Modélisation de la variabilité inter-individuelle dans les populations de plantes par un modèle non linéaire mixte et estimation par variants stochastiques de l'algorithme EM.

Charlotte Baey [1] & Samis Trevezas [1] & Amélie Mathieu [2] & Alexandra Jullien [2] & Paul-Henry Cournède [1]

[1] *Laboratoire MAS - Grande Voie des Vignes 92290 Châtenay-Malabry charlotte.baey@gmail.com,*
[2] *INRA AgroParisTech - 78850 Thiverval-Grignon France*

Résumé. Il existe une forte variabilité génétique entre plantes, même au sein de la même variété, ce qui, combiné à la variation locale des conditions climatiques dans le champ, peut conduire deux plantes voisines à se développer de façon très différentes. C'est l'une des raisons pour lesquelles les approches populationnelles dans les modèles de croissance de plantes suscitent un grand intérêt. Nous proposons dans cette étude une extension du modèle individu-centré Greenlab à l'échelle de la population dans le cas du colza, à l'aide d'un modèle non linéaire mixte. Deux variants stochastiques de l'algorithme EM (Espérance-Maximisation), le Monte-Carlo EM automatique (MCEM) et le SAEM seront comparés, en utilisant le fait que le modèle complet appartient à la famille exponentielle.

Mots-clés. variabilité, modèles de croissance de plantes, algorithme EM, Monte Carlo par Chaîne de Markov

Abstract. There is a strong genetic variability among plants, even of the same variety, which, combined with the locally varying climatic effects in a given field, can lead to the development of highly different neighboring plants. This is one of the reasons why population-based methods for modeling plant growth are of great interest. In this study, we extend the formulation of the individual-based model Greenlab to the population level for rapeseed plants using nonlinear mixed models. Stochastic variants of an EM-type algorithm (Expectation-Maximization) will be compared (the automated Monte-Carlo EM (MCMC-EM) and the SAEM algorithm), using the fact that the complete data distribution belongs to the exponential family of distributions.

Keywords. variability, plant growth models, EM algorithm, Markov Chain Monte Carlo

1 Introduction

Il existe à l'état naturel une forte variabilité génétique entre plantes, parfois même au sein de la même espèce, gage d'une meilleure résistance aux maladies ou aux insectes ravageurs,

et d'une meilleure capacité d'adaptation à de nouvelles conditions environnementales. De la même façon, même au sein d'une parcelle agricole donnée, des variations locales dans le sol ou les conditions climatiques peuvent entraîner de fortes disparités entre les plantes.

Cette variabilité peut avoir un fort impact à l'échelle agronomique, pourtant elle est rarement prise en compte dans la pratique. En effet l'estimation se fait souvent à partir d'une plante moyenne obtenue à partir d'observations individuelles, et l'extrapolation à l'échelle de la population n'est pas si immédiate.

Nous proposons dans cet article une extension du modèle de croissance de plantes individu-centré Greenlab [3] à l'échelle de la population en utilisant la structure hiérarchique des modèles mixtes. Les paramètres sont estimés par maximum de vraisemblance, à l'aide de l'algorithme d'Espérance-Maximisation (EM) [5]. Cependant, l'étape E n'est en général pas explicite, à cause de la non linéarité du modèle, et il est alors nécessaire de recourir à des approximations stochastiques. Dans notre cas, la densité des données complètes appartient à la famille exponentielle, et l'étape M sera alors explicite. Nous comparons dans cette étude les deux algorithmes MCMC-EM [10, 7, 8] et SAEM [4].

2 Modèle

Le modèle Greenlab [3] est un modèle de type structure fonction, prenant en compte à la fois l'architecture de la plante et son développement à l'échelle de l'organe, et les processus écophysiologiques impliqués dans la croissance de la plante. La production de biomasse au jour t est donnée par :

$$q_{pl}(t) = 0.95 \cdot \mu \cdot s^{pr} \cdot \mathrm{PAR}(t) \cdot \left(1 - \exp\left(-k_b \, \frac{s^{act}(t)}{s^{pr}}\right)\right), \tag{1}$$

où μ correspond à l'efficience de conversion en biomasse, s^{pr} est un paramètre relié à la surface qu'occupe la plante au sol, PAR est le rayonnement photosynthétiquement actif, k_b est un paramètre (connu) lié à l'absorption de la lumière, et s^{act} la surface foliaire photosynthétiquement active de la plante au début du jour t.

