I}_{IJ} \end{aligned}$$

Le nombre de composantes à retenir est déterminé par le PRESS, *Prediction Error Sum of Squares* [7].

Nous avons calculé les intervalles de confiance des coefficients estimés par la méthode *bootstrap* [8]. Cette technique permet d'estimer la loi inconnue d'un estimateur par une loi empirique obtenue à partir d'une procédure de rééchantillonnage fondée sur des tirages aléatoires avec remise des données. Les intervalles de confiance construits sont de type percentile-t [9]. Soit $z_{i,\mathrm{PLS}}^{(p)\star b}$ la variable aléatoire définie par :

$$z_{i,\mathrm{PLS}}^{(p)\star b} = \frac{\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b} - \tilde{\beta}_{i,\mathrm{PLS}}^{(p)}}{\tilde{s}^{\star}(\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b})} \tag{7}$$

où $\tilde{\beta}_{i,\mathrm{PLS}}^{(p)}$ est le $(p \cdot i)^{\mathrm{e}}$ élément de $\tilde{\boldsymbol{\beta}}_{\mathrm{PLS}}$, $\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b}$ obtenu au b^{e} tirage avec $b = 1, \ldots, \mathrm{B}$ et $\tilde{s}^{\star}(\tilde{\beta}_{i,\mathrm{PLS}}^{(p)\star b})$ l'écart-type estimé de $\tilde{\boldsymbol{\beta}}_{\mathrm{PLS}}^{\star b}$. Soit $\widehat{\mathrm{F}}_{\mathrm{B}}$ la fonction de répartition empirique des $z_{i,\mathrm{PLS}}^{(p)\star b}$. Le fractile d'ordre α, $\widehat{\mathrm{F}}_{\mathrm{B}}^{-1}(\alpha)$ est estimé par la valeur $\hat{t}(\alpha)$ telle que :

$$\frac{1}{\mathrm{B}} \sum_{b=1}^{\mathrm{B}} \mathbb{1}_{\{z_{i,\mathrm{PLS}}^{(p)\star b} \leqslant \hat{t}(\alpha)\}} = \alpha$$

Donc un intervalle de confiance percentile-t pour le $(p.i)^{\mathrm{e}}$ élément de $\boldsymbol{\beta}$ peut s'écrire :

$$\left[\tilde{\beta}^{(p)}_{i,\mathrm{PLS}} - \tilde{s}\left(\tilde{\beta}^{(p)}_{i,\mathrm{PLS}}\right)\cdot\hat{t}(1-\alpha),\ \tilde{\beta}^{(p)}_{i,\mathrm{PLS}} - \tilde{s}\left(\tilde{\beta}^{(p)}_{i,\mathrm{PLS}}\right)\cdot\hat{t}(\alpha)\right] \tag{8}$$

L'évaluation de la qualité du modèle proposé est faite avec l'erreur quadratique moyenne de prédiction MSEP, *Mean Squared Error of Prediction* [10]. La MSEP est utilisée comme critère pour comparer différents modèles dont le modèle moyen [11], défini pour nos donnés par :

$$Y_{ij} = m + g_i + E_j + \delta_{ij} \tag{9}$$

où m est la moyenne de la population et g_i l'effet génotype. L'effet E_j de l'environnement j est supposé aléatoire, d'espérance nulle et de variance σ^2_E. Les erreurs δ_{ij} sont indépendantes, d'espérance nulle et de variance σ^2_δ. De plus, E_j et δ_{ij} sont supposés indépendants.

Le logiciel R [12] a été utilisé la fonction qui a servi pour les régression est de J.-F. Durand [13].

4. Les données utilisées

Nous avons des résultats d'essais agronomiques d'arachide menés de 1994 à 1998 sur la station expérimentale du Ceraas, située à Bambey (14°42N et 16°28O), au Sénégal. Ces essais pluriannuels ont concerné au total 26 génotypes à cycle de développement de 90 jours et répondaient à l'objectif de recherche de génotypes physiologiquement adaptés à la sécheresse.

