Modélisation de la variabilité inter-individuelle dans les modèles de croissance de plantes et sélection de modèles pour la prévision

Charlotte Baey

ECOLE CENTRALE DES ARTS
ET MANUFACTURES
"ECOLE CENTRALE PARIS"

Thèse
préparée par Charlotte BAEY
pour l'obtention du grade de Docteur

Spécialité : Mathématiques appliquées
Laboratoire d'accueil : Mathématiques Appliquées aux Systèmes (MAS)

Modélisation de la variabilité inter-individuelle dans les modèles de croissance de plantes et sélection de modèles pour la prévision

Soutenue le 28 février 2014 N° d'ordre : 2014ECAP0024

Devant un jury composé de :

M. Paul-Henry Cournède	Directeur de thèse
M. Philippe de Reffye	Rapporteur
M. Jean-Louis Foulley	Examinateur
Mme. Anne Goelzer	Examinatrice
Mme. Estelle Kuhn	Examinatrice
M. Marc Lavielle	Rapporteur
M. Samis Trevezas	Examinateur

« La vie n'est pas ce que tu crois. C'est une eau que les jeunes gens laissent couler sans le savoir, entre leurs doigts ouverts. Ferme tes mains, ferme tes mains, vite. Retiens-la. Tu verras, cela deviendra une petite chose dure et simple qu'on grignote, assis au soleil. »

Jean Anouilh, ***Antigone***.

Remerciements

Enfin ! Après plusieurs mois intenses, épuisants mais également grisants, au cours desquels on grandit finalement plus que pendant tous les mois qui ont précédé cette phase de rédaction, après le point culminant de la soutenance, voici venu le moment de remercier ceux sans qui tout ceci n'aurait pas pu aboutir.

Tout d'abord, je tiens à remercier Paul-Henry, mon directeur de thèse, pour m'avoir fait confiance il y a maintenant presque quatre ans, lorsque j'ai décidé de me (ré)orienter vers la recherche. Quand on sait à quel point les relations entre directeur de thèse et doctorant peuvent parfois être difficiles, je mesure la chance que j'ai eue d'avoir pu compter sur ton soutien, tes encouragements, ta confiance, tout au long de ma thèse. En première année d'abord, quand tu m'as encouragée à poursuivre mes idées, en deuxième année ensuite, lorsque tu as su me montrer ce que parfois je voyais un peu moins nettement, et en troisième année, enfin, quand tu m'as laissée prendre petit à petit mon envol. Merci pour ta confiance, ta disponibilité, tes conseils, et pour avoir su créer au sein de l'équipe Digiplante une ambiance agréable et chaleureuse. Cette belle aventure n'aurait pas été possible sans toi.

Je remercie également Philippe de Reffye et Marc Lavielle, pour avoir accepté de rapporter ma thèse, pour leurs remarques et conseils avisés qui m'ont permis d'améliorer le manuscrit et de me poser encore plus de questions (c'est cela la recherche !). Merci également à Anne Goelzer, Jean-Louis Foulley et Estelle Kuhn pour avoir accepté de participer à mon jury de thèse, pour l'intérêt que vous avez porté à mon travail, pour vos remarques et commentaires, et pour les échanges que nous avons pu avoir.

Merci à Samis, pour ses multiples relectures minutieuses (à la virgule près !), et pour les échanges scientifiques que nous avons pu avoir sur la dernière partie de ma thèse. Cela n'a pas toujours été facile, mais même avec un ordinateur qui a bu une tasse de thé et à distance, on peut parfois faire des miracles ! Merci pour tes conseils et tes encouragements.

Un grand merci également à Lionel Gabet, Erick Herbin et Gilles Faÿ, pour m'avoir permis de faire mes premiers pas de chargée de TD. Cette expérience a été révélatrice pour moi. J'espère avoir l'occasion à l'avenir de continuer dans cette voie. La passion que l'on peut développer pour une matière naît souvent de celle qu'ont pu nous transmettre nos enseignants ; à ce titre, j'aimerais remercier, même si elle ne lira sûrement pas ces lignes, mon professeur de mathématiques du lycée, qui m'a donné envie de poursuivre dans cette voie.

Merci à Sylvie pour ta bonne humeur, tes photos de la semaine, et aussi pour tous les petits sachets de thé qui remplacent bien le café ! Merci à Annie pour les conversations que nous avons pu avoir sur un peu tout et rien, et pour ton avis éclairé sur les méandres administratifs de l'École ... Enfin, un immense merci à Mélanie, pour sa bonne humeur matinale, et pour avoir pris soin de mes plantes quand je n'étais pas là !

Merci aussi à tous les doctorants ou non-doctorants du labo MAS, Alexandre, Benjamin, Benoît, Blandine, Gautier, Marion, Paul, PA, pour avoir parfois délaissé le baby-foot pour une partie de tarot, et pour

avoir rendu la vie quotidienne au labo plus agréable ! Merci à Marion pour m'avoir fait découvrir, entre autres, les joies des expérimentations, et pour les multiples échanges sur les noms de variables improbables, ainsi que pour tous les petits précieux conseils en C++. Ma vie chez Digiplante a un peu changé du jour au lendemain quand tu as commencé ton stage ! Merci !! Merci à Benoît sans qui je serais sans doute encore en train de débugger la version 1 de mon code. Tu as été d'une patience à toute épreuve face à mes questions parfois triviales. Avant de venir ici je ne savais même pas ce qu'était un template, et maintenant je sais même faire du OpenMP ! (bon, version basique, mais quand même !). Merci à Véronique pour ces multiples conseils, pour m'avoir permis d'encadrer des projets Enjeu passionnants, et grâce à qui je me suis mise à l'escalade ! Merci aux membres de l'équipe Digiplante, passés et présents, Cédric, Claire, Corina, Fenni, Qiongli, Robert, Yuting, Zhongping, pour les bons moments passés dans le bureau des doctorants, à dessiner des planisphères, coller des cartes postales, ou partager les spécialités culinaires des uns et des autres, et grâce à qui je garde de si bons souvenirs de mon passage chez Digiplante.

Merci également à Sébastien Lemaire et surtout à Anne Didier, sans qui cette thèse n'aurait pas été possible ... et oui, sans données expérimentales, les statistiques ne sont pas grand chose. Pour avoir un peu expérimenté une journée type de prélèvements en champ, je mesure à quel point la charge de travail est énorme. Merci d'avoir su répondre à toutes mes questions sur les données parfois en un temps record, et pour avoir rendu ce travail possible.

Last but not least, j'aimerais remercier enfin tous les membres de ma famille. Tout d'abord mes parents, bien sûr ... Merci d'avoir toujours été là pour moi, de m'avoir offert le meilleur, d'avoir cru en moi, de m'avoir supportée dans mes choix, quels qu'ils soient, en me laissant toute la liberté dont j'avais besoin. Merci d'avoir été là pour moi dans les moments difficiles, de m'avoir (r)ouvert votre porte. Votre soutien sans faille m'a aidé à contourner chaque difficulté et sans vous, je n'en serais pas là où j'en suis aujourd'hui. Et merci aussi pour ce petit (gros ?) grain de folie qui vous anime, et qui rend les réunions familiales particulièrement hautes en couleurs ! Maman, merci de m'avoir transmis (entre autres) ton goût pour la science, même si je n'ai pas tout à fait choisi la même voie. Papa, tu as su faire face à toutes les difficultés de la vie sans te décourager, même si cela a sans douté été très difficile parfois ... ta force et ton courage sont un exemple pour moi.

Merci à toi Shirley, grande sœur, cousine, meilleure amie, confidente ... comment te définir ?! un merveilleux mélange de tout cela, quelqu'un sur qui l'on peut compter, qui a su avoir les mots justes au bon moment, toujours souriante, de bonne humeur, avec un sacré caractère qui te va si bien ... tu es sous bien des aspects un modèle pour moi, et pour toutes ces années et celles qui viennent, je te remercie du fond du cœur ...

Cam, tu es passée avant moi par cet exercice, et tu sais à quel point il est difficile pour moi de trouver les mots justes pour te remercier. Merci *petite* sœur, d'avoir toujours été là pour moi, quoiqu'il arrive, en toutes circonstances. Merci d'avoir, par ordre chronologique et de façon non exhaustive : accepté de me prêter tes jouets, (...), discuté des nuits entières de tout et de rien, (...) été là pour partager mes états d'âme d'adolescente, (...), eu les petites et grandes attentions qui pouvaient manquer par ailleurs, (...), su m'écouter et me conseiller, (...) accepté mes choix quels qu'ils soient sans me juger, (...) été là pour partager pleins de bons moments, week-ends, vacances (dans l'Outback abandonné ... awesome !), soirées, mais aussi d'avoir été là quand cela n'allait pas, alors que tu étais toi même en période de stress pré-soutenance ... merci enfin d'être également complètement délurée, une vraie tête brûlée, en somme ... ☺

Enfin, merci à toi, Benjamin, pour tout ce que tu as fait pour moi au cours de ces derniers mois. Merci pour les merveilleux moments que nous avons partagés et qui sont, je l'espère, le début d'une longue série ;

pour tes attentions quotidiennes, même dans les moments difficiles ; pour avoir été là pour moi à toute heure du jour et de la nuit ; pour avoir su trouver les mots justes pour m'apaiser lors de mes accès de colères intempestifs ou mes périodes de stress ; pour avoir même tout fait pour me faciliter les choses et me permettre de finir ma rédaction sereinement. Tu es l'épaule sur laquelle je peux me reposer, le soutien sur lequel je peux compter. J'espère que je serais à la hauteur quand viendra ton tour de rédiger. Sans toi, je ne sais pas comment j'aurai vécu ces dix-huit derniers mois de thèse : pour ceci et pour tout le reste, merci ...

Table des matières

Introduction

"Les plantes semblent avoir été semées avec profusion sur la terre, comme les étoiles dans le ciel, pour inviter l'homme par l'attrait du plaisir et de la curiosité à l'étude de la nature."

Jean-Jacques Rousseau, *Rêveries du promeneur solitaire*

Les plantes jouent un rôle essentiel dans l'équilibre fragile de notre planète. Sources d'oxygène, d'énergie, avec l'émergence des agro-carburants, à la base de nos régimes alimentaires, utilisées dans la construction ou dans l'habillement, et même à l'origine de nombreux médicaments, nous dépendons entièrement du monde végétal pour notre survie. C'est pourquoi nous exerçons parfois de fortes pressions sur notre environnement afin d'optimiser le rendement de certaines cultures, en ayant recours par exemple à l'irrigation intensive, aux engrais, ou aux pesticides. Or, si ces apports peuvent parfois s'avérer utiles voire nécessaires, leur impact sur l'écosystème, à petite et à grande échelle, n'est pas toujours bien maitrisé.

Face à ces différentes problématiques, et dans un contexte mondial de changement climatique, la diminution de l'empreinte écologique de l'agriculture est l'un des enjeux majeurs du XXI^{ème} siècle. Et, à cette fin, les mathématiques apparaissent comme un outil essentiel nous permettant de mieux appréhender les phénomènes mis en jeu. C'est ainsi que la modélisation de la croissance des plantes a vu le jour à la fin du XX^{ème} siècle, à l'intersection de trois disciplines : la botanique, l'agronomie et l'informatique (de Reffye et al., 2009b).

Nous présentons tout d'abord quelques pré-requis sur le développement et le fonctionnement d'une plante, puis un bref historique des modèles de croissance de plantes. Nous présenterons ensuite les différentes problématiques auxquelles nous nous sommes intéressés, et les axes de développement que nous avons suivis.

1 Développement et fonctionnement d'une plante

Les plantes sont les seuls êtres vivants, avec les algues et certaines bactéries, à pouvoir fabriquer des composés organiques complexes (des glucides) à partir de dioxyde de carbone, d'eau et de sels minéraux, et avec la seule aide de l'énergie lumineuse. Tous les éléments de la plante participent à ce processus de photosynthèse : les racines puisent dans le sol l'eau et les sels minéraux nécessaires, et les feuilles captent l'énergie lumineuse (grâce à leurs cellules chlorophylliennes) et le dioxyde de carbone (grâce aux stomates).

1.1 Élements de morphogenèse végétale

La morphogenèse (du grec *morphê* - forme, et *genesis* - naissance) correspond à l'ensemble des mécanismes qui participent à l'édification d'un organisme vivant. Chez les plantes, elle démarre avec la germi-

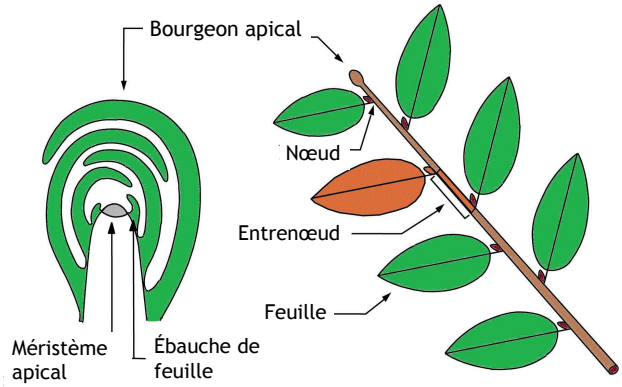


Fig. 1 – Schéma de la structure d'une plante (d'après Barthélémy et Caraglio (2007)). Le métamère (en orange sur la figure de droite) est composé d'un entrenœud, d'une feuille et d'un bourgeon axillaire.

nation de la graine et s'arrête à la mort de la plante, et dépend à la fois des caractéristiques génétiques de la plante, et de son environnement.