L'allocation se fait ensuite proportionnellement aux forces d'attraction de chaque organe. Dans le cas du colza au stade rosette, seules les feuilles sont prises en compte dans le modèle. La fonction puits de la feuille de rang k au temps u est donnée par :

$$s_k(u) = c \, \left(\frac{\tau(u) - \tau_k}{\tau_k^e}\right)^{a-1} \left(1 - \frac{\tau(u) - \tau_k}{\tau_k^e}\right)^{b-1} \mathbf{1}_{\tau_k \leq \tau(u) \leq \tau_k + \tau_k^e}, \tag{2}$$

où $\tau(u)$ est le temps thermique (somme cumulée de températures depuis semis) au temps u, τ_k est le temps thermique d'initiation de la feuille k, et τ_k^e est le temps thermique d'expansion de la feuille k, et c est une constante de normalisation.

La somme des puits de toutes les feuilles au jour u définit la demande totale de la plante en biomasse $d(u)$, et le rapport $s_k(u)/d(u)$ détermine le pourcentage de biomasse produite qui sera allouée à la feuille k à la fin du jour u.

Nous notons $y_i = (y_i(t_1), \ldots, y_i(t_{n_i}))$ le vecteur d'observation des biomasses des feuilles initiées aux temps $t_1, \ldots, t_{n_i}$ pour la plante i. Nous considérons deux formulations pour le modèle de population en fonction du type d'erreur considérée (additive ou log-additive) :

$$\begin{cases} y_i(t_n) = G_n(\phi_i)\ e^{\varepsilon_{i,n}} \quad \text{ou} \quad y_i(t_n) = G_n(\phi_i) + \varepsilon_{i,n}, \\ \varepsilon_{i,n} \sim \mathcal{N}(0, \sigma_l^2). \\ \phi_i = \beta + \xi_i, \qquad \xi_i \sim \mathcal{N}_P(0, \Gamma). \end{cases} \tag{3}$$

où la fonction G_n représente la biomasse théorique de la feuille initiée au temps t_n (donnée par le modèle Greenlab), et $\phi_i = (\phi_{i,1}, \ldots, \phi_{i,P})^t$ le vecteur de paramètres spécifiques de la plante i. Le vecteur de paramètres à estimer est $\theta = (\beta, \Gamma, \sigma_l^2)$.

3 Estimation

L'ensemble des paramètres θ peut se décomposer $\theta_1 = (\beta, \Gamma)$ et $\theta_2 = \sigma_l^2$. En supposant que chaque effet aléatoire a une variance non nulle, la vraisemblance complète appartient à la famille exponentielle et s'écrit :

$$f(\tilde{y}, \phi; \theta) = \exp\left\{\langle s_1(\theta_1), t_1(\phi)\rangle - a_1(\theta_1)\right\} \exp\left\{\langle s_2(\theta_2), t_2(\tilde{y}, \phi)\rangle - a_2(\theta_2)\right\}, \tag{4}$$

$$s_1(\theta_1) = \begin{pmatrix} \Gamma^{-1}\beta \\ \Gamma^{-1} \end{pmatrix} \quad t_1(\phi) = \begin{pmatrix} \sum_{i=1}^s \phi_i \\ -\frac{1}{2}\sum_{i=1}^s \phi_i \phi_i^t \end{pmatrix}, \tag{5}$$

$$s_2(\theta_2) = \sigma_l^{-2} \qquad t_2(\tilde{y}, \phi) = -\frac{1}{2}\sum_{i=1}^{s}\sum_{n=1}^{n_i}\left(\tilde{y}_i(t_n) - \tilde{G}_n(\phi_i)\right)^2, \tag{6}$$

$$a_1(\theta_1) = \frac{sP}{2}\log 2\pi + \frac{s}{2}\log|\Gamma| + \frac{s}{2}\beta^t\Gamma^{-1}\beta, \qquad a_2(\theta_2) = \frac{N}{2}\log 2\pi + \frac{N}{2}\log\sigma_l^2. \tag{7}$$