La variété de référence choisie est la 55-437, c'est une variété hâtive de 90 jours ; elle a donc une longueur de cycle proche de celle des autres variétés utilisées. Elle a été choisie parce que ses données étaient disponibles.

Dans ce milieu à forte variabilité des pluies dans l'espace et même dans le temps pour un même lieu, nous avons considéré chacune des cinq années d'expérimentation comme un environnement (Fig. 1).

Pour valider notre modèle, nous avons réservé successivement chacune des années et estimé les paramètres des génotypes sur les années restantes. Pour chaque année, les rendements observés ont été comparés à ceux prédits par la méthode APLAT. Les rendements sont exprimés en kilogrammes de gousses par hectare.

SarraH a été utilisé pour calculer **X**. Compte tenu du nombre de données disponibles, seuls deux paramètres ($P = 2$) ont été considérés parmi les 61 de SarraH. Le premier paramètre est en fait un coefficient multiplicateur qui agit sur cinq paramètres de SarraH : coefficient

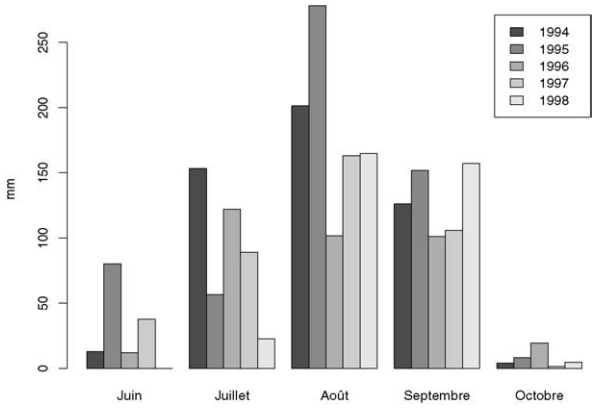


Fig. 1. Répartition des pluies sur la station de Bambey, au Sénégal, de 1994 à 1998.

moyen d'angle des feuilles, coefficient de conversion en assimilat, coefficient d'efficience d'assimilation des feuilles à la phase végétative juvénile, coefficient d'efficience d'assimilation des feuilles à la première phase de maturation, phase sensible de remplissage des grains et coefficient d'efficience d'assimilation des feuilles à la deuxième phase de maturation, phase non sensible. Le deuxième paramètre est le poids moyen des gousses.

5. Résultats

Au Sahel, l'interaction G×E est largement due aux aléas climatiques, dont la probabilité peut être estimée à l'aide de longues chroniques de relevés météo au sol. Cependant, relier l'interaction G×E et la pluviométrie à l'aide d'un modèle de simulation de cultures n'est habituellement possible que pour des variétés dont on a estimé les paramètres, au prix d'une expérimentation spécifique. Le modèle APLAT permet de prédire cette interaction avec les seules données d'une expérimentation multilocale classique, sans autre instrumentation que des stations météo simples.

Pour les modèles sans les données respectivement de 1994, 1995 et 1997, le PRESS minimal est atteint avec six composantes. Pour les deux autres modèles, le PRESS est minimal avec neuf composantes, mais nous avons réduit leur espace à cinq dimensions, car le PRESS n'y est pas trop différent de ses valeurs minimales (Fig. 2).

Les coefficients des régressions PLS et les intervalles de confiance qui leur sont associés sont représentés sur la Fig. 3.

Les MSEP estimées pour les modèles APLAT, sauf celle sans les données de l'année 1998, sont inférieures aux MSEP des modèles moyens correspondants (Tableau 1). Ce qui signifie que, pour ces modèles, pré-