Lorsque les conditions nécessaires à la germination de la graine se trouvent réunies, les premiers organes de la plante peuvent commencer à se développer. Les éléments nécessaires à son développement pré-existent déjà au stade embryonnaire : le méristème racinaire, le méristème caulinaire, responsable du développement de la partie aérienne, et une première ébauche foliaire (ce sont les cotylédons). Les méristèmes sont des tissus constitués d'amas de cellules non différenciées qui peuvent se diviser un grand nombre de fois, pour permettre à la plante d'élaborer de nouveaux organes ou d'allonger des organes existants. On distingue les méristèmes apicaux, situés aux extrémités des tiges et des racines, et les méristèmes axillaires, situés au niveau des bourgeons axillaires, à l'aisselle des feuilles (voir Figure 1).

Au cours du processus d'organogenèse (création de nouveaux organes), les méristèmes apicaux et axillaires de la partie aérienne (méristèmes caulinaires) créent de nouvelles briques élémentaires qui, mises bout à bout, constituent l'architecture de la plante. Ces unités élémentaires, aussi appelées métamères ou phytomères, sont composées d'un *nœud*, auquel sont associés une ou plusieurs feuilles et un méristème, de l'*entrenœud* sous-jacent et d'un bourgeon axillaire situé à la base du nœud, à l'aisselle des feuilles.

Dans cette thèse, nous nous intéresserons en particulier à la betterave sucrière et au colza, dont la culture représente un enjeu économique majeur en France et dans le monde. Nous présentons ces deux plantes de façon un peu plus détaillée dans les deux paragraphes suivants.

La betterave sucrière

Environ 20% de la production mondiale de sucre provient de la betterave sucrière (*Beta vulgaris*), et la France est le premier producteur mondial de sucre de betterave (environ 4,5 millions de tonnes en 2012-2013). Une partie seulement de cette production est destinée à l'industrie agro-alimentaire, et le surplus de production sert entre autres à produire du bioéthanol dont la France est le premier producteur au niveau européen. À l'heure actuelle, 38% du bioéthanol produit en France provient des surplus de production de la betterave sucrière. Les enjeux économiques liés à la culture de la betterave sucrière sont donc majeurs.

Fig. 2 – Betterave sucrière

La betterave est une plante dicotylédone, semée en général au mois de mars ou d'avril, et dont on récolte la racine charnue entre septembre et novembre (voir Figure 2). En réalité, le cycle de croissance de la betterave sucrière est bisannuel : la première année correspond à la phase végétative, pendant laquelle la betterave développe son bouquet foliaire et accumule du sucre dans sa racine, et la deuxième année correspond à la phase reproductive, au cours de laquelle la plante puise dans les réserves de sa racine pour former une inflorescence. Cependant, la betterave étant récoltée pour le sucre contenu dans sa racine, elle n'est cultivée que sur une année (sauf dans le but particulier de produire des semences).

La morphologie de la betterave sucrière est assez simple, puisqu'au cours de la phase végétative, seuls trois types d'organes sont présents sur la plante : la racine (aussi appelée « pivot »), les pétioles et les limbes. Le collet, qui relie le pivot au bouquet foliaire, contient également du sucre, mais celui-ci est plus difficilement extractible. Il est utilisé, de même que le bouquet foliaire, pour l'alimentation du bétail. En pratique donc, le rendement de l'agriculteur se mesure en tonnes de racine par hectare.

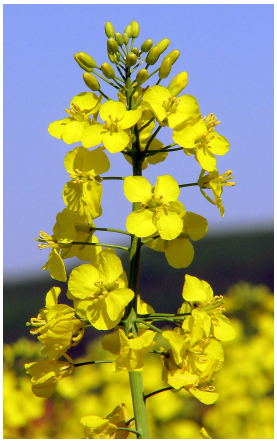


Fig. 3 – Colza

Colza

Le colza (*Brassica napus*) est principalement cultivé pour l'huile contenue dans ses graines, qui peut être utilisée pour l'alimentation humaine ou, à l'instar de la betterave, comme bio-carburant pour les moteurs Diesel. La France est le cinquième pays mondial producteur d'huile de colza, et près de 80% de l'huile récoltée sert à produire du bio-carburant. Le colza peut également être utilisé comme culture de couverture en hiver, et permettre ainsi de retenir une partie de l'azote contenue dans le sol, avant d'être utilisé comme engrais pour la culture suivante.

Toute comme la betterave, le colza est une plante dicotylédone, généralement semée à la fin de l'été au mois de septembre, et dont on récolte les graines au début de l'été suivant, après fécondation des ovules des fleurs. Cinq principaux stades peuvent être distingués dans la croissance du colza : le stade rosette, la montaison, la ramification, la floraison et le remplissage des graines. Jusqu'au début de l'hiver, la plante développe un bouquet foliaire en rosette avant d'entrer en repos hivernal. À la reprise de la végétation, au printemps suivant, la plante entre en phase de montaison, pendant laquelle la tige s'allonge pour atteindre sa taille maximale. La floraison intervient quelques semaines après, puis les fleurs fécondées se transforment en petites gousses contenant les graines.

Dans cette thèse, nous nous intéresserons uniquement au stade rosette, pendant lequel la morphologie de la plante s'apparente donc à celle de la betterave, car elle ne contient que des feuilles et une racine pivotante.

1.2 Fonctionnement

Les plantes sont des organismes autotrophes, c'est-à-dire capables de transformer de la matière minérale en matière organique, grâce au processus de *photosynthèse*. Ce mécanisme a été définitivement identifié au début du XX^{ème}, mais l'homme s'est de tout temps intéressé à la façon dont les plantes se « nourrissent ».

Depuis Aristote, on pense en effet que l'essentiel de la « nourriture » de la plante provient du sol, et ce n'est qu'au XVII^{ème}, avec les travaux du médecin belge Jan Baptista van Helmont, que cette certitude commence à s'ébranler. En observant la croissance d'un saule poussant dans un pot, celui-ci observe que l'arbre a pris environ 75 kg en 5 ans, quand la terre contenue dans le pot n'a perdu que quelques grammes

sur la même période. Le sol ne serait donc pas le principal fournisseur de matière pour la plante, et van Helmont fait alors l'hypothèse que cette matière provient de l'eau d'arrosage. À la fin du XVIII^{ème} siècle, Joseph Priestley met en évidence la production d'oxygène par les plantes, à travers diverses expériences sur la combustion et la respiration. Quelques années plus tard, le médecin hollandais Jan Ingen-Housz affine les résultats de Priestley et montre que cette production d'oxygène ne peut se faire qu'en présence de lumière. Puis, Jean Sénebier, en s'appuyant sur les travaux d'Antoine Laurent de Lavoisier, montre que les plantes absorbent du dioxyde de carbone et rejettent de l'oxygène. Enfin, c'est au suisse Nicolas Théodore de Saussure que revient la découverte, au début du XIX^{ème}, du lien définitif entre air, eau et lumière au cours de ce processus, en montrant que la plante consomme également, en plus du dioxyde de carbone atmosphérique, l'eau contenue dans le sol (King, 2004).

Chez les plantes, la lumière du soleil est absorbée par des pigments, dont le plus important est la *chlorophylle*, située dans les chloroplastes des cellules végétales, et responsable de leur couleur verte. Une première phase *photochimique* permet la transformation de l'énergie lumineuse en énergie chimique, et la transformation des molécules d'eau contenues dans les chloroplastes en molécules d'hydrogène et de dioxygène. Puis, au cours d'une deuxième phase appelée *cycle de Calvin*, l'énergie chimique obtenue précédemment est utilisée pour fabriquer du glucose, à partir du dioxyde de carbone atmosphérique absorbé par la plante et de l'hydrogène produit lors des réactions photochimiques. Les sucres ainsi formés vont participer à la fois au fonctionnement de la plante en lui fournissant de l'énergie, et lui permettre de fabriquer de la cellulose, le matériau dont elles sont principalement constituées. L'oxygène est quant à lui rejeté dans l'atmosphère par l'intermédiaire des stomates.

Plusieurs éléments sont donc nécessaires à la croissance de la plante. Parmi eux, on retrouve la lumière, l'eau, le dioxyde de carbone, mais également la température, qui joue un rôle sur l'ouverture ou la fermeture des stomates et régule ainsi les échanges gazeux, l'azote, qui permet à la plante de construire les acides aminés nécessaires à l'élaboration des protéines, et d'autres minéraux comme le potassium, qui favorise notamment le transfert des assimilats vers les organes de réserve (il est donc particulièrement important chez la betterave, par exemple) ou le phosphore qui joue un rôle dans la photosynthèse.

Chaque feuille [1] participe ainsi à la production de biomasse, qui sera ensuite distribuée à chaque organe en expansion ou nouvellement créé, via l'activité des méristèmes (voir section 1.1). Dans la suite, nous regrouperons donc sous le terme de *fonctionnement* l'ensemble des mécanismes de production de biomasse par photosynthèse, et d'allocation de biomasse aux différents organes de la plante.

2 Bref historique des modèles de croissance de plantes

Très tôt, l'importance capitale des plantes a poussé l'homme à étudier leurs caractéristiques, d'abord d'un point de vue botanique, depuis Aristote et l'un de ses disciples Théophraste, à qui l'on doit le plus ancien traité de botanique, puis agronomique, notamment grâce à Olivier de Serres et son ouvrage pionnier *Théâtre d'agriculture* publié en 1600.

Les premiers modèles de croissance de plantes sont, eux, beaucoup plus récents, et remontent au début des années 1970. Depuis, ils n'ont cessé de gagner en précision et complexité grâce aux progrès constants de l'informatique. Nous présentons ici un bref aperçu de l'histoire des modèles de croissance de plantes, avec d'abord les modèles géométriques d'un côté et agronomiques de l'autre, puis l'approche récente consistant à combiner ces deux types d'approches dans ce que l'on appelle les modèles structure-fonction.

1. Les feuilles sont les principaux organes photosynthétiques, même si on observe de façon marginale une activité photosynthétique dans les tiges

D'un point de vue botanique, la structure modulaire des plantes sous la forme d'une succession de métamères a permis l'émergence de modèles architecturaux, dont l'objectif est de classer les végétaux en fonction de leurs modes de développement. Dans les années 1970, les botanistes Hallé et Oldeman proposent notamment un système de classification permettant de répartir toutes les espèces d'arbre connues en 23 catégories, en fonction de leur mode de croissance, de ramification, de la différentiation morphologique ou de la position des organes reproducteurs (Hallé et Oldeman, 1970 ; Hallé et al., 1978). Puis, dans les années 1980, l'avènement de l'informatique a permis le développement de modèles entièrement basés sur la simulation. Parmi eux, on retrouve les L-systèmes, introduits par Lindenmeyer à la fin des années 1960, d'abord pour décrire la croissance d'organismes multicellulaires (Lindenmayer, 1968), puis appliqués plus tard à la croissance des plantes dans Prusinkiewicz et al. (1988) (voir aussi l'ouvrage de référence *The Algorithmic Beauty of Plants* (Prusinkiewicz et Lindenmayer, 1990)). Plusieurs extensions ont ensuite été proposées, notamment les L-systèmes stochastiques ou les grammaires relationnelles (Kurth, 1994), et de nombreux logiciels basés sur ces grammaires formelles ont vu le jour depuis les années 1990 : L-studio (Federl et Prusinkiewicz, 1999 ; Karwowski et Prusinkiewicz, 2004), GroIMP (Kniemeyer et al., 2007), … Une autre approche est celle développée au CIRAD [2] dans les modèles AMAP (de Reffye et al., 1988 ; de Reffye et Dinouard, 1990 ; de Reffye et al., 1991). Cependant, si ces modèles permettent d'obtenir une représentation fidèle de l'architecture de la plante, ils ne permettent pas de prendre en compte l'interaction avec le fonctionnement (Vos et al., 2007) [3].

