Estimation des paramètres d'un modèle de culture à partir de données de plein champ et de données de plateforme de phénotypage

Jean-Benoist Leger, Estelle Kuhn, Boris Parent, François Tardieu, Claude Welcker

▶ **To cite this version:**

Jean-Benoist Leger, Estelle Kuhn, Boris Parent, François Tardieu, Claude Welcker. Estimation des paramètres d'un modèle de culture à partir de données de plein champ et de données de plateforme de phénotypage. 52èmes Journées de Statistique de la Société Française de Statistique, Société Française de Statistique; Université Côte d'Azur, Jun 2021, Nice, France. hal-03326210

HAL Id: hal-03326210
https://hal.inrae.fr/hal-03326210

Submitted on 25 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimation des paramètres d'un modèle de culture à partir de données de plein champ et de données de plateforme de phénotypage

Jean-Benoist Leger [1] & Estelle Kuhn [2] & Boris Parent [3] & François Tardieu [4] Claude Welcker [5]

[1] *UTC, CNRS, UMR 7253 Heudiasyc, Compiègne, France*
[2] *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France, estelle.kuhn@inrae.fr*
[3] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, boris.parent@inrae.fr*
[4] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, francois.tardieu@inrae.fr*
[5] *INRAE, LEPSE, 2 Place Pierre Viala, 34000 Montpellier, claude.welcker@inrae.fr*

Résumé. Les modèles de culture élaborés par des écophysiologistes décrivent les processus de développement d'une plante. Ils permettent en particulier de rendre compte des différences de comportement de plusieurs variétés dans différents environnements, dues aux interactions génotype-environnement. Pour les utiliser à des fins prédictives, il est nécessaire de calibrer auparavant leurs paramètres. Nous considérons le modèle de culture APSIM et proposons un modèle joint bayésien à effets mixtes dans lequel nous inférons la valeur des paramètres inconnus à partir de données issues d'expérience de plein champ et mesurées en plateforme de phénotypage. Nous choisissons des lois *a priori* informatives pour intégrer les connaissances d'expert et implémentons un algorithme de type Gibbs hybride pour simuler la loi *a posteriori*. Les résultats obtenus sur données simulées et réelles mettent en évidence le gain obtenu sur la précision des estimations en utilisant les données issues de plateforme de phénotypage en sus des données du champ.

Mots-clés. Modèle de culture, données hétérogènes, modèles à effets mixtes, modèle bayésien, algorithme Gibbs hybride.

Abstract. Crop models were developed by ecophysiologists to describe plant development. They allow in particular to report difference existing between several genotypes in several environments, due to genotype by environment interaction. It is first necessary to calibrate these models to use them for prediction purpose. We consider the crop model APSIM and present a joint bayesian model with mixed effects. We infer models parameter values from data collected in the field and in phenotyping platform. Prior distribution are chosen in order to integrate expert knowledge. We implement an hybrid Gibbs algorithm to simulate the posterior distribution. Results obtained from simulated and real data highlight clearly the advantage of using phenotyping platform data in addition to field data.

Keywords. crop model, heterogeneous data, mixed effects models, bayesian model, Gibbs hybrid algorithm.

1 Introduction

1.1 Contexte de l'amélioration des plantes

Un des enjeux actuels en sciences du végétal vise à mieux comprendre les mécanismes impliqués dans le développement des plantes et leurs réponses aux conditions environnementales. Ces mécanismes diffèrent d'une espèce à l'autre, chacune ayant des phases de développement spécifiques. Au sein d'une même espèce, les différents processus qui se succèdent au cours de la croissance ont lieu de façon différente selon la variété considérée : ils se réalisent plus ou moins rapidement, à des périodes plus ou moins précoces, donnant lieu à une importante variabilité de comportements. Mieux comprendre comment cette variabilité dans les processus de croissance est reliée au génotype caractérisant la variété est un objectif essentiel en amélioration des plantes.

De plus, une forte interaction existe entre la variété considérée et l'environnement, incluant non seulement les aspects météorologiques mais également la composition du sol, les intrants, etc. Ainsi, une même variété va évoluer différemment dans différents environnements et différentes variétés vont se comporter différemment dans un même environnement (cf. Millet et al. (2016)). Mieux comprendre ces interactions génotype-environnement-conduite de culture est un levier important pour fournir de meilleures recommandations dans le choix des variétés selon l'environnement, en particulier dans un contexte de changement climatique fort, incluant de plus en plus d'événements climatiques extrêmes (cf. Millet et al. (2019)).