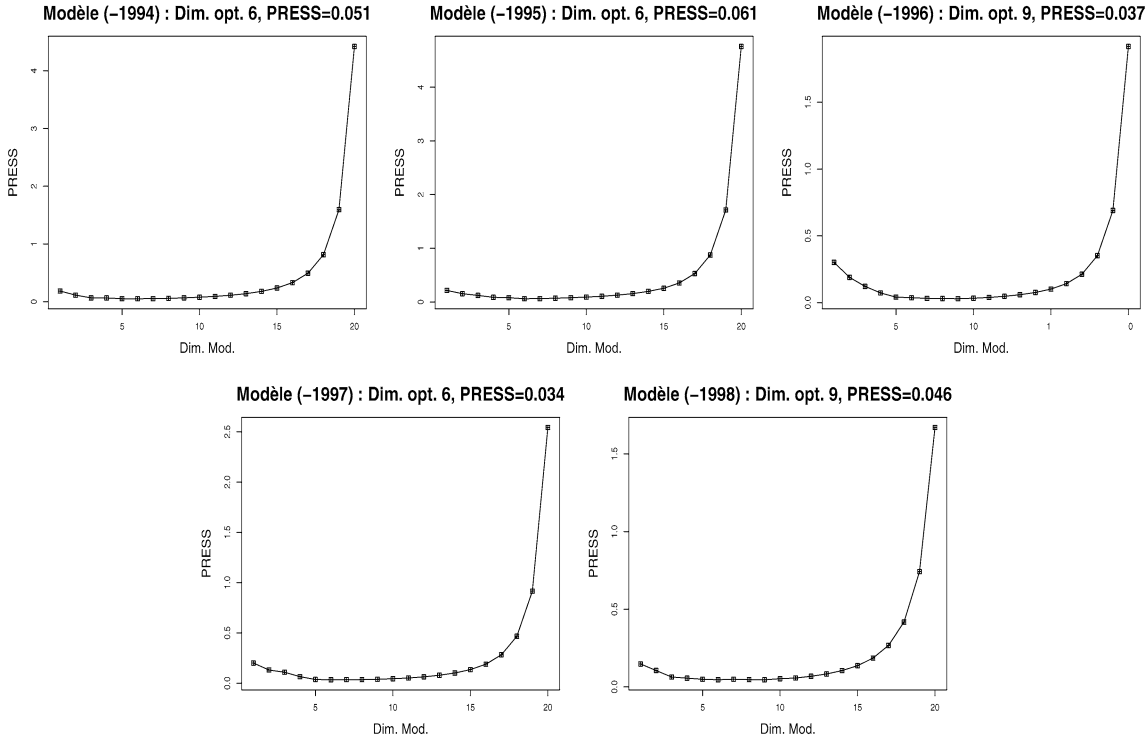


Fig. 2. Evolution du PRESS en fonction du nombre de composantes. Le modèle (-1994) utilise les données, sauf celles de l'année 1994, et ainsi de suite.