3.1 Étape E

L'étape E peut s'écrire en fonction des statistiques exhaustives [9]. Dans notre cas, l'étape E n'est pas explicite, car la loi $f(\phi \mid y; \theta)$ est inconnue, et est remplacée par une étape de simulation pour l'algorithme MCMC-EM, ou d'approximation stochastique pour l'algorithme SAEM. Dans les deux cas, cela nécessite de simuler une chaîne de Markov de loi stationnaire $f(\phi \mid y; \theta)$, et pour cela nous utilisons un algorithme de Metropolis-Hastings à marche aléatoire adaptative composante par composante [1]. À partir de la chaîne de Markov de taille m_k générée à l'itération k de l'algorithme, l'étape E s'écrit :

- pour l'algorithme MCMC-EM,

$$t_1^{(k)} = \frac{1}{m_k}\sum_{m=1}^{m_k} t_1(\phi^{k,(m)}), \qquad t_2^{(k)} = \frac{1}{m_k}\sum_{m=1}^{m_k} t_2(\tilde{y}, \phi^{k,(m)}) \tag{8}$$

- pour l'algorithme SAEM,

$$t_1^{(k)} = t_1^{(k-1)} + \gamma_k \left[\frac{1}{m_k} \sum_{m=1}^{m_k} t_1(\phi^{k,(m)}) - t_1^{(k-1)} \right] \quad t_2^{(k)} = t_2^{(k-1)} + \gamma_k \left[\frac{1}{m_k} \sum_{m=1}^{m_k} t_2(\tilde{y}, \phi^{k,(m)}) - t_2^{(k-1)} \right]. \tag{9}$$

Nous utilisons la version automatique de l'algorithme MCMC-EM proposée par [2], qui repose sur la propriété de monotonie de l'algorithme EM et qui permet d'augmenter à chaque itération de l'algorithme la taille de la chaîne de Markov, pour diminuer l'erreur de Monte Carlo et assurer la convergence de l'algorithme. Un critère d'arrêt peut également être dérivé.

3.2 Étape M

Lorsque tous les éléments du vecteur ϕ_i sont considérés comme aléatoires, l'étape de maximisation est explicite. Grâce à la formulation sous forme de modèle exponentiel et à la décomposition du vecteur de statistiques exhaustives en deux sous-vecteurs, maximiser θ revient à résoudre les deux équations suivantes :

$$\mathbb{E}_\theta(t_1(x)) = t_1^{(k)}, \qquad \mathbb{E}_\theta(t_2(x)) = t_2^{(k)}.$$

On obtient les équations suivantes, en supposant $\Gamma = \text{diag}(\sigma_1^2, \ldots, \sigma_P^2)$:

$$\hat{\beta}_j = \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}_{\theta^k}(\phi_{i,j} \mid \tilde{y}_i), \qquad \hat{\sigma}_j^2 = \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}_{\theta^k}(\phi_{i,j}^2 \mid \tilde{y}_i) - \hat{\beta}_j^2, \quad j = 1, \ldots, P, \tag{10}$$

$$\hat{\sigma}_l^2 = \frac{1}{N} \sum_{i=1}^{s} \sum_{n=1}^{n_i} \mathbb{E}_{\theta^k} \left[(\tilde{y}_{i,n} - \tilde{G}_n(\phi_i))^2 \mid \tilde{y}_i \right]. \tag{11}$$

4 Application au cas du colza

Nous disposons des masses sèches individuelles des feuilles de 34 plants de colza, provenant d'expérimentations réalisées en 2012-2013 à la station expérimentale de l'INRA à Grignon (N 48°51'20" E1°56'25") sur la variété Pollen [6].

Les résultats obtenus avec les deux algorithmes MCMC-EM et SAEM sont similaires, l'algorithme MCMC-EM (automatique) étant plus rapide en temps d'exécution que le SAEM (non automatique). Les deux types d'erreur ont été comparées, ainsi que le nombre de paramètres aléatoires. L'erreur additive permet d'obtenir des valeurs beaucoup plus faibles de AIC et BIC. Le meilleur modèle au sens de ces critères est ensuite celui où les deux paramètres μ et a sont considérés comme aléatoires (voir résultats Tableau 1.

Table 1: Résultats obtenus avec les algorithmes MCMC-EM automatique et SAEM.