Parallèlement au développement de ces modèles architecturaux, des modèles agronomiques ou « process-based » ont émergé, avec pour objectif de quantifier la production végétale au niveau du mètre carré, en fonction des conditions environnementales. Dans ce type de modèle, l'architecture de la plante n'est pas prise en compte, celle-ci étant simplement divisée en plusieurs compartiments d'organes (feuilles, tiges, racines, fruits, …). La production de biomasse s'obtient ensuite grâce à un système d'équations mettant en jeu les processus biologiques de photosynthèse, respiration, allocation, … En général, la quantité de rayonnement reçue par la plante y est modélisée par l'intermédiaire de la loi de Beer-Lambert (de Wit et al., 1970 ; Monteith, 1977), empruntée à la physique optique, et qui permet de relier la quantité de lumière absorbée à l'épaisseur du milieu traversé. Ces modèles peuvent être spécifiques à une espèce donnée (par exemple, CERES-MAÏZE (Jones et Kiniry, 1986)), ou génériques (PILOTE, (Mailhol et al., 1996)), et peuvent prendre en compte un grand nombre de processus écophysiologiques (STICS, (Brisson et al., 1998)). S'ils permettent en général une bonne estimation du rendement des cultures, plusieurs auteurs (Le Roux et al., 2001 ; Kurth, 1994) ont montré qu'une prise en compte de l'architecture permettrait d'augmenter leurs performances, à cause de l'interaction forte qui existe entre la structure de la plante et son fonctionnement.

C'est au carrefour de ces deux pratiques que sont nés les modèles de type structure-fonction, avec la première conférence internationale *Funtional-Structural Plant Growth Models* sur le sujet en 1996 (Korpilaht, 1997) (voir aussi (Sievänen et al., 2000) pour une revue détaillée). Ces modèles permettent de combiner la description du développement de la structure de la plante au cours du temps, et les processus éco-physiologiques mis en jeu (photosynthèse, respiration, allocation), eux-mêmes dépendants des conditions environnementales. Deux approches sont alors possibles pour construire ce type de modèles : soit étendre les modèles architecturaux en y ajoutant le fonctionnement de la plante, soit raffiner les modèles agronomiques afin de prendre en compte l'architecture de la plante. La première approche a mené par

2. Centre de coopération Internationale en Recherche Agronomique pour le Développement

3. Notons toutefois que les modèles AMAP ont évolué au cours du temps, pour prendre en compte l'interaction entre développement et fonctionnement. Nous faisons ici référence aux premières versions qui prenaient seulement en compte l'architecture.

exemple à la création du langage L+C (Karwowski et Prusinkiewicz, 2003), basé sur une extension des L-systèmes, ou encore à la création du modèle Greenlab, initié au LIAMA par de Reffye et Hu (2003) comme une suite logique des modèles AMAP. Basé sur un pas de temps discret (le cycle de croissance), le modèle permet de déterminer à chaque cycle le nombre d'organes créés, puis en déduit la production de biomasse par photosynthèse et l'allocation de cette biomasse aux organes existants ou nouvellement créés. La deuxième approche a abouti par exemple à la construction du modèle LIGNUM (Perttunen et al., 1996), dont le fonctionnement est proche de celui du modèle Greenlab.

3 Problématiques

Après ce premier élan qui a donné naissance à un grand nombre de modèles, un deuxième courant a vu le jour lorsque l'on a commencé à vouloir appliquer ces modèles et les confronter à des données expérimentales. Il est en effet nécessaire pour cela d'avoir recours à des outils mathématiques et statistiques qui requièrent en général une formulation mathématique rigoureuse des modèles. Beaucoup de progrès ont alors été faits en ce sens. Dans leur ouvrage, Wallach et al. (2006) proposent notamment plusieurs axes d'étude, dont l'analyse de sensibilité, l'estimation paramétrique, et l'évaluation des modèles.

L'analyse de sensibilité permet de déterminer de quelles façons les entrées d'un modèle peuvent influencer les sorties de ce modèle (Saltelli et al., 2004), et est utilisée en particulier pour identifier les paramètres les plus influents d'un modèle. Dans le contexte des modèles de croissance de plantes, ces approches ont déjà été appliquées avec succès : Ruget et al. (2002) pour le modèle STICS, Colbach et al. (2004) pour le modèle GeneSys, Garnier (2006) pour le modèle PASTIS, Bertheloot et al. (2011) pour le modèle NEMA, … De récents développements dans le cas des modèles de type structure-fonction ont été proposés dans les travaux de thèse de Wu (2012). Une extension de ces modèles dans le cas particulier où les paramètres sont corrélés, ce qui est souvent le cas dans les modèles considérés, est actuellement en cours de développement par Wu et al. (2013).

Dans le cadre de l'estimation paramétrique, au-delà du cadre général présenté dans Makowski et al. (2006), plusieurs auteurs ont proposé des méthodes d'estimation adaptées aux types d'observations et aux modèles utilisés. On peut citer notamment Zhan et al. (2003) et Guo et al. (2006) dans le cas du modèle Greenlab, de Reffye et al. (1999) pour la calibration de modèles hydrauliques, ou Hillier et al. (2005) pour les modèles de croissance d'organes. Plus récemment, Cournède et al. (2011) ont proposé une revue des méthodes d'estimation utilisables dans les modèles de type structure-fonction, en présentant une adaptation de l'estimateur d'Aitken, et en ouvrant la voie aux méthodes basées sur la théorie des modèles de Markov cachés, appliquées plus tard par Trevezas et Cournède (2013) dans le cas du modèle Greenlab. Des approches bayésiennes sont également possibles, et l'on peut citer notamment Makowski et al. (2002), ou Gaucherel et al. (2008) qui comparent deux méthodes d'estimation Bayésiennes (filtrage particulaire ou simulations de type Monte Carlo par chaîne de Markov) à une approche fréquentiste basée sur une méthode d'optimisation, ou les travaux de thèse de Yuting Chen sur les méthodes de filtrages particulaires et d'assimilation de données (Chen et al., 2013b).

Cependant, un certain nombre de questions restent en suspens. Par exemple, nous avons évoqué plus haut les différents types de modèles existants, chacun ayant ses propres objectifs, ses propres spécificités. Dans un tel contexte, le modélisateur se retrouve souvent confronté à un choix qui peut s'avérer délicat : quel modèle choisir parmi ceux existants ? comment le choisir ? Un modèle mécaniste visant à décrire des processus complexes pourra s'avérer peu robuste comme outil de prévision, car trop complexe, et inverse-

ment, un modèle prédictif peut échouer à décrire certains phénomènes fins impliqués dans le processus de croissance de la plante, et ainsi s'avérer moins performant dans des conditions non optimales de croissance, par exemple. Ainsi, si l'approche structure-fonction apparaît comme une amélioration des modèles agronomiques, avec une meilleure prise en compte des processus impliqués dans la croissance de la plante, ce type de modèle est souvent plus complexe à calibrer, et l'effort consenti pour obtenir des jeux de données expérimentales satisfaisants peut ne pas être récompensé par les performances des modèles. Il paraît alors indispensable de comparer les différents modèles candidats, au moyen de critères appropriés. En pratique cependant, chaque modèle est défini dans un contexte précis, et pour un objectif particulier. Ils peuvent donc être difficilement comparables *a priori*. Peu d'études portent d'ailleurs sur la comparaison de modèles de croissance de plantes et l'évaluation de leurs performances en accord avec un objectif donné. En particulier, aucun « benchmark » n'est disponible, qui permettrait d'obtenir une sorte de performance de référence pour de tels modèles.

Dans ce contexte, nous proposons dans cette thèse une méthodologie s'inscrivant dans les « bonnes pratiques » de modélisation pour les modèles de croissance de plantes, spécifiées par Vos et al. (2007), dans le but de construire et d'évaluer des modèles utilisés comme outils prédictifs. Cinq modèles de croissance de plantes génériques, ou dont les concepts peuvent facilement s'étendre à d'autres plantes, ont été étudiés dans le cas de la betterave sucrière, l'objectif étant de prédire le rendement de la culture et la biomasse totale de la plante. Ces modèles n'ayant pas nécessairement été construits comme des outils de prévision, une première étape consiste à élaborer pour chacun d'eux une version plus robuste, à l'aide d'une analyse de sensibilité dont l'objectif est de déterminer les paramètres les plus influents du modèle. Ces paramètres seront ensuite ajustés sur un premier jeu de données, les autres paramètres étant fixés à des valeurs de référence disponibles dans la littérature. La deuxième étape correspond à la comparaison à proprement dite des capacités prédictives des modèles ainsi obtenus. Nous obtenons ainsi une première comparaison des performances de ces modèles, en fonction de leurs caractéristiques et de leur niveau de prise en compte, et nous pouvons proposer un premier « benchmark » dans le cas de la betterave sucrière.

Un autre point crucial concerne l'estimation paramétrique, ou plus exactement, la façon dont sont actuellement prises en compte les observations expérimentales. Il existe en effet une forte variabilité entre plantes, due en partie à la variabilité génétique, mais également aux micro-variations dans les conditions environnementales, même au sein d'une même parcelle agricole. Souvent, le protocole expérimental prend en compte cette variabilité, et prévoit de mesurer, à chaque date d'observation, plusieurs plantes choisies aléatoirement dans le champ et censées représenter cette variabilité. Cependant, au moment de l'estimation, ces mesures indépendantes sont souvent résumées en une plante « moyenne » correspondant aux moyennes, à chaque date d'observation, des plantes mesurées (voir par exemple Guo et al. (2006), Lemaire et al. (2008), Letort (2008), Bertheloot et al. (2008), Jullien et al. (2011), Cournède et al. (2011)). Outre la perte d'information qu'implique l'utilisation de cette plante moyenne, il n'est pas possible, à partir de ces données résumées, d'estimer la variabilité inter-plantes. Or, cette variabilité est d'une importance capitale. Elle permet notamment aux plantes de mieux s'adapter aux conditions environnementales, de mieux résister aux attaques des insectes ou aux maladies. En étudiant le rendement agricole d'exploitations situées en Afrique de l'Ouest, où le climat est semi-aride, Brouwer et al. (1993) ont montré que cette variabilité entre plantes, associée à une variabilité locale des conditions environnementales, avait pour conséquence de rendre certaines parties du champ plus adaptées à la sécheresse, et capables de compenser les mauvaises performances obtenues dans d'autres parties du champ. Dans ce cas précis, la variabilité inter-individuelle apporte à l'agriculteur la garantie d'un rendement minimum. De la même façon, Renno et Winkel (1996) ont montré que lorsque la période de floraison n'est pas synchronisée pour toutes les plantes d'un champ,

la culture sera moins susceptible de subir les effets d'un stress hydrique ponctuel, ou d'une attaque d'insectes. Ici encore, la variabilité inter-individuelle peut donc avoir des effets bénéfiques sur le rendement, par exemple dans le cas du colza, où l'on s'intéresse à la récolte des graines, elles-mêmes issues des fleurs. Cependant, cette variabilité peut également avoir des effets néfastes pour l'agriculteur, notamment dans le cas de la betterave sucrière, car celle-ci peut se traduire par une trop forte variabilité du calibre du pivot, ce qui peut entraîner notamment des problèmes logistiques. De même, certains agriculteurs cherchent à obtenir des fruits ou légumes dont le calibre sera plutôt homogène, afin de minimiser les problèmes logistiques (c'est le cas par exemple du poivron, ou du concombre). Plus généralement, dans une optique de prévision de rendement, ou d'analyse de risque, il peut être intéressant de fournir en sortie des modèles, non pas une valeur unique, comme c'est le cas des modèles actuels, mais plutôt une plage de valeurs.

Une approche basée sur des « différentielles statistiques » a notamment été proposée par de Reffye et al. (2009a) dans le cas de la betterave, et par Feng et al. (2014) dans le cas du maïs, à l'aide du modèle Greenlab. À partir d'un développement en série de Taylor du modèle de croissance de plante, il est possible de propager l'incertitude entourant certains des paramètres du modèle vers une ou plusieurs variables d'intérêt (par exemple, le rendement, ou la production de biomasse). Cette approche permet également d'estimer les moments de certains paramètres considérés comme aléatoires (typiquement, moyenne et variance) à l'aide de la variabilité observée dans la population. Si cela permet d'obtenir une première estimation de la variabilité de ces paramètres dans la population, plusieurs limites apparaissent. D'une part, l'approximation du modèle par un développement de Taylor peut s'avérer mauvaise, en particulier si le modèle est fortement non linéaire. D'autre part, même si la variance des observations est utilisée en plus de la moyenne pour calibrer le modèle, ce qui constitue sans aucun doute une amélioration par rapport aux méthodes basées uniquement sur une plante moyenne, il s'agit encore de données agrégées, et une partie de l'information est perdue par rapport au cas où les observations de toutes les plantes seraient utilisées. Enfin, si une première estimation de la variabilité des paramètres dans la population peut être obtenue avec cette méthode, il n'est en revanche pas possible de tester statistiquement si cette variabilité est significative.

L'approche que nous proposons dans cette thèse est basée sur l'utilisation de modèles à effets mixtes, qui permettent justement de prendre en compte à la fois la variabilité intra-individuelle, c'est-à-dire la façon dont varient les mesures d'un même individu (au cours du temps, par exemple), mais également la variabilité inter-individuelle. Dans une première étape, un même modèle est défini pour chaque individu, mais à l'aide de paramètres qui lui seront spécifiques. Puis, les paramètres individuels obtenus à l'étape précédente sont considérés comme des réalisations de variables aléatoires, dont on peut calculer la distribution dans la population. Cette approche a été appliquée dans un premier temps à l'organogenèse chez la betterave sucrière, puis au modèle Greenlab pour lequel nous avons proposé une version à l'échelle de la population.