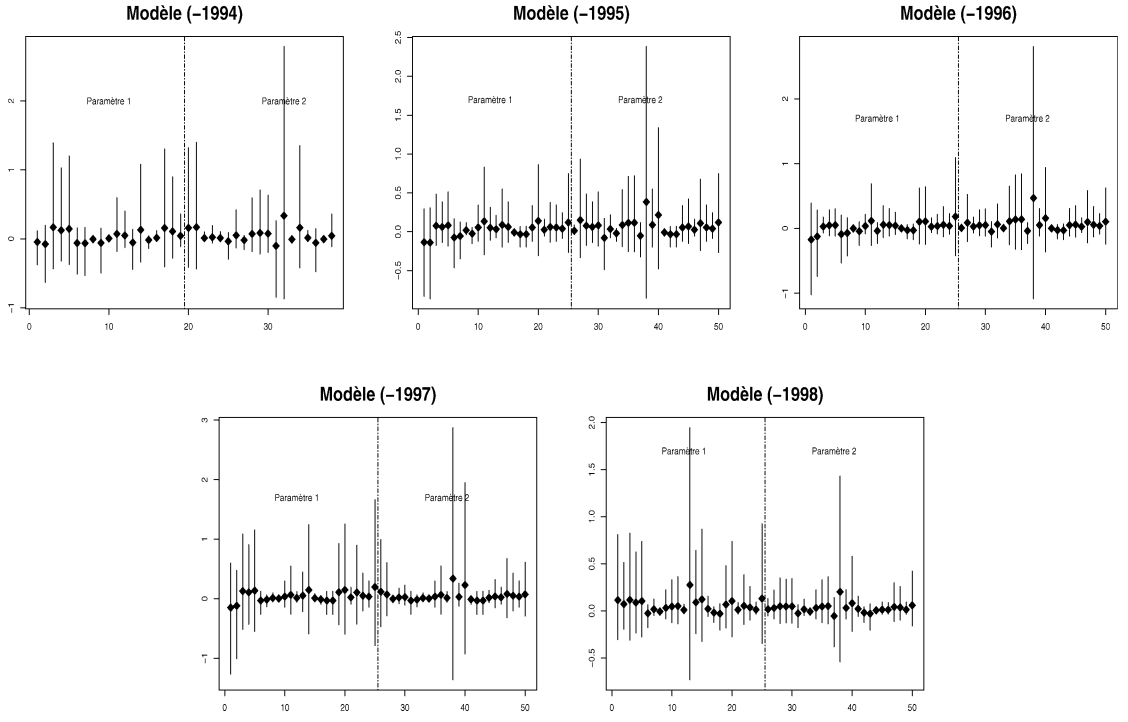


Fig. 3. Intervalle de confiance percentile-t à 95 % des coefficients estimés. Le modèle (-1994) utilise les données, sauf celles de l'année 1994, et ainsi de suite. Sur l'axe des abscisses figurent les génotypes par ordre alphabétique pour chacun des deux paramètres. Le symbole ♦ représente l'estimation des coefficients.

Tableau 1
MSEP des différents modèles APLAT et modèles moyens correspondants. Le modèle (-1994) utilise les données, sauf celles de l'année 1994, et ainsi de suite

	APLAT	Modèle moyen
Modèle (-1994)	24 687,3	64 651,6
Modèle (-1995)	5915,0	7160,6
Modèle (-1996)	35 446,1	37 814,8
Modèle (-1997)	10 038,3	18 201,1
Modèle (-1998)	118 304,9	84 963,6

dire le rendement par la méthode APLAT est meilleur que par la moyenne des rendements du passé. Ainsi, quatre fois sur cinq, la méthode APLAT s'est révélée meilleure que le modèle moyen. Toutefois, cette étude souffre de la faible taille de notre échantillon.

6. Conclusion

La méthode APLAT peut être vue comme un outil d'aide à la décision pour la sélection au Sahel. Dans l'exemple où un sélectionneur doit tester plusieurs génotypes dans un nouvel environnement, cette méthode lui permettra d'écarter d'emblée certains génotypes qui donneront une production faible, en lieu et place d'essais multilocaux ou pluriannuels dans ces environnements contrastés ou d'une tentative de paramétrisation d'un modèle de simulation de cultures qui implique un coût élevé. Son attention sera portée par la suite sur l'ensemble restreint des génotypes retenus avec APLAT, où il pourra appliquer les schémas classiques de sélection.

Cette nouvelle approche semble prometteuse, mais il faut des études supplémentaires. Notamment disposer de données agronomiques plus conséquentes pour l'éprouver.

Remerciements

Nous remercions Danièle Clavel pour les données de l'étude et Jean-Claude Combres pour toutes les discussions autour du modèle SarraH.

