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

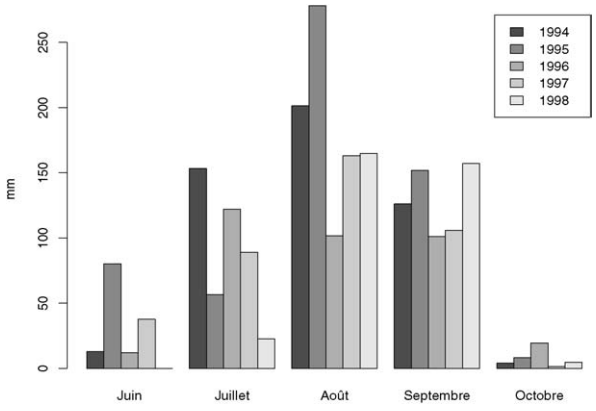

Fig. 1. Répartition des pluies sur la station de Bambey, au Sénégal, de 1994 à 1998.

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-

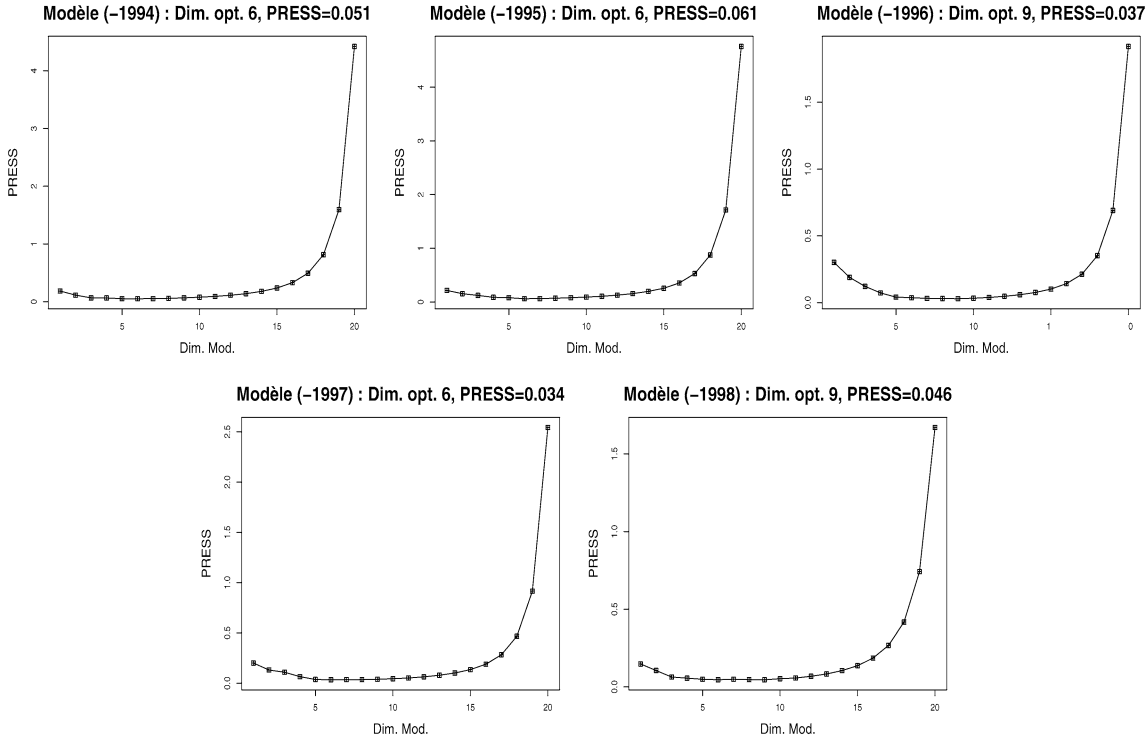

Fig. 2. Evolution du PRESS en fonction du nombre de composantes. Le modèle (-1994) utilise les données, sauf celles de l'année 1994, et ainsi de suite.

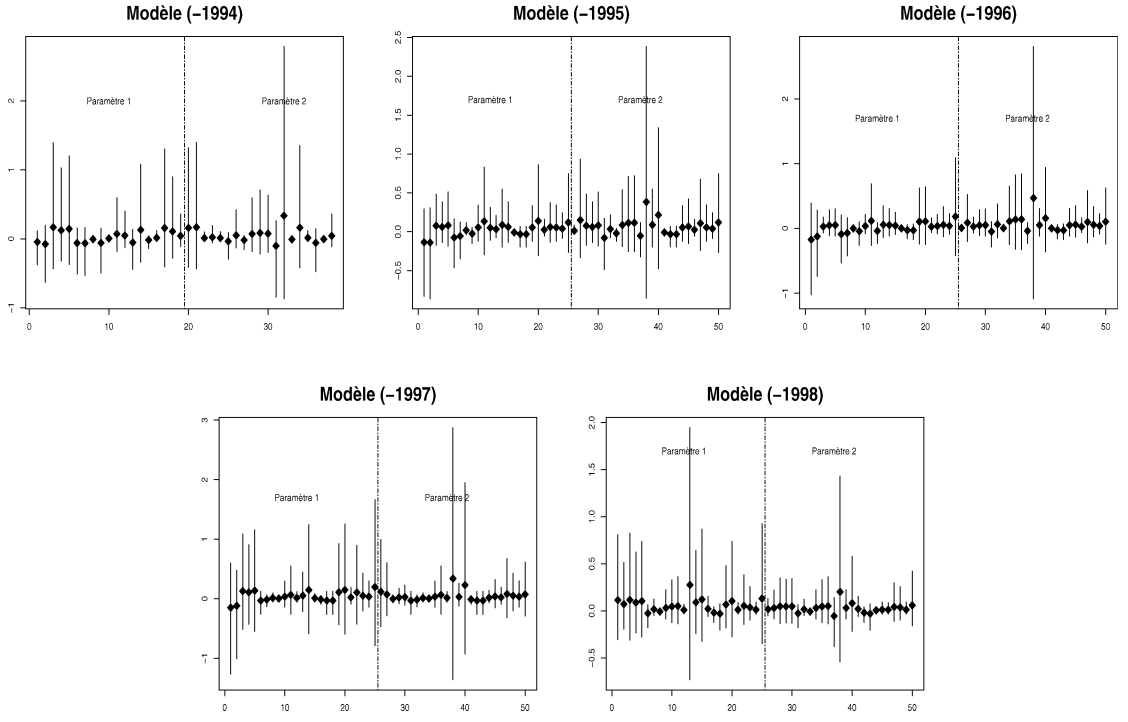

Fig. 3. Intervalle de confiance percentile-t à 95 % des coefficients estimés. Le modèle (-1994) utilise les données, sauf celles de l'année 1994, et ainsi de suite. Sur l'axe des abscisses figurent les génotypes par ordre alphabétique pour chacun des deux paramètres. Le symbole ♦ représente l'estimation 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.