PyGMAlion

Tous les résultats présentés dans le chapitre 1 et dans la section 2 du chapitre 3 ont été obtenus grâce à la plateforme de modélisation de l'équipe Digiplante du laboratoire MAS (Mathématiques Appliquées aux Systèmes) de l'École Centrale Paris, PyGMAlion (Plant Growth Models Analysis, Identification and Optimization), dans laquelle les modèles et les méthodes ont été implémentés. Cette plateforme est développée en C++ et est découpée en plusieurs sous-modules correspondant aux différentes méthodes implémentées (Cournède et al., 2013).

Un modèle est considéré sous PyGMAlion comme un système dynamique discret, c'est-à-dire qui peut être représenté sous la forme $X_{n+1} = F_n(X_n, U_n, P)$, où X_n représente les variables d'état du système

au temps n, U_n les variables environnementales, P le vecteur de paramètres et F_n représente les processus éco-physiologiques mis en jeu dans le modèle. On suppose ici que le processus est déterministe, mais il est également possible de considérer des bruits de modélisation.

Pour chaque modèle, il est donc nécessaire de spécifier les trois classes suivantes : une classe de paramètres, une classe d'environnement, et une classe de variables d'état, puis d'implémenter la fonction F_n. Ce cadre est générique et peut également être utilisé pour tout modèle pouvant s'écrire comme un système dynamique discret. On définit ensuite pour chaque modèle un ensemble d'observeurs, qui correspondent à l'observation de certaines variables d'état du système dynamique à un temps donné, à partir des variables environnementales et de conditions initiales. Une fois le modèle défini, on peut lui appliquer un certain nombre de méthodes : analyse de sensibilité, identification paramétrique, et évaluation.

4 Organisation du manuscrit et remarques préliminaires

Le manuscrit est organisé en trois chapitres :

1. Le chapitre 1 concerne l'évaluation de cinq différents modèles de croissance de plantes utilisés comme outils de prévision. Les cinq modèles comparés diffèrent par l'échelle de prise en compte, que ce soit au niveau de la plante individuelle ou du mètre carré, et par le processus de répartition de la biomasse, qu'il soit empirique, à l'aide d'un simple indice de récolte, ou basé sur une allocation dynamique de la biomasse produite en cours de croissance. Parmi ces cinq modèles, quatre sont des modèles agronomiques, tels que définis précédemment, et seul le modèle Greenlab appartient à la classe des modèles structure-fonction. Nous proposons dans ce chapitre une méthodologie permettant de réduire le nombre de paramètres à estimer, et ainsi d'augmenter la robustesse des modèles, en utilisant une analyse de sensibilité. Puis, nous comparons les capacités prédictives des modèles ainsi construits sur deux jeux de données indépendants.

2. Le chapitre 2 contient une présentation des modèles non linéaires mixtes qui seront ensuite appliqués dans le chapitre 3. Nous présentons la formulation générale du modèle, et les méthodes d'estimation existantes, en nous concentrant sur celles basées sur l'algorithme EM (Espérance-Maximisation). Plus spécifiquement, nous développons deux versions stochastiques de cet algorithme, avec d'une part l'algorithme MCMC-EM, et d'autre part l'algorithme SAEM.

3. Le chapitre 3 présente deux applications des méthodes présentées au chapitre 2, permettant de prendre en compte et d'estimer la variabilité entre plantes. Dans un premier temps (section 1) nous proposons un modèle permettant de modéliser la variabilité du processus d'organogenèse chez la betterave sucrière, puis nous proposons une version population du modèle individuel Greenlab (section 2). Dans le premier cas, l'estimation a été faite sous le logiciel Monolix (The Monolix Team, 2011) qui repose sur l'utilisation de l'algorithme SAEM, mais dans le deuxième cas, nous avons implémenté les deux algorithmes MCMC-EM et SAEM dans la plateforme de modélisation de l'équipe PyGMAlion, ce qui nous a permis en particulier de comparer les performances des deux algorithmes.

Remarques préliminaires

Le premier chapitre peut se lire indépendamment des deux autres, même si le modèle d'organogenèse développé au Chapitre 3 y est évoqué brièvement. En revanche, la lecture du Chapitre 2 est requise pour la compréhension du Chapitre 3.

Certains termes sont définis de façon plus précise dans le Glossaire situé en fin de manuscrit. Ils sont identifiés par une police et une couleur différentes, par exemple : métamère.

Chapitre 1

Sélection de modèles pour la prévision

"All models are wrong, but some are useful."

Georges E. P. Box, ***Empirical Model-Building and Response Surfaces.***

Un grand nombre de modèles de croissance de plantes ont été développés depuis les années 1970, chacun ayant ses propres spécificités et ses propres objectifs. Certains modèles sont notamment conçus pour prédire le rendement d'une culture, ou comme des outils d'aide à la décision, alors que d'autres s'attachent à décrire de façon précise les processus éco-physiologiques mis en jeu au cours du développement de la plante (Fourcaud et al., 2008).

En fonction de leurs objectifs, ces modèles n'auront pas tous les mêmes niveaux de modélisation (à l'échelle de l'organe, de la parcelle, ...), et pourront intégrer un nombre variable de mécanismes (effets des différents stress environnementaux, architecture, prise en compte des stratégies d'allocation, ...), augmentant ainsi la complexité du modèle. Celle-ci s'accompagne alors souvent d'une hausse du nombre de paramètres à estimer, et d'une plus grande variance du modèle. Or, dans un objectif de prévision, les modèles doivent être les plus robustes possible, et pouvoir s'adapter à des situations différentes de celles dans lesquelles ils ont été paramétrés. C'est pourquoi il convient de réaliser un compromis entre le biais et la variance d'un modèle qui sera utilisé comme outil de prévision.

Dans ce contexte, nous proposons une étude comparative des capacités prédictives de cinq modèles de croissance de plante pour la betterave sucrière : Greenlab (de Reffye et Hu, 2003 ; Yan et al., 2004 ; Lemaire et al., 2008), CERES (Jones et Kiniry, 1986 ; Leviel, 2000), Pilote (Mailhol et al., 1997 ; Taky, 2008), STICS (Brisson et al., 2003, 2008 ; Launay et Guérif, 2003) et un cinquième modèle nommé LNAS (Cournède et al., 2013). Une première comparaison des modèles Greenlab, CERES et Pilote est disponible dans les travaux de thèse de Lemaire (2010). Ces cinq modèles diffèrent d'une part sur l'échelle de prise en compte, qu'elle soit au niveau de l'organe (Greenlab ou CERES) ou au niveau de la plante (Pilote, STICS, LNAS), et d'autre part sur la stratégie adoptée pour la répartition de la biomasse produite aux différents organes de la plante. Dans les modèles CERES et Pilote, l'indice de surface foliaire évolue de façon empirique, indépendamment de la biomasse créée et disponible, et un indice de récolte permet de répartir la biomasse produite entre feuilles et racines à la fin de la période de croissance. Dans les modèles LNAS et Greenlab, la répartition de la biomasse se fait de manière dynamique tout au long de la période de croissance de la plante, à l'aide de fonctions d'allocations de biomasse. L'approche utilisée dans STICS peut être vue comme un intermédiaire entre les deux précédentes, car elle est basée sur une évolution empirique

Tab. 1.1 – Classement des cinq modèles selon l'échelle de description et la stratégie de répartition de la biomasse

	Échelle de modélisation	
	compartiment	**organe**
Répartition empirique	Pilote	CERES
Allocation de biomasse	STICS LNAS	Greenlab

de l'indice de surface foliaire couplée à un mécanisme de type source-puits induisant une rétroaction de l'allocation par un indice de stress trophique. Le tableau 1.1 résume les différences entre les cinq modèles.

L'objectif ici est de construire des modèles robustes qui puissent être utilisés comme outils prédictifs. C'est pourquoi, dans un premier temps, une analyse de sensibilité a été conduite sur chaque modèle afin d'identifier les paramètres les plus influents, qui ont été estimés sur un premier jeu de données d'apprentissage. Puis, les capacités prédictives des cinq modèles ont été comparées sur un jeu de données indépendant à l'aide de différents critères d'évaluation. Tous les résultats présentés ici ont été obtenus grâce à la plateforme de modélisation de l'équipe, PyGMAlion (Cournède et al., 2013), dans laquelle les modèles ont été implémentés, et qui dispose des différents outils dont nous avons eu besoin : estimation paramétrique, analyse de sensibilité, calcul de critères de prédiction. Les cinq modèles sont décrits en détails dans la section 1, puis la procédure de calibration est présentée en section 2, et les critères d'évaluation des qualités prédictives en section 3. Les résultats sont présentés en section 4.

1 Modèles

La plante produisant la biomasse nécessaire à sa croissance à travers le processus de la photosynthèse, c'est tout naturellement que les modèles de croissance de plantes cherchent à relier la production de biomasse à la quantité de rayonnement reçue par la plante. C'est cette approche énergétique de la production que met en avant Monteith (1977), en supposant que l'accumulation de biomasse au temps t est proportionnelle au cumul de rayonnement photosynthétiquement actif absorbé (PAR_a en $\mathrm{MJ.m}^{-2}$) par la culture. En considérant la production journalière de biomasse, cela donne :

$$q(t) = \mathrm{RUE} \cdot \mathrm{PAR}_a(t), \tag{1.1}$$

où RUE (Radiation Use Efficiency, en $\mathrm{g.MJ}^{-1}$) est l'efficience liée à la conversion en biomasse de l'énergie lumineuse.

Le rayonnement absorbé par le feuillage est ensuite supposé proportionnel au rayonnement reçu par la plante grâce à la loi de Beer-Lambert (Marcelis et al., 1998), en considérant le couvert végétal comme un milieu absorbant traversé par un rayonnement de longueur d'onde constante. Damay et Le Gouis (1993) donnent la formulation suivante dans le cas de la betterave sucrière :

$$\mathrm{PAR}_a(t) = 0.95 \cdot \mathrm{PAR}(t) \cdot \left(1 - \exp\left(-k_b\, \mathrm{LAI}(t)\right)\right), \tag{1.2}$$

avec $\mathrm{LAI}(t)$ l'indice de surface foliaire (LAI, pour Leaf Area Index, en anglais) et k_b le coefficient d'absorption de la loi de Beer-Lambert, le coefficient 0.95 correspondant à l'efficience maximale de l'interception lumineuse. Ainsi, l'absorption lumineuse augmente lorsque la surface foliaire augmente, mais

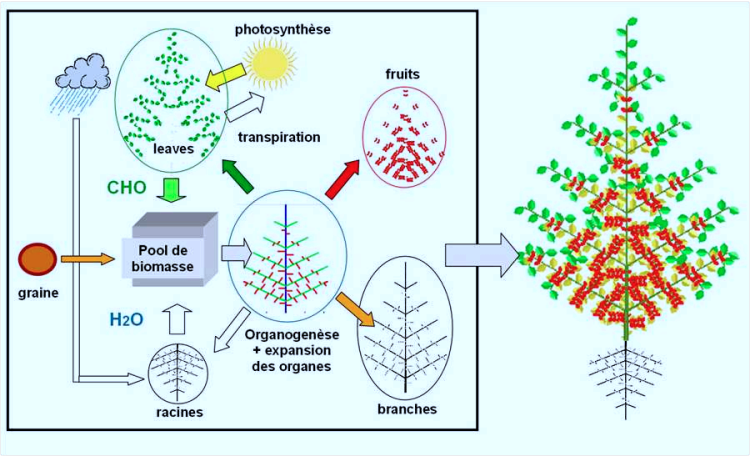


FIG. 1.1 – Croissance de la plante dans le modèle Greenlab

un effet de saturation apparaît lorsque les feuilles commencent à se superposer, dû à l'ombre que les feuilles situées au-dessus du couvert projettent sur leurs voisines situées plus près du sol.

On obtient finalement l'équation de production suivante, qui sert de base aux cinq modèles qui seront comparés dans ce chapitre :

$$q(t) = 0.95 \cdot \text{RUE} \cdot \text{PAR}(t) \cdot (1 - \exp(-k_b \, \text{LAI}(t))) \, . \tag{1.3}$$

Notons toutefois que cette équation de production est valable à l'échelle de la culture et non pas de la plante individuelle, l'indice de surface foliaire LAI étant défini comme le ratio entre la surface supérieure des feuilles vertes et la surface de sol sur laquelle se développe la culture (Watson, 1947). Si cela ne pose pas de problème pour les modèles ayant pour niveau de modélisation le mètre carré, quelques ajustements sont nécessaires dans le cas des modèles individus-centrés à l'échelle de l'organe Greenlab et CERES (voir sections 1.1 et 1.5 respectivement).