La modélisation mathématique est un outil extrêmement pertinent pour mieux comprendre, quantifier et prédire ces interactions. Des modèles linéaires ont tout d'abord été utilisés, donnant lieu à un cadre relativement limité du point de vue de la modélisation des effets génotypiques et environnementaux. Plus récemment, des modèles descriptifs des mécanismes de croissance des plantes, appelés modèles de culture, ont été développés par des écophysiologistes des plantes, comme par exemple le modèle APSIM (cf. Keating et al (2003)). Ces modèles dynamiques rendent compte des processus qui interviennent lors du développement de la plante. Ils utilisent en entrée des variables environnementales et des paramètres dépendant du génotype et fournissent en sortie des caractères plus ou moins intégrés de la plante comme par exemple la date de floraison, le rendement, la biomasse au cours du temps. Ces modèles descriptifs permettent également de modéliser les interactions génotype-environnement-conduite de culture. Utiliser à des fins prédictives, ils sont un outil efficace pour prédire ces interactions. Cependant un grand nombre de paramètres de ces modèles sont généralement inconnus et doivent être calibrés. La valeur des paramètres peut être ajustée manuellement en comparant les sorties du modèle à des données. Cette approche requiert néanmoins beaucoup de temps, d'autant plus si le nombre de paramètres est important. Une approche plus rapide consiste à ajuster un modèle statistique basé sur le modèle de culture à partir des données disponibles, en inférant la valeur des paramètres via un estimateur statistique. Ce type d'approche reste néanmoins

difficile à mettre en oeuvre du fait de la complexité du modèle de culture et demande la mise en place de méthodes numériques efficaces. Des premières approches ont été proposées (cf. Cooper et al (2016)). Cependant elles ne permettent de traiter qu'un nombre réduit de données et d'estimer un petit nombre de paramètres.

1.2 Le modèle de culture APSIM

Nous considérons le modèle de culture Agricultural Production Systems sIMulator (APSIM) avec le module maïs et une extension du module feuille réalisée par l'unité INRAE LEPSE (Lacube et al. (2017)). Il s'agit d'un modèle à pas de temps journalier qui simule un couvert constitué de plantes moyennes en mettant à jour un vecteur de descripteurs de la plante qui évolue au cours du temps. Les entrées de ce modèle sont des covariables météorologiques, des covariables descriptives du sol et de la conduite de culture. Il fournit en sortie la date de floraison et les composantes du rendement. Certains paramètres ont un sens physique, comme par exemple l'efficacité d'interception lumineuse, d'autres ne sont pas interprétables. Par ailleurs, certains processus sont communs à toute l'espèce, leurs paramètres ont une valeur commune à tous les génotypes de cette espèce, tandis que d'autres processus sont variables génétiquement, leurs paramètres ont des valeurs différentes selon le génotype, comme par exemple le nombre final de feuilles du plant.

1.3 Les données de sources hétérogènes

Nous disposons d'un riche jeu de données issu du projet DROPS European Project incluant un panel de diversité composé de 230 génotypes hybrides observés en plein champ dans 13 conditions environnementales. Une condition environnementale est définie par un lieu et une année. Les lieux sont répartis en Europe du nord au sud, rendant compte de conditions climatiques très contrastées. Les années considérées varient de 2012 à 2013. Les données comprennent la date de floraison, les composantes du rendement (nombre de grains, poids d'un grain) et les conditions environnementales.

De plus, des expériences auxiliaires complémentaires ont été réalisées sur la plateforme INRAE PhénoArch à Montpellier (Cabrera-Bosquet et al. (2016)). Ces expériences ont permis d'obtenir des données supplémentaires sur des paramètres mécanistes intervenant dans le modèle de culture. Ainsi, des mesures de quantités telles que le nombre final de feuilles observées ont été effectuées en plateforme, apportant des informations complémentaires au jeu de données obtenu en plein champ.

L'objectif est d'utiliser les deux sources d'information champ et plateforme dans la procédure d'inférence statistique des paramètres du modèle de culture.