Références

[1] AGRHYMET, Bulletins décadaires et mensuels de suivi de la campagne agricole pluviale, Niamey, 1991.
[2] FAO, IRSIS, Irrigation scheduling information system, Rome, 1987.
[3] C. Baron, Modèle de bilan hydrique et de croissance des plantes céréales : Mil Sorgho et Arachide, Cirad, 2002.
[4] M. Vargas, J. Crossa, F.v. Eeuwijk, K.D. Sayre, M.P. Reynolds, Interpreting treatment × environment interaction in agronomy trials, Agron. J. 93 (2001) 949–960.
[5] J.-B. Denis, P. Vincourt, Panorama des méthodes statistiques d'analyse des interactions génotype × milieu, Agronomie 2 (1982) 219–230.
[6] S.A. Eberhart, W.A. Russel, Stability parameters for comparing varieties, Crop Sci. 6 (1966) 36–40.
[7] M. Tenenhaus, La Régression PLS : théorie et pratique, Technip, Paris, 1998.
[8] B. Efron, Bootstrap methods: another look at the jackknife, Ann. Stat. 7 (1979) 1–26.
[9] S. Aji, S. Tavolaro, F. Lantz, A. Faraj, Apport du *bootstrap* à la régression PLS : application à la prédiction de la qualité des gazoles, Oil Gas Sci. Technol.–Rev. IFP 58 (2003) 599–608.
[10] D. Wallach, B. Goffinet, Mean squared error of prediction in models for studying ecological and agronomic systems, Biometrics 43 (1987) 561–573.
[11] J. Colson, D. Wallach, A. Bouniols, J. Denis, J. Jones, Mean squared error of yield prediction by SOYGRO, Agron. J. 87 (1995) 397–407.
[12] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, 2004, URL http://www.R-project.org.
[13] J.-F. Durand, Calcul matriciel et analyse factorielle des données, université Montpellier-2, Montpellier, France, 2002.

RESUME

Ce travail porte sur la prédiction de l'interaction entre génotype et environnement (GxE) et est appliqué au contexte sahélien. Après un tour d'horizon des principales méthodes d'analyse de la littérature, nous proposons la méthode APLAT. Le rendement de génotypes prédit à l'aide de covariables d'environnement par un modèle de simulation de cultures est développé en série de Taylor à l'ordre 1 au voisinage du vecteur de paramètres d'un génotype de référence. Nous nous ramenons alors approximativement à un modèle linéaire où la matrice des régresseurs est remplacée par la matrice des dérivées partielles par rapport aux paramètres. Le très grand nombre de paramètres variétaux généralement constaté dans les modèles de simulation de cultures conduit à un nombre important de régresseurs ; d'où une estimation par régression Partial Least Squares (PLS). Par la suite, nous proposons APLAT-mixte, une extension de APLAT. Pour ce modèle mixte, nous maintenons le rendement des génotypes linéarisé dans la partie fixe, les interactions G×E résiduelles étant aléatoires, de variances inconnues. Nous introduisons à cet effet la technique PLS-Mixte pour estimer les composantes de variance dans un modèle où il y a plus de régresseurs que d'observations. L'algorithme itératif proposé, qui consiste à imbriquer la régression PLS dans l'algorithme Expectation Maximization (EM), est fondé sur les méthodes de maximisation de la vraisemblance Maximum Likelihood (ML) et Restricted Maximum Likelihood (REML).

TITLE

Prediction of the G×E interaction by linearisation and PLS-Mixed regression.

ABSTRACT

This work deals with the prediction of the interaction between genotype and environment (G×E) and is applied to the context of the Sahel. After a literature review of the main methods of analysis, we propose the APLAT method. The yield predicted by a crop-simulation model using environment covariates is expanded as a Taylor series in the neighborhood of a parameter vector of a control genotype. We obtain an approximate linear model where the regressors matrix is replaced by the matrix of the derivatives with respect to the parameters.The sizeable number of varietal parameters generally noted in the crop-simulation models makes the number of regressors exceed the number of observations ; hence an estimatation by Partial Least Squares (PLS) regression is performed. Thereafter, we propose APLAT-mixed, an extension of APLAT. For this mixed model, we keep the yield of the genotypes linearized in the fixed part, the residual G×E interactions being random, of unknown variances. We introduce to that end the PLS-Mixed technique to estimate the variance components in a model where there are more regressors than observations. The iterative algorithm proposed, which consists in imbricating PLS regression in the Expectation Maximization (EM) algorithm, is based on the Maximum Likelihood (ML) and on the Restricted Maximum Likelihood (REML).

DISCIPLINE

Biostatistique

MOTS-CLES

Interaction G×E ; linéarisation ; régression PLS ; alogorithme EM ; ML ; REML.

INTITULE ET ADRESSE DU LABORATOIRE D'ACCUEIL

Unité de recherche Aide à la décision et Biostatistique (Upr 13) - CIRAD
TA 70/09, Avenue dAgropolis, 34398 Montpellier cedex 5 France.