1.1 Greenlab

Le modèle Greenlab appartient à la famille des modèles structure-fonction (Sievänen et al., 2000 ; Vos et al., 2007), qui combinent la description de l'architecture de la plante et son fonctionnement écophysiologique (production et allocation de biomasse). C'est également un modèle générique, c'est-à-dire qu'il n'est pas spécifique à une espèce de plante donnée. Le schéma général du modèle est résumé sur la Figure 1.1. Après germination et émission des cotylédons par la graine, la plante absorbe l'eau contenue dans le sol par ses racines, et capte le dioxyde de carbone atmosphérique pour faire la photosynthèse. Elle va ainsi alimenter un 'pool' commun de biomasse, qui sera ensuite redistribué aux différents organes de la plante (feuilles, entrenœuds, branches, fruits, fleurs, racines, ...), soit pour en former de nouveaux (c'est l'organogenèse), soit pour permettre leur expansion.

Introduit par de Reffye et Hu (2003) comme une suite logique des modèles AMAP (de Reffye et al., 1997), le développement de la plante y est initialement déterministe et indépendant du fonctionnement. Plus précisément, deux plantes de même type produisent toujours le même nombre d'organes, la taille des organes étant par contre influencée par l'environnement. Ce modèle a été et continue d'être largement utilisé pour une grande variété de plantes : le maïs (Guo et al., 2006), la tomate (Dong et al., 2008), la betterave (Lemaire et al., 2008), ... D'autres versions ont depuis été proposées, intégrant une organogenèse stochastique pour la version 2 (Kang et al., 2008), ou une rétro-action du fonctionnement sur le développement pour la version 3 (Mathieu, 2006 ; Mathieu et al., 2009).

À l'origine, le modèle Greenlab s'écrit comme un système dynamique *discret*, dont le pas de temps correspond au cycle de croissance de la plante. La notion de cycle de croissance est intimement liée à celle de temps thermique, correspondant à l'accumulation de températures dépassant un certain seuil, qui sert alors de base au système. Il existe en effet une relation fortement linéaire entre le nombre de feuilles présentes sur la plante (visibles ou non) et le temps thermique. Cela correspond au cumul de températures nécessaire au méristème pour former un nouveau métamère. Le modèle Greenlab donne alors l'état du système au cycle de croissance n en fonction de son état au cycle $n-1$ et de variables exogènes (environnementales).

Cependant, malgré l'apparente légitimité du cycle de croissance comme échelle de temps, plusieurs difficultés apparaissent lorsqu'il s'agit de prendre en compte l'effet de l'environnement. Notamment, les processus écophysiologiques mis en jeu au cours de la croissance de la plante dépendent de conditions bioclimatiques (température, PAR, ...) qui varient de façon continue, et qui sont en général collectées quotidiennement. De même, la prise en compte des stress environnementaux, par exemple à l'aide de modèles de bilan hydrique, se fait également de façon continue. Il paraît donc intéressant de synchroniser ces phénomènes avec la croissance de la plante. De plus, les quatre autres modèles utilisés dans ce chapitre sont des modèles journaliers. Il nous est donc apparu nécessaire de proposer une version journalière du modèle Greenlab, correspondant à une discrétisation du modèle continu proposé par Li et al. (2009) avec un pas de temps journalier pour les fonctions de production et d'allocation. L'organogenèse reste quant à elle rythmée par le cycle de croissance architectural.

Comme précisé plus haut, le modèle Greenlab est un modèle individu-centré et de ce fait, quelques ajustements de l'équation (1.3) ont été faits. Tout d'abord, le coefficient d'efficience est adapté à la plante individuelle, et peut donc être différent de la RUE utilisée au niveau de la culture. Une surface foliaire spécifique s^{pr} exprimée en $\text{m}^2.\text{pl}^{-1}$ et pouvant s'interpréter comme la projection orthogonale de la plante sur le sol, est également introduite dans le modèle. Finalement, la production de biomasse d'une plante individuelle au jour t est donnée par la relation suivante :

$$q_{pl}(t) = 0.95 \cdot \mu \cdot s^{pr} \cdot \text{PAR}(t) \cdot \left(1 - \exp\left(-k_b \frac{s^{act}(t)}{s^{pr}}\right)\right), \tag{1.4}$$

où μ correspond à l'efficience « individuelle » (en g.pl^{-1}), et s^{act} à la surface foliaire photosynthétiquement active de la plante au début du jour t en m^2 (voir équation 1.15). La production de biomasse par mètre carré s'obtient ensuite en multipliant la production individuelle par la densité de population d :

$$q(t) = d \cdot q_{pl}(t), \tag{1.5}$$

et l'efficience au niveau du mètre carré peut s'approcher de la façon suivante : RUE $= \mu\ s^{pr}\ d$, où d est la densité de population.

La surface foliaire photosynthétiquement active se déduit de la masse des limbes des feuilles photosynthétiquement actives au jour t et de la masse surfacique des limbes, e_b. Afin de définir proprement cette quantité, nous avons d'abord besoin d'introduire un certain nombre de notions, concernant l'organogenèse et l'allocation.

1.1.1 Organogenèse

Définissons tout d'abord la notion de temps thermique, qui a été introduite brièvement au début de la section 1, et qui nous permettra de définir le cycle de croissance. À partir d'une température de base T_b,

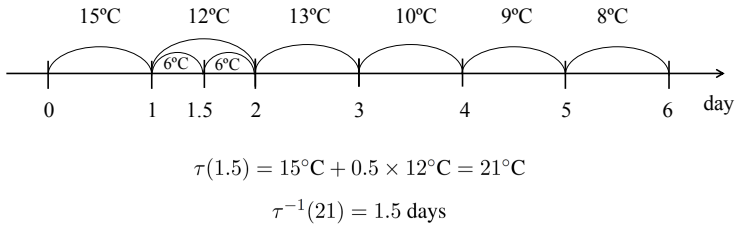


FIG. 1.2 – Calcul du temps thermique par interpolation linéaire.

le temps thermique au temps t correspond au cumul des températures ayant dépassé ce seuil :

$$\tau(t) = \int_0^t \max(0, T(s) - T_b)\, ds, \quad t \geq 0, \tag{1.6}$$

où $T(s)$ est la température au temps s. Comme nous disposons en général des températures moyennes journalières, nous utiliserons par la suite une interpolation linéaire du temps thermique (voir Figure 1.2). Nous définissons également l'inverse généralisée du temps thermique τ^{-1} par :

$$\tau^{-1}(u) = \inf\{t \in \mathbb{R} \mid \tau(t) \geq u\}, \quad u \geq 0.$$

Seuls trois types d'organes sont considérés dans le cas de la betterave sucrière (voir Introduction, section 1.1) : les limbes, les pétioles et la racine. Le cycle de croissance correspond alors simplement au phyllochrone, c'est-à-dire au temps thermique s'écoulant entre l'apparition de deux feuilles successives. Chaque feuille, et donc chaque limbe et chaque pétiole, est ainsi entièrement déterminée par son rang, correspondant au cycle de croissance au cours duquel elle a été initiée. Nous faisons alors les hypothèses suivantes : le limbe et le pétiole d'une même feuille sont initiés simultanément, ont le même temps d'expansion et la même durée de vie. Nous supposons également que la racine est initiée lors du premier cycle, qui correspond à la mise en place des cotylédons, et qu'elle ne tombera pas en sénescence au cours de la période d'observation. Les cotylédons apparaissant en même temps sur la plante, ils seront considérés par la suite comme une feuille unique dont la masse sera égale à la somme des masses des deux feuilles cotylédonaires.

Nous notons $\mathcal{O} = \{b, p, r\}$ l'ensemble des organes de la plante, où b = limbe, p = pétiole et r = racine, et pour une feuille de rang k, nous notons τ_k son temps thermique d'initiation, τ_k^e son temps thermique d'expansion, et τ_k^s sa durée de vie. Le temps thermique d'initiation de la racine (correspond au temps de germination) est alors égal à τ_1, et son expansion à τ_r^e.

Deux phases distinctes se succèdent dans le développement foliaire de la betterave sucrière, comme l'ont montré Milford et al. (1985b) et Lemaire et al. (2008), conduisant à la définition de deux phyllochrones. On observe en effet un ralentissement du rythme d'apparition des feuilles lorsque le couvert végétal devient plus dense, correspondant à une plus forte compétition pour la lumière (Lemaire et al., 2008 ; Lemaire, 2010). Nous notons alors τ^{rupt} le temps thermique de rupture correspondant au changement de phyllochrone, et γ_1 et γ_2 les phyllochrones de la première et de la seconde phase. En s'appuyant sur ces notations, nous pouvons définir le temps thermique d'initiation de la feuille de rang k de la façon suivante :

$$\tau_k = \begin{cases} \tau_1 + (k-1)\,\gamma_1 & \text{si } 1 \leq k \leq N^{rupt}, \\ \tau_1 + (N^{rupt} - 1)\,\gamma_1 + (k - N^{rupt})\,\gamma_2 & \text{si } k > N^{rupt}, \end{cases} \tag{1.7}$$

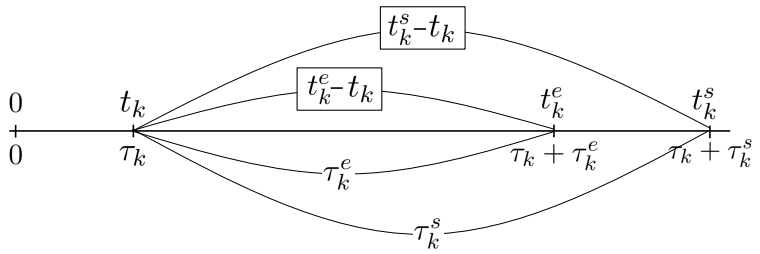


FIG. 1.3 – Correspondances entre le temps calendaire et le temps thermique d'initiation, d'expansion et de durée de vie de la feuille de rang k.

où N^{rupt} est le rang de la dernière feuille initiée avant le temps de rupture,

$$N^{rupt} = 1 + \left\lfloor \frac{\tau^{rupt} - \tau_1}{\gamma_1} \right\rfloor .$$

Le nombre de feuilles au temps t est donné par :

$$N_l(t) = \left(1 + \left\lfloor \frac{\tau(t) - \tau_1}{\gamma_1} \right\rfloor\right) \mathbf{1}_{\tau(t) \leq \tau^{rupt}} + \left(N^{rupt} + \left\lfloor \frac{\tau(t) - \tau_{N^{rupt}}}{\gamma_2} \right\rfloor\right) \mathbf{1}_{\tau(t) > \tau^{rupt}}.$$

Pour des raisons de concision, nous adoptons dans la suite du document les notations suivantes : $t_k := \tau^{-1}(\tau_k)$, $t_k^e := \tau^{-1}(\tau_k + \tau_k^e)$ et $t_k^s = \tau^{-1}(\tau_k + \tau_k^s)$, correspondant aux temps calendaires d'initiation, de fin d'expansion et de fin de vie de la feuille de rang k (voir Figure 1.3).

1.1.2 Allocation

La biomasse produite par la plante au temps t est distribuée à chaque organe de la plante par l'intermédiaire de relations d'allocation de type source-puits, indépendamment de sa position sur la plante. C'est l'hypothèse d'un pool commun de biomasse, auquel tous les organes en expansion peuvent s'alimenter. Les organes sources correspondent aux organes producteurs de biomasse (les feuilles), et les organes puits sont ceux qui consomment de la biomasse, c'est-à-dire tous les organes en expansion. Chaque organe puits possède une force d'attraction de la biomasse, et recevra chaque jour une quantité de biomasse proportionnelle à sa demande, qui est mesurée à l'aide d'une fonction puits. Lors du premier cycle de croissance, cependant, la plante ne possède pas encore de feuilles et ne peut donc pas réaliser la photosynthèse. Dans ce cas, la biomasse provient uniquement de la graine, et nous supposons que cette biomasse est distribuée uniformément au cours du cycle aux différents organes de la plante (racine et cotylédons). Lors des cycles suivants, nous supposons que la biomasse est produite uniquement par photosynthèse selon l'équation (1.3). Ainsi :

- entre 0 et τ_1, la graine n'a pas encore commencé à germer, rien ne se passe
- entre τ_1 et τ_2, la plante produit ses premières feuilles et racines à partir de la graine
- à partir de τ_2, la plante produit sa biomasse par photosynthèse.

Dans la suite, nous notons p_{al}^o le vecteur de paramètres d'allocation de l'organe o, et $p_{al} = (p_{al}^o)_{o \in \mathcal{O}}$ le vecteur contenant tous les paramètres d'allocation de tous les organes.