2 Modélisation statistique

2.1 Modélisation de la variabilité génotypique

Nous disposons pour chaque génotype du panel DROPS de mesures répétées du rendement, du nombre de grains et de la date de floraison dans M conditions environnementales. Nous considérons un modèle statistique à effets mixtes (cf. Pinheiro et Bates (2000)) basé sur le modèle de culture APSIM qui permet de prendre en compte simultanément les variabilités inter-génotype et intra-génotype. On note Y_{gm} la mesure vectorielle dans la condition expérimentale m pour le génotype g et on modélise pour tout g et tout m :

$$\log Y_{gm} = \log G(e_m, \beta_g, \gamma_g) + \varepsilon_{gm} \tag{1}$$

où G représente la sortie du modèle de culture APSIM, e_m le vecteur de variables descriptives de l'environnement m, β_g le vecteur des paramètres du génotype g non mesurées en plateforme, γ_g le vecteur des paramètres du génotype g mesurées en plateforme, ε_{gm} un terme d'erreur supposé gaussien centré de variance diagonale $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$. On suppose que les vecteurs d'effets aléatoires (β_g) et (γ_g) sont indépendants identiquement distribués de loi gaussienne d'espérance $\bar{\beta}$ resp. $\bar{\gamma}$, et de matrice de covariance diagonale Σ_β, resp. Σ_γ.

2.2 Modélisation des données de plateforme de phénotypage

Nous considérons un modèle joint pour intégrer les données issues de la plateforme à l'inférence des paramètres du modèle de culture. Pour cela, nous modélisons les mesures des paramètres effectuées en plateforme de phénotypage via un modèle linéaire en fonction de la valeur des paramètres du modèle de culture APSIM. On note Z_g le vecteur de taille p des mesures de paramètres du génotype g. Pour $1 \leq l \leq p$, on a :

$$Z_{g,l} = \mu_l + \zeta_l \gamma_{g,l} + \eta_{g,l} \tag{2}$$

où μ et ζ sont des vecteurs inconnus de $\mathbb{R}^p$ et $\eta_{g,l}$ un terme d'erreur résiduel supposé gaussien centré de variance τ_l^2.

3 Inférence bayésienne du modèle joint

Du fait de la complexité du modèle de culture APSIM et du grand nombre de paramètres à estimer, nous faisons le choix d'une approche bayésienne qui va permettre de régulariser la procédure d'estimation. Nous souhaitons choisir des lois *a priori* uniformes pour les paramètres mécanistes du modèle de culture qui prennent leurs valeurs dans des intervalles bornés fixés suivant des dires d'expert. Toutefois, pour des raisons computationnelles, nous avons effectué une reparamétrisation du modèle, et les nouveaux paramètres sont à valeurs réelles. Nous choisissons pour ces paramètres des lois *a priori* normales, telles que leurs transformées par la reparamétrisation inverse soient les plus proches au sens de

la divergence de Kullback-Leibler des lois uniformes de départ. Pour les paramètres μ et ζ du modèle des données issues de la plateforme, nous choisissons des lois *a priori* normales centrées sur la valeur attendue, 0 pour l'ordonnée à l'origine et 1 pour la pente. Nous fixons des lois inverse gamma qui sont conjuguées pour les lois *a priori* des paramètres de variances des bruits.

Nous appliquons un algorithme de Monte Carlo Markov Chain de type Gibbs hybride (cf. Carlin et Louis (2008)) pour générer une chaine de Markov qui sous des hypothèses de régularité du modèle est ergodique et a pour loi stationnaire la loi *a posteriori*. A partir des réalisations de cet algorithme, nous construisons des estimateurs empiriques des lois *a posteriori* des paramètres du modèle.