Param.	MCMC-EM			SAEM		
	Estimation	ET	IC	Estimation	ET	IC
β_μ	1.1151	0.0155	[1.0848; 1.1454]	1.1151	0.0151	[1.0854 ; 1.1448]
σ_μ	0.0866	0.0123	[0.0625 ; 0.1107]	0.0873	0.0111	[0.0656 ; 0.1091]
β_a	0.5799	0.0204	[0.5400 ; 0.6198]	0.5804	0.0119	[10.5571 ; 0.6037]
σ_a	0.0631	0.0174	[0.0290 ; 0.0972]	0.0574	0.0089	[0.0400 ; 0.0747]
σ_l^2	0.0043	0.00037	[0.0036 ; 0.0050]	0.0043	0.00035	[0.0036 ; 0.0050]

Les résultats obtenus suggèrent que le paramètre d'allocation a permet de mieux prendre en compte la variabilité inter-individuelle que le paramètre b, puisque les deux modèles contenant ce paramètre sont meilleurs que ceux contenant b, au sens des deux critères AIC et BIC. Ce paramètre permet de modéliser différentes stratégies d'allocation de la biomasse aux feuilles : plus la valeur de a est faible, plus l'allocation se fait tôt. Il porte sur la première partie de la courbe d'allocation, qui correspond au début de l'expansion des feuilles et est donc particulièrement importante puisque certains organes sont encore en expansion au moment où les mesures ont été faites. En revanche, l'ajout du paramètre s^{pr} comme effet aléatoire ne paraît pas pertinent. Cela pourrait suggérer que la variabilité inter-individuelle est suffisamment bien prise en compte par l'introduction de deux effets aléatoires sur les paramètres μ et a. Ceci pourrait s'expliquer par un faible effet de la compétition dans la première phase de croissance du colza, lorsque la plante est encore au stade rosette. Et en effet dans la pratique ce paramètre est souvent supposé constant et égal à l'inverse de la densité [6].

La figure 1 représente la distribution des observations prédites par le modèle Greenlab de population. Nous avons également représenté sur la figure les observations correspondant à la plante moyenne utilisée habituellement pour calibrer le modèle Greenlab individuel.

Les axes futurs de recherche et d'amélioration incluent notamment l'introduction d'effets fixes dans le modèle, ou l'utilisation d'une matrice Γ non diagonale. De même, l'implémentation d'une version automatique de SAEM permettrait une meilleure comparaison de deux algorithmes.

References

[1] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18:343–373, 2008.

[2] B. S. Caffo, W. Jank, and G. L. Jones. Ascent-based Monte Carlo expectation-maximization. 67(2):235–251, 2005.

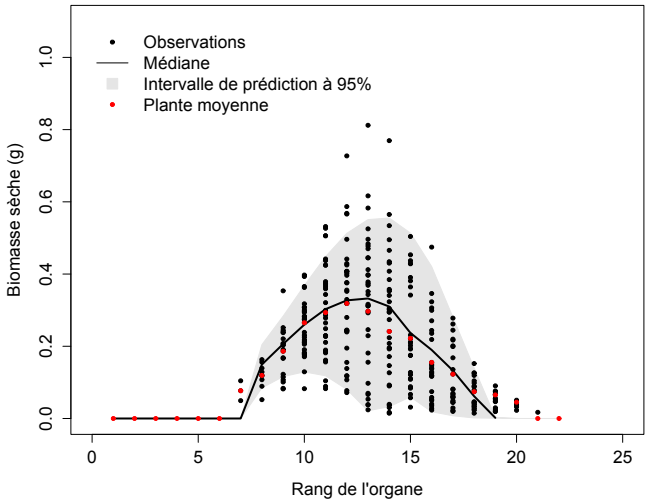

Figure 1: Prédictions du modèle Greenlab de population pour le colza. Les points rouges correspondent à la plante moyenne qui est utilisée classiquement pour calibrer le modèle Greenlab individuel, la ligne continue correspond à la médiane des prédictions et la zone grisée aux quantiles d'ordre 5% et 95%.

[3] P. de Reffye and B.-G. Hu. Relevant qualitative and quantitative choices for building an efficient dynamic plant growth model: GreenLab case. pages 87–107. Tsinghua University Press and Springer, 2003.

[4] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[5] A. Dempster, N. M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. 39(1):1–38, 1977.

[6] A. Jullien, A. Mathieu, J.-M. Allirand, A. Pinet, P. de Reffye, P.-H. Cournède, and B. Ney. Characterisation of the interactions between architecture and source:sink relationships in Winter Oilseed Rape (Brassica Napus L.) using the GreenLab model. *Annals of Botany*, 107(5):765–779, 2011.

[7] C. E. McCulloch. Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89(425):330–335, 1994.

[8] C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997.

[9] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 2nd edition, 2007.

[10] G. C. G. Wei and M. A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.