Fonctions puits

La fonction puits d'un organe o de rang k au temps u est donnée par la formulation suivante, en utilisant

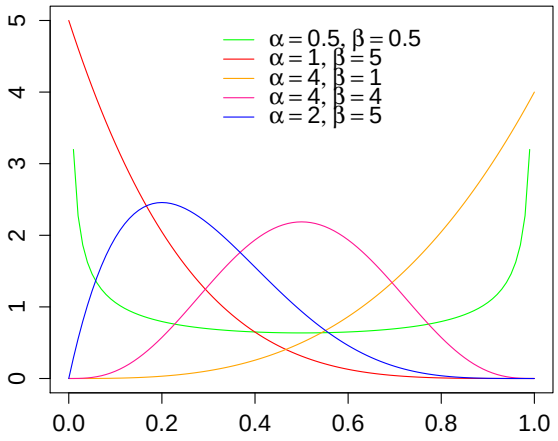


FIG. 1.4 – Densité de la loi bêta B(α, β) en fonction des valeurs des paramètres α et β

la convention $s_r = s_{r,1}$:

$$s_{o,k}(u; p^o_{al}) = c_o \, p_o(u) \left(\frac{\tau(u) - \tau_k}{\tau^e_k} \right)^{a_o - 1} \left(1 - \frac{\tau(u) - \tau_k}{\tau^e_k} \right)^{b_o - 1} \mathbf{1}_{\tau_k \leq \tau(u) \leq \tau_k + \tau^e_k}, \tag{1.8}$$

$$p_o(u) = \begin{cases} p_o & \text{if } o \in \{b, r\}, \\ p_p + q_p I_c(u) & \text{if } o = p, \end{cases} \tag{1.9}$$

où p_b, p_r, p_p, et q_p représentent la force de puits des organes, avec la convention $p_b = 1$, et c_o est un constante de normalisation définie par $c_o := \max_{x \in [0,1]} (x)^{a_o - 1} (1 - x)^{b_o - 1}$.

Nous supposons, comme Lemaire et al. (2008), que la force de puits des pétioles n'est donc pas constante, et varie en fonction d'un indice de compétition pour la lumière $I_c(u)$, défini par :

$$I_c(u) = 1 - \frac{s^{pr}}{k_b \, s^{act}(u)} \left(1 - \exp\left(-k_b \, \frac{s^{act}(u)}{s^{pr}} \right) \right), \quad u \geq 0, \tag{1.10}$$

où $s^{act}(u)$ est la surface foliaire photosynthétiquement active au temps u (voir équation 1.15). Cet indice permet de prendre en compte une éventuelle augmentation de la production des organes de soutien de la plante lorsque le feuillage commence à couvrir le sol. Cet indice de compétition tend vers 0 lorsque la surface foliaire tend vers 0, et vers 1 lorsqu'elle tend vers $+\infty$.

Les fonctions puits sont donc proportionnelles à des densités de lois bêta, dont l'allure dépend de la valeur des paramètres a_o et b_o. Les avantages de cette loi sont multiples (Yin et al., 2003), et résident notamment dans sa grande flexibilité, qui permet de modéliser aussi bien des courbes symétriques qu'asymétriques. De plus, seuls deux paramètres sont nécessaires pour obtenir les multiples formes possibles, ce qui rend l'estimation plus stable. À titre d'illustration, la figure 1.4 présente différentes formes possibles pour la densité de la loi bêta, en fonction de la valeur de ses paramètres.

Nous avons $p^o_{al} = (p_o, a_o, b_o) \in \mathbf{R}^*_+ \times [1, +\infty)^2$ pour $o \in \{r, b\}$ et $p^o_{al} = (p_p, q_p, a_p, b_p) \in (\mathbf{R}^*_+)^2 \times [1, +\infty)^2$ pour $o = p$.

Demande

La demande totale en biomasse de la plante au temps u est égale à la somme des demandes de tous les organes :

$$d(u; p_{al}) = s_r(u; p^r_{al}) + \sum_{o\in\{b,p\}} \sum_k s_{o,k}(u; p^o_{al}). \tag{1.11}$$

Biomasse des organes

Nous supposons que la plante accumule de la biomasse tout au long de la journée, et qu'elle distribue sa production aux différents organes à la fin de la journée. La biomasse allouée à chaque organe à la fin de la journée, ou de façon équivalente, au début du jour suivant, est donc proportionnelle à la biomasse créée ce jour-là. En notant q_0 la masse de la graine, la biomasse des organes de rang 1 (racine et cotylédons considérés comme une feuille unique) est donnée par :

$$q_{o,1}(t) = \begin{cases} 0 & \text{si } t \in [0, t_1), \\ \dfrac{q_0}{\gamma_1} \displaystyle\int_{t_1}^{t} \frac{s_{o,1}(u, p^o_{al})}{d(u, p_{al})}\, du & \text{si } t \in [t_1, t_2). \end{cases} \tag{1.12}$$

Si $t_2 \notin \mathbb{N}$, la biomasse des organes de rang 1 au début du jour $\lfloor t_2 \rfloor + 1$ continue à dépendre de la masse de la graine entre $\lfloor t_2 \rfloor$ et t_2, puis dépend uniquement de la biomasse produite par photosynthèse entre t_2 et $\lfloor t_2 \rfloor + 1$. Nous avons la formulation suivante :

$$q_{o,k}(\lfloor t_2 \rfloor + 1) = q_{o,k}(t_2) + \frac{s_{o,k}(\lfloor t_2 \rfloor, p^o_{al})}{d(\lfloor t_2 \rfloor, p_{al})}\, q(\lfloor t_2 \rfloor), \tag{1.13}$$

où $q(\lfloor t_2 \rfloor)$ résulte d'une adaptation de l'équation (1.4) à un pas de temps inférieur au jour et égal à $\lfloor t_2 \rfloor + 1 - t_2$.

Puis, à partir du jour t, pour $t > \lfloor t_2 \rfloor + 1$, la biomasse de l'organe o au début du jour t est donnée par :

$$q_{o,k}(t) = q_{o,k}(t-1) + \frac{s_{o,k}(t-1; p^o_{al})}{d(t-1, p_{al})}\, q(t-1), \tag{1.14}$$

avec la convention $q_r = q_{r,1}$.

Surface foliaire photosynthétiquement active

La surface foliaire photosynthétiquement active, qui participe à la production de biomasse par la plante, se déduit de la masse des limbes photosynthétiquement actifs et de la masse spécifique associée e_b. Plus précisément, au début du jour t, les feuilles photosynthétiquement actives sont celles qui sont apparues avant cette date et qui ne sont pas encore entrées en sénescence. Cela inclut notamment les feuilles qui sont encore en expansion, c'est-à-dire pour lesquelles $t_k \le t \le t^e_k$, et celles qui ne sont plus en expansion, mais toujours présentes sur la plante, c'est-à-dire pour lesquelles $t^e_k < t < t^s_k$. Pour ces dernières, la biomasse associée est celle atteinte à la fin de l'expansion.

La surface foliaire photosynthétiquement active au début du jour t est donc donnée par :

$$s^{act}(t) = \frac{1}{e_b} \sum_k q_{b,k}(t)\ \mathbf{1}_{[t_k, t^s_k)}(\tau(t)) \tag{1.15}$$

1.2 LNAS

Le modèle LNAS (Log Normal Allocation and Senescence) a été introduit par Cournède et al. (2013), et peut être vu comme une simplification du modèle Greenlab, dans lequel les organes ne sont plus considérés individuellement mais globalement, à l'échelle du compartiment. Dans le cas de la betterave, pour laquelle il a été initialement créé, seuls deux compartiments sont alors à prendre en compte (les feuilles et la racine). Le modèle peut être facilement généralisable à d'autres espèces de plantes, en ajoutant simplement de nouveaux compartiments d'organes pour les fruits, les entrenœuds, ... La masse des feuilles est ici considérée dans son ensemble, contrairement au modèle Greenlab où il est possible de faire la distinction entre les limbes et les pétioles. C'est donc la masse surfacique des feuilles et non celle des limbes, qui est utilisée pour calculer la surface foliaire photosynthétiquement active.

La production de biomasse au jour t, $q(t)$, se fait selon l'équation (1.3), avec :

$$\mathrm{LAI}(t) = \frac{q_g(t)}{e_g}, \tag{1.16}$$

où $q_g(t)$ est la masse des feuilles vertes au jour t, et e_g la masse surfacique correspondante.

La biomasse produite au jour t est ensuite allouée aux différents compartiments d'organes à la fin de la journée ou, de façon équivalente, au début du jour $t+1$:

$$\begin{aligned} q_l(t) &= q_l(t-1) + \gamma(t-1) \cdot q(t-1) \\ q_r(t) &= q_r(t-1) + (1-\gamma(t-1)) \cdot q(t-1), \end{aligned}$$

où q_l et q_r désignent respectivement la masse totale des feuilles (vertes et sénescentes) et des racines. La fonction γ s'obtient par transformation affine de la fonction de répartition G_a d'une loi log-normale :

$$\gamma(t) = \gamma_0 + (\gamma_f - \gamma_0) \cdot G_a(\tau(t)),$$

où γ_0 et γ_f correspondent respectivement à la proportion initiale et finale de biomasse allouée aux feuilles (voir Figure 1.5).

Pour permettre une meilleure interprétation biologique des paramètres du modèle, la loi log-normale sous-jacente est paramétrée par sa médiane μ_a et son écart-type σ_a. La médiane est en effet plus aisément interprétable que la moyenne dans notre cas, car elle correspond au temps thermique au-delà duquel la biomasse produite est allouée majoritairement aux racines. Il est ainsi plus facile d'obtenir des plages de variation pour ce paramètre, ce qui s'avèrera utile plus loin lors de l'analyse de sensibilité.

La masse de feuilles vertes q_g intervenant dans l'équation 1.16 s'obtient ensuite en soustrayant à la masse totale des feuilles celle des feuilles sénescentes, qui s'exprime à son tour selon la fonction de répartition G_s d'une loi log-normale de médiane μ_s et d'écart-type σ_s :

$$q_g(t) = \left(1 - G_s(\tau(t) - \tau_{sen})\right) q_l(t)$$

où τ_{sen} est le temps thermique de début de sénescence.

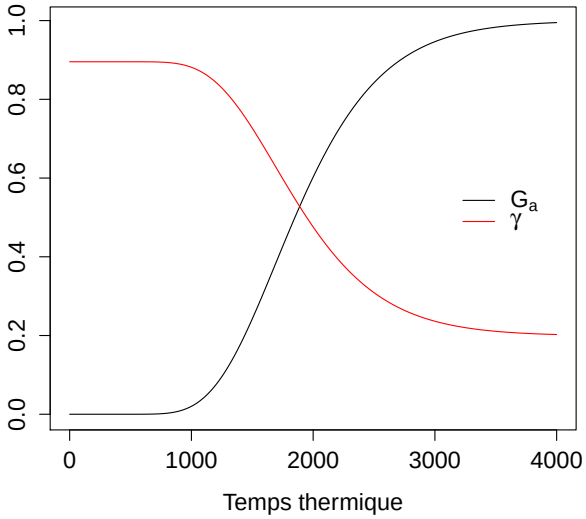


FIG. 1.5 – Fonction d'allocation du modèle LNAS (γ, en rouge) et fonction de répartition de la loi log-normale (G_a, en noir), avec une proportion initiale de biomasse allouée aux feuilles de 0.9 et finale de 0.2.

1.3 STICS

Le modèle STICS (Simulateur mulTIdisciplinaire pour les Cultures Standard (Brisson et al., 2003, 2008)) est un modèle journalier générique, qui est déjà largement utilisé pour une grande variété de cultures (tomate, maïs, vigne, blé, betterave sucrière, ...). Il est organisé en plusieurs modules eux-mêmes composés de sous-modules, chacun d'entre eux s'intéressant à l'un des processus spécifiques impliqués dans la croissance de la plante. Si la prise en compte de tous les modules permet d'aboutir à un modèle relativement complet, il n'est pas toujours nécessaire de tous les intégrer, et un modèle plus simple peut alors être construit. C'est le cas ici, où nous n'avons pas considéré les modules de bilans hydrique et azoté ou de transferts au niveau racinaire, ni l'effet du microclimat ou l'intervention agricole.

1.3.1 Production de biomasse

Dans la formulation initiale du modèle STICS, la production de biomasse dépend quadratiquement de la quantité de radiation interceptée, et non pas linéairement comme dans l'équation (1.3), avec l'introduction d'un coefficient de saturation. De plus, l'efficience de conversion RUE n'est pas considérée comme constante et varie en fonction du stade de développement de la plante.

Cependant, l'objectif de notre étude étant de comparer les modèles, d'une part sur l'échelle de prise en compte, et d'autre part sur la stratégie de répartition de la biomasse, il nous a paru nécessaire d'adopter pour STICS la même formulation pour la production de biomasse que pour les autres modèles, afin d'éviter l'introduction d'un possible biais de confusion dans l'étude comparative.

C'est donc l'équation (1.3) qui sert de référence pour la production journalière de biomasse, avec un indice de surface foliaire qui varie empiriquement de la façon suivante :

$$\mathrm{LAI}(t) = \sum_{j=t_2}^{t-1} (\Delta\mathrm{LAI}(j) - \Delta\mathrm{LAI}_{\mathrm{sen}}(j))$$

où t_2 est le jour d'émergence (voir section 1.1), $\Delta\mathrm{LAI}(j)$ la croissance nette de l'indice foliaire au jour j et $\Delta\mathrm{LAI}_{\mathrm{sen}}(j)$ l'indice de surface foliaire sénescent au jour j.