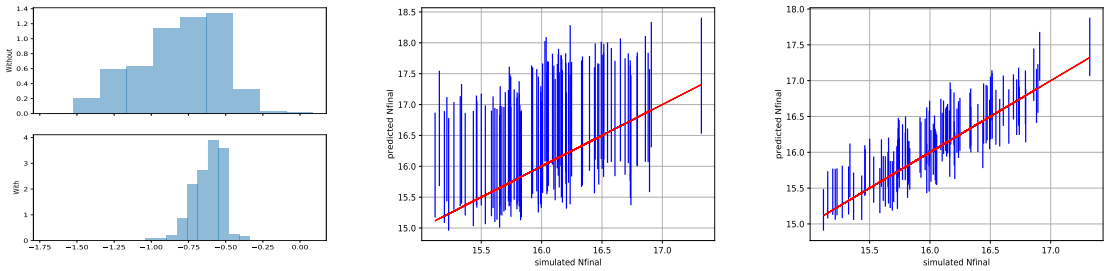

Figure 1: Histogramme de la loi *a posteriori* de Nfinal sans (gauche, haut) et avec (gauche, bas) les données plateforme supplémentaires ; Simulations versus prédictions via intervalles de crédibilité à 90% pour Nfinal sans (centre) et avec (droite) les données plateforme supplémentaires.

4 Expériences numériques

Nous effectuons une étude de simulation en considérant les 13 environnements réels du jeu de données DROPS et les valeurs des paramètres proches de ceux du génotype de référence *B*73. Nous simulons 100 génotypes. Nous estimons les trois paramètres du modèle correspondants au nombre final de feuilles (noté Nfinal), au premier ligulochrone et au poids moyen potentiel d'un grain, les autres étant fixés à la valeur de référence. Nous mettons en évidence que les estimateurs obtenus à partir des données issues du champ et de la plateforme dans le modèle joint sont plus précis que les estimateurs obtenus à partir des seules données issues du champ dans le modèle initial (cf Figures 1 et 2 gauche).

Nous ajustons ensuite le modèle proposé aux données réelles. Les prédictions obtenues à partir du modèle avec les paramètres estimés à partir des données issues du champ et de la plateforme sont meilleures que celles obtenues avec les paramètres estimés à partir des seules données champ (cf Figure 2 droite).

Ce travail a été financé par le projet AMAIZING ANR-10-BTBR-01. Les auteurs remercient la plateforme MIGALE, INRAE, 2020, Migale bioinformatics Facility pour les moyens de calcul et les capacités de stockage.

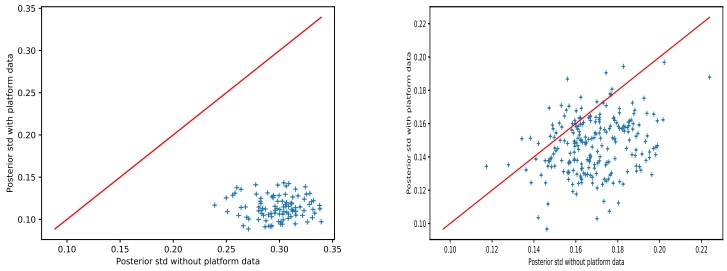

Figure 2: Ecart-types de la distribution *a posteriori* de Nfinal obtenus sans (abscisse) et avec (ordonnée) les données plateforme supplémentaires en simulation (gauche) et sur données réelles (droite).

Bibliographie

Cabrera-Bosquet, L., et al., (2016), High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. *New Phytologist*, 212, (1), 269-281.

Carlin, B.P. and Louis, T. (2008), Bayesian methods for data analysis, *Chapman and Hall/CRC*.

Cooper, M. and Technow, F. and Messina, C. and Gho, C and Totir, L. R. (2016), Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial, *Crop Science*, 56, (5), 2141–2156.

Keating, B. and Carberry, P. and Hammer, G. and Probert, M. and Robertson, M. and Holzworth, D. and Huth, N.and Hargreaves, J. and *et al.*, (2003), An overview of APSIM, a model designed for farming systems simulation, *European journal of agronomy*, 18, (3-4), 267–288.

Lacube, S., et al., (2017) Distinct controls of leaf widening and elongation by light and evaporative demand in maize, *Plant Cell and Environment*, 40, (9), 2017-2028.

Millet E., Welcker C, Kruijer W, Negro S, Coupel-Ledru A, et al., (2016), Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios, *Plant Physiol*, 172, 749-764.

Millet, E. and Kruijer, W. and Coupel-Ledru, A. and Prado, S.A. and Cabrera-Bosquet, L. and Lacube, S. and Charcosset, A. and Welcker, C. and van Eeuwijk, F. and Tardieu, F., (2019), Genomic prediction of maize yield across European environmental conditions, *Nature genetics*, 51, (6), 952–956.

Pinheiro, J.C. and Bates D.M. (2000), Mixed-Effects Models in S and S-PLUS, *Springer*.