Les différentes formulations possibles pour la production de matière seront tout de même comparées à cette version simplifiée :

1. la version initiale du modèle, avec une efficience de conversion variable en fonction du stade de développement, et une relation quadratique entre la production de biomasse et la radiation interceptée
2. une version dans laquelle la production de biomasse dépend linéairement de la radiation interceptée, mais avec une efficience variable
3. la version avec production de biomasse linéaire et efficience constante (celle qui sera comparée aux autres modèles).

1.3.2 Croissance foliaire

L'évolution du LAI se fait généralement en trois phases : une première phase de croissance, une deuxième phase de stabilité, puis une dernière phase de sénescence où le LAI décroît linéairement. La deuxième phase de stabilité n'existe cependant que chez les plantes à croissance déterminée.

Lors de la première phase, la croissance du LAI suit une courbe logistique, depuis la levée (stade « ILEV ») jusqu'à sa valeur maximum (stade « ILAX »). Le point d'inflexion de la courbe marque la fin de la phase juvénile de croissance et le début d'une accélération de la croissance foliaire (stade « IAMF »). L'évolution du LAI dépend d'une unité de développement foliaire $u(j)$ qui vaut 1 au stade ILEV, 3 au stade ILAX, et u_{mat} au stade IAMF (Brisson et al., 2008). Entre ces trois stades, la valeur de $u(j)$ s'obtient par interpolation linéaire. Nous avons donc la formule suivante :

$$\Delta \text{LAI}(j) = \frac{\alpha}{1 + \exp(\beta(u_{\text{mat}} - u(j)))} \cdot f_d(j) \cdot f_T(j) \cdot s(j), \quad \text{si } 1 \leq u(j) \leq 3 \tag{1.17}$$

où f_d correspond à l'effet de la densité, f_T est la température effective de la culture, et s est un indice de stress trophique. Il est également possible d'introduire une diminution progressive de la croissance foliaire à l'approche du stade ILAX, au lieu d'un arrêt brutal comme c'est le cas ici (Brisson et al., 2008).

L'effet de la densité f_d permet de prendre en compte l'éventuelle superposition entre feuilles de plantes adjacentes, lorsque la densité dépasse un certain seuil fixé pour chaque espèce. En dessous de ce seuil, il n'y a pas de compétition, et au-delà, la surface foliaire par plante décroit de façon exponentielle. La température effective f_T dépend à la fois de la température de la culture, et de trois seuils de température : TCMIN et TCMAX, correspondant respectivement aux températures minimale et maximale de croissance foliaire, et TCXSTOP, correspondant au seuil critique de température au-delà de laquelle le développement de la plante est stoppé. La température effective est nulle avant TCMIN et après TCXSTOP, elle vaut $T_{cult}(j)$ - TCMIN entre TCMIN et TCMAX, et est réduite par un facteur multiplicatif entre TCMAX et TCXSTOP. En effet, au-delà d'une température optimale, les stomates de la plante ont tendance à se fermer pour minimiser l'évaporation, ce qui a pour effet d'inhiber la photosynthèse et donc la croissance foliaire.

L'indice de stress trophique s a quant à lui été introduit dans le modèle STICS pour les plantes à croissance indéterminée, pour lesquelles la compétition trophique entre feuilles et organes récoltés est un élément moteur de la croissance de la plante (Brisson et al., 2008). Rappelons que la betterave, en tant que plante bisanuelle (voir Introduction, 1.1), appartient à la catégorie des plantes à croissance déterminée, mais que par son mode de culture elle est considérée dans le modèle STICS comme une plante à croissance indéterminée. L'indice de stress dépend du rapport entre la biomasse créée (offre) et les forces de puits des organes de la plante (demande), et varie entre 0 et 1 grâce à une renormalisation du ratio source-puits entre

deux valeurs seuils minimale et maximale. Lorsque la biomasse produite est insuffisante pour répondre aux demandes des organes, le ratio source-puits s'approche de la valeur seuil minimale, et l'indice de stress trophique s'approche de 0. A l'inverse, lorsque la plante produit de la biomasse en quantité suffisante, le ratio augmente et l'indice de stress trophique tend vers 1.

L'indice de surface foliaire sénescente au jour j, $\Delta\text{LAI}_{sen}(j)$, correspond à l'indice de surface foliaire des feuilles qui sont tombées en sénescence entre le jour $j-1$ et le jour j, pour lesquelles la durée de vie est calculée en fonction du stade de développement de la plante.

1.3.3 Croissance racinaire

Dans STICS, la croissance des organes récoltés, aussi appelés « fruits », dépend du ratio entre la force de puits de ces organes, et la demande totale de la plante. Dans le cas de la betterave, qui est cultivée sur la première moitié de son cycle de vie, les organes « fruits » se réduisent au pivot de la racine. Au début du jour t, la masse de la racine peut donc s'écrire :

$$q_r(t) = q_r(t-1) + \frac{p_r(t-1)}{d(t-1)}\, q(t-1),$$

où p_r est la force de puits de la racine, et d la demande totale de la plante, i.e. $d = p_r + p_v$, avec p_v la force de puits de la partie végétative.

1.4 Pilote

Conçu à l'origine comme un outil d'aide à la décision pour l'irrigation, le modèle Pilote permettait de modéliser la teneur en eau du sol, divisé en deux réservoirs, l'un racinaire et évolutif, et l'autre permettant de capter le surplus d'eau provenant du premier réservoir, par drainage. Le modèle a cependant rapidement évolué vers une version plus complète, couplant un module « sol » à trois réservoirs dédié à l'estimation du bilan hydrique, à un module « plante » permettant de simuler le rendement (Mailhol et al., 1996, 1997). C'est un modèle journalier qui, s'il a d'abord été construit pour le sorgho et le tournesol, peut se généraliser facilement à d'autres types de plantes. Une version betterave a notamment été développée par Taky (2008).

Le module sol permet de gérer les transferts d'eau dans le sol à l'aide de différents réservoirs, l'eau pouvant passer de l'un à l'autre de ces réservoirs par drainage ou par évapotranspiration. Le paramètre fondamental du module sol est alors la réserve utile, définie comme la quantité d'eau que le sol peut contenir et qui peut être absorbée par la plante, ou encore, comme la différence entre la quantité d'eau à la capacité au champ et la quantité d'eau au point de flétrissement. Le module plante permet quant à lui de prédire le rendement de la culture, grâce à la modélisation de l'évolution de l'indice de surface foliaire ou LAI. Le LAI évolue alors selon une courbe empirique, et peut être directement affecté en cas de stress hydrique. Comme précisé en Section 1.6, nous adoptons ici une version simplifiée du modèle, contenant uniquement le module plante. Dans ce cas, le stress hydrique n'intervient pas dans le modèle, et le LAI s'exprime de la façon suivante :

$$\text{LAI}(t) = \text{LAI}_{\max}\left(\frac{\tau(t)-\tau_e}{\tau_{\max}}\right)^{\beta} \exp\left[\frac{\beta}{\alpha}\left(1-\left(\frac{\tau(t)-\tau_e}{\tau_{\max}}\right)^{\alpha}\right)\right] \tag{1.18}$$

où $\text{LAI}_{\max}$ est la valeur maximale que peut atteindre le LAI dans des conditions non-limitantes de croissance, $\tau_{\max}$ est le temps thermique nécessaire à la plante pour atteindre ce LAI maximal, τ_e est le

temps thermique d'émergence (voir Section 1.1), et α et β sont deux paramètres de calage de la courbe empirique. Il est également possible de modéliser la croissance du LAI et sa sénescence séparément, en utilisant deux valeurs différentes pour le paramètre α, selon que l'on se situe avant ou après τ_{max}.

La production de biomasse se fait ensuite, comme pour les autres modèles, grâce à l'équation (1.3), et la répartition de biomasse entre les différents organes de la plante se fait à l'aide d'un coefficient empirique appelé « Harvest Index » (indice de récolte). Plus précisément, nous avons :

$$q_r(t_{fin}) = \text{HI} \sum_{j=0}^{t_{fin}} q(j)$$

avec q_r la masse des racines et t_{fin} le jour de la récolte (dernier jour de mesure). Cette relation n'est donc valable qu'à la récolte, et il n'est pas possible de connaître la masse des racines en cours de croissance grâce au modèle.

1.5 CERES

Le modèle CERES (Crop Environment REsource Synthesis) a été construit dans les années 1980 par Jones et Kiniry (1986) sur le maïs, et une version betterave a été développée par Leviel (2000). Tout comme STICS, le modèle CERES est divisé en plusieurs sous-modèles qui interagissent entre eux, parmi lesquels un sous-modèle de sol permettant d'intégrer les effets de l'irrigation et de la fertilisation. Un seul stade de développement est considéré dans le cas de la betterave, et donc, il n'est pas possible de déterminer une date optimale de récolte correspondant au stade de maturité de la plante, comme cela se fait dans la version maïs de CERES, par exemple (Leviel, 2000).

La production de biomasse se fait à l'échelle du mètre carré, à partir de l'équation (1.3), mais la construction de l'indice foliaire se fait à partir des surfaces de feuilles de la plante individuelle. Plus précisément, le LAI est obtenu comme la somme des surfaces de toutes les feuilles de la plante, multipliée par la densité de plantation d :

$$\text{LAI}(t) = d \cdot \sum_k S_k(t),$$

où $S_k(t)$ est la surface de la feuille de rang k au temps t. Celle-ci se déduit de la surface maximale qu'elle peut atteindre au cours de sa croissance, en supposant une croissance linéaire de la feuille jusqu'à la fin de son expansion, la feuille gardant alors cette surface maximale jusqu'à la fin de sa durée de vie :

$$S_k(t) = \begin{cases} 0 & \text{si } \tau(t) \in [0, \tau_k[\\ \dfrac{S_{k,\text{max}}}{\tau_k^e - \tau_k} \left(\tau(t) - \tau_k^i\right) & \text{si } \tau(t) \in [\tau_k, \tau_k + \tau_k^e[\\ S_{k,\text{max}} & \text{si } \tau(t) \in [\tau_k + \tau_k^e, \tau_k + \tau_k^s[\\ 0 & \text{sinon} \end{cases}$$

où $\tau(t)$ est le temps thermique au temps t, et τ_k, τ_k^e et τ_k^s sont respectivement les temps thermiques d'initiation, d'expansion et de fin de vie d'une feuille de rang k (voir Section 1.1).

À la récolte, la biomasse des racines se déduit de la production cumulée de biomasse à l'aide d'un indice de récolte empirique HI comme dans Pilote :

$$q_r(t_{fin}) = \text{HI} \sum_{j=0}^{t_{fin}} q(j)$$

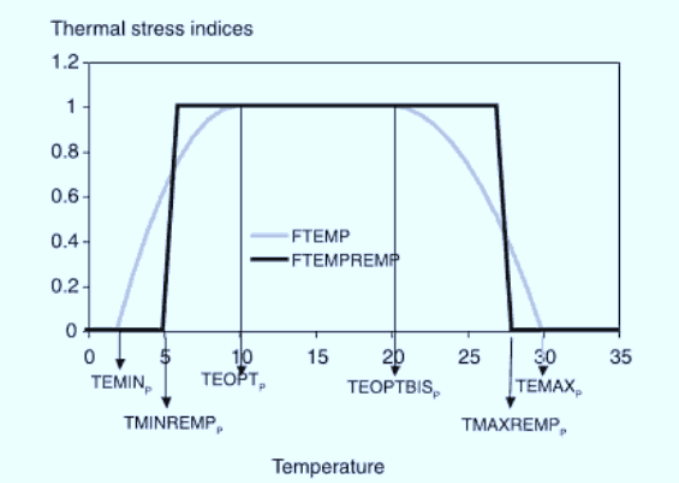


FIG. 1.6 – Prise en compte du stress thermique dans le modèle STICS. La courbe FTEMP correspond à l'indice de stress thermique agissant sur la RUE, et la courbe FTEMPREMP est l'indice agissant sur la croissance racinaire.

avec q_r la masse des racines et t_{fin} le jour de la récolte (dernier jour de mesure).Comme pour le modèle Pilote, il n'est pas possible de connaître la masse des racines en cours de croissance.

1.6 Prise en compte des stress

La prise en compte des stress dans les modèles de culture est primordiale, pour tenir compte de l'altération du fonctionnement de la plante dans des conditions non optimales de croissance. Ces stress peuvent être d'origines multiples : froid ou chaleur extrêmes, manque d'eau ou d'azote, pollution des sols, ... Parmi les modèles qui ont été présentés dans les sections précédentes, certains intègrent les effets de différents stress environnementaux, notamment les stress hydrique (STICS, Pilote, CERES), thermique (STICS) et azoté (STICS, CERES).

Les données dont nous disposons proviennent d'expérimentations au cours desquelles les plantes ont été cultivées sous des conditions non limitantes en eau et en azote (voir Sections 2.1 et 3.1), rendant l'étude de ces stress peu pertinente ici. En revanche, il nous a paru intéressant d'étudier l'effet d'un éventuel stress thermique sur le développement de la culture. Draycott (2008) suggère en effet que la température est un des facteurs climatiques ayant le plus d'influence sur la croissance de la betterave. Le modèle STICS permet d'intégrer facilement l'effet des différents stress environnementaux, grâce à son module dédié, et a donc été choisi dans un premier temps pour évaluer l'effet de la prise en compte du stress thermique.

Dans le modèle STICS, le stress thermique a un effet à la fois sur l'efficience de conversion de l'énergie lumineuse, et sur la croissance des organes de récolte (la racine dans le cas de la betterave). Grâce à l'introduction d'un indice de stress compris entre 0 et 1, l'efficience décroît en dehors d'une plage de températures optimales, et devient nulle lorsque la température est trop faible ou au contraire trop élevée. S'agissant de la croissance racinaire, elle est tout simplement inhibée en cas de stress thermique. Six seuils de températures sont ainsi nécessaires : les températures minimales et maximales pour assurer la croissance racinaire et pour permettre la photosynthèse et les températures optimales permettant d'obtenir une efficience maximale (voir Figure 1.6). Des valeurs par défaut sont proposées par Brisson et al. (2008) : TEMIN = 2°C, TEMAX = 30°C, TEOPT = 15°C, TEOPTBIS = 26°C, TMINREMP = 2°C, TMAXREMP = 38°C (voir Figure 1.6 pour les notations).

Finalement, une quatrième version du modèle STICS sera donc comparée aux trois premières, et permettra d'évaluer l'effet de la prise en compte du stress thermique sur la qualité des prédictions.

2 Calibration

Une fois le modèle et ses hypothèses définis de façon précise, le modélisateur ou le statisticien se retrouve confronté à la nécessité de le calibrer à l'aide d'un jeu de données expérimentales. Comme nous l'avons vu dans les sections précédentes, certains modèles de croissance de plantes comportent cependant un grand nombre de paramètres.

Toutefois, certains d'entre eux correspondent à des processus biologiques ou physiques connus pour lesquels nous disposons d'un certain nombre d'informations. Par exemple, le coefficient d'extinction de la loi de Beer-Lambert présent dans l'équation (1.3) a déjà été étudié par Andrieu et al. (1997) dans le cas de la betterave et pourra donc être considéré comme constant. De la même façon, la masse spécifique des limbes peut se déduire des mesures de masses et de surfaces des feuilles, et peut être mesurée indépendamment des autres paramètres du modèle.

Malgré cette première réduction du nombre de paramètres à estimer, l'estimation simultanée des paramètres restants peut s'avérer délicate. Tout d'abord, les méthodes d'estimation utilisées, dont la résolution est rarement explicite dans un cadre non-linéaire, dépendent d'algorithmes numériques qui peuvent ne pas converger lorsque la dimension de l'espace des paramètres est trop grande. De la même façon, de fortes corrélations peuvent exister entre les paramètres (et ceci est d'autant plus vrai que l'on introduit un grand nombre de paramètres dans le modèle), et des problèmes d'identifiabilité peuvent également survenir. Enfin, et c'est ce qui va nous préoccuper principalement ici, augmenter le nombre de paramètres à estimer peut avoir un effet délétère sur la robustesse du modèle.

Ce phénomène s'explique par la décomposition de l'erreur quadratique moyenne d'un prédicteur en un terme de biais et un terme de variance qui ne peuvent être minimisés simultanément. En notant θ le vecteur de paramètres du modèle, et $\hat{\theta}$ un estimateur de θ, et en supposant que le modèle s'écrit comme une fonction f du vecteur de paramètres θ, l'erreur quadratique moyenne associée au prédicteur $f(\hat{\theta})$ s'écrit :

$$\begin{aligned} \mathrm{MSE}(f(\hat{\theta})) &= \mathbb{E}[(f(\hat{\theta}) - f(\theta))^2] \\ &= \mathrm{Biais}^2(f(\hat{\theta})) + \mathrm{Var}(f(\hat{\theta})). \end{aligned}$$

Ainsi, si l'on diminue le biais du modèle en augmentant le nombre de paramètres estimés (en particulier si la valeur estimée est assez éloignée de la valeur de référence, ou si au contraire une telle valeur n'existe pas), tout processus d'estimation est entaché d'erreurs qui se répercutent sur la variance finale du modèle. Il convient donc d'aboutir à un compromis satisfaisant sur le nombre de paramètres à estimer pour garantir à la fois une bonne adéquation du modèle aux données, et une bonne robustesse (voir Figure 1.7). C'est le principe du rasoir d'Ockham [1] ou principe de parcimonie.

Certains critères de comparaison de modèles, tels le critère d'Akaike (AIC) ou le Bayesian Information Criterion (BIC), mettent ce principe en application en intégrant à la fois une composante prenant en compte l'ajustement du modèle, et un terme de pénalisation qui augmente avec le nombre de paramètres. Cependant, dans notre cas, on ne souhaite pas seulement déterminer le nombre P optimal de paramètres à estimer, mais le *meilleur* sous-ensemble de paramètres de taille P répondant aux objectifs du modèle. Pour cela, on va tout d'abord chercher à ordonner les paramètres selon un critère d'intérêt, avant d'utiliser les critères évoqués ci-dessus pour déterminer le sous-ensemble final de paramètres à estimer.

Cette approche a déjà été utilisée dans le contexte des modèles de plantes (Wallach et al., 2001), et a été détaillée dans (Wallach et al., 2006)(Chapitre 4). Elle consiste donc en deux étapes successives :

1. *« Pluralitas non est ponenda sine necessitate »* - Une pluralité ne doit pas être posée sans nécessité (Thorburn, 1918).

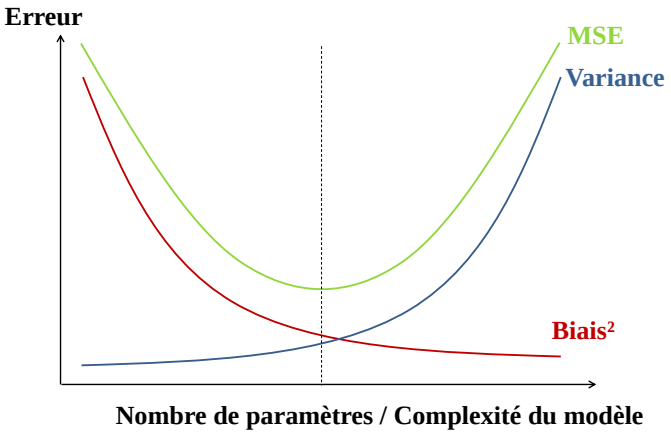


FIG. 1.7 – Évolution de l'erreur quadratique moyenne en fonction des termes de biais et de variance.

Étape 1 : Les paramètres sont ordonnés selon un critère d'intérêt défini par le modélisateur. Cela peut être un critère d'ajustement (tel le coefficient de détermination dans le cas de la régression linéaire), ou un critère rendant compte de l'influence du paramètre sur la variance du modèle, obtenu par une analyse de sensibilité.

Étape 2 : Une fois l'ordre des paramètres déterminé, un second critère est utilisé pour sélectionner le nombre de paramètres qui seront estimés. Les autres seront alors fixés à des valeurs nominales, provenant de la littérature ou d'informations dont on dispose *a priori*. Plusieurs critères peuvent être utilisés au cours de cette étape. Wallach et al. (2006) utilisent l'erreur quadratique moyenne de prédiction, évaluée par validation croisée, tandis que Tremblay et Wallach (2004) ont montré que dans leur cas, une version du BIC corrigée pour de petits échantillons donnait de meilleurs résultats. Ce critère est à définir en fonction de l'objectif du modèle.

Nous avons adopté une approche similaire, en utilisant une analyse de sensibilité lors de la première étape et les critères d'information AIC et BIC lors de la deuxième étape. En effet, nous disposons de peu de mesures sur notre jeu de données d'apprentissage, ce qui rend difficile le calcul du MSEP par validation croisée. Ces critères, même s'ils ne permettent pas nécessairement de minimiser l'erreur de prédiction du modèle, pénalisent toutefois les modèles avec un trop grand nombre de paramètres, et peuvent même s'avérer de meilleurs critères que le MSEP dans certains cas (Tremblay et Wallach, 2004).

Dans un premier temps nous décrivons le jeu de données d'apprentissage ayant servi à calibrer les cinq modèles en Section 2.1, puis les paramètres qui sont déduits directement de la littérature ou mesurés séparément en Section 2.2. Nous présentons en Section 2.3 les principes généraux de l'analyse de sensibilité, et la procédure de sélection de modèles permettant de sélectionner le nombre de paramètres qui seront estimés en Section 2.4.

2.1 Données d'apprentissage

Le jeu de données utilisé pour calibrer les modèles correspond aux expérimentations de l'Institut Technique de la Betterave (ITB) de 2010 (Didier, 2013), qui ont lieu à La Selve, en France (N49°34'22" E3°59'24"), avec la variété Python. Le semis a eu lieu le 15 avril 2010, et la culture a été fertilisée avec 136 kg d'azote (N) par hectare. La densité finale de population a été estimée à 11.82 pl/m^2 (la densité finale de population est en général légèrement plus faible que la densité de plantation car certaines graines

ne germent pas, et certaines plantules meurent prématurément). Les données météorologiques de température (en °C) et de rayonnement (en MJ.m^2) ont été obtenues à partir d'une station météorologique de Météo France située à Courcy, à 7 km du site expérimental.

Au total, quinze prélèvements ont été effectués entre le 7 juin et le 21 septembre 2010, au cours desquels les différents organes de la plante ont été pesés. Deux types de mesures ont été faites, d'une part des mesures globales où étaient pesés le bouquet foliaire entier et la racine de la plante, et d'autre part des mesures plus détaillées où chaque limbe et chaque pétiole de la plante ont été pesés séparément, en plus de la racine. Ces mesures détaillées étant plus fastidieuses à réaliser, elles ont été faites sur un plus petit nombre de plantes (10, au lieu des 50 plantes pesées lors des mesures globales), et à seulement cinq dates. Les surfaces foliaires individuelles ont été mesurées sur un sous-échantillon de cinq plantes parmi ces dix, à chacune des cinq dates de mesures détaillées. Une plante « moyenne » a ensuite été construite à partir des valeurs moyennes de poids secs à chaque date et pour chaque organe.

La construction de cette plante moyenne requiert quelques ajustements pour les masses foliaires individuelles ; en particulier, le nombre de feuilles est celui de la plante ayant le plus de feuilles, et lorsqu'une plante possède un plus petit nombre de feuilles, les masses foliaires des feuilles d'après sont considérées comme nulles. Cela permet de lisser la courbe de masse foliaire pour les rangs de feuilles élevés (Lemaire et al., 2008).

La surface foliaire, qui est une donnée d'entrée de trois des cinq modèles considérés, n'a pas été mesurée directement sur le champ, mais a été estimée à partir de la masse totale des limbes et de la masse surfacique des limbes. La masse moyenne de limbes par plante obtenue grâce aux mesures décrites dans le paragraphe précédent a été multipliée par la densité de population pour obtenir la masse totale de limbes dans le champ. Plus précisément :

$$\mathrm{LAI}_{\mathrm{exp}}(t) = \frac{q_b(t) \cdot d}{e_b}.$$

2.2 Paramètres considérés comme fixes

Comme précisé plus haut, les valeurs de certains paramètres ayant une signification biologique peuvent être déduites de la littérature. Il en va ainsi du coefficient d'extinction de la loi de Beer-Lambert, qui a été étudié dans le cas de la betterave par Andrieu et al. (1997), dans différentes conditions de densité et pour différentes dates de semis. Ce paramètre s'est révélé être assez stable dans chacune des conditions analysées, et se situe autour de 0.7 pour une densité d'environ 10 pl/m^2. Nous avons donc fixé $k_b = 0.7$ pour chaque modèle, sauf dans le cas où nous avons utilisé la version de STICS prenant en compte une relation quadratique entre la production de biomasse et l'interception lumineuse, auquel cas nous avons retenu la valeur de référence indiquée dans Brisson et al. (2008), soit $k_b = 0.58$.

Tous les autres paramètres considérés comme fixes ont été mesurés à partir des données expérimentales à notre disposition. Ainsi, la masse surfacique des limbes a été déduite des mesures de masse et surface de chaque feuille individuelle. À chaque date de prélèvement individuel, les surfaces foliaires individuelles et les poids secs correspondants ont été mesurés sur cinq plantes, puis la pente de la régression linéaire entre ces différents points a été utilisée comme approximation. Nous avons obtenu une valeur de 0.0083g/cm^2, avec un coefficient de détermination de 0.96. L'indice de récolte empirique utilisé dans les modèles Pilote et CERES a également été mesuré sur les données d'apprentissage, comme le rapport entre la masse sèche de la racine (collet et pivot) et la masse sèche totale (verte et sénescente), le jour de la récolte. Toutes les plantes pour lesquelles nous disposions de ces deux mesures au jour de la récolte, soit 60, ont été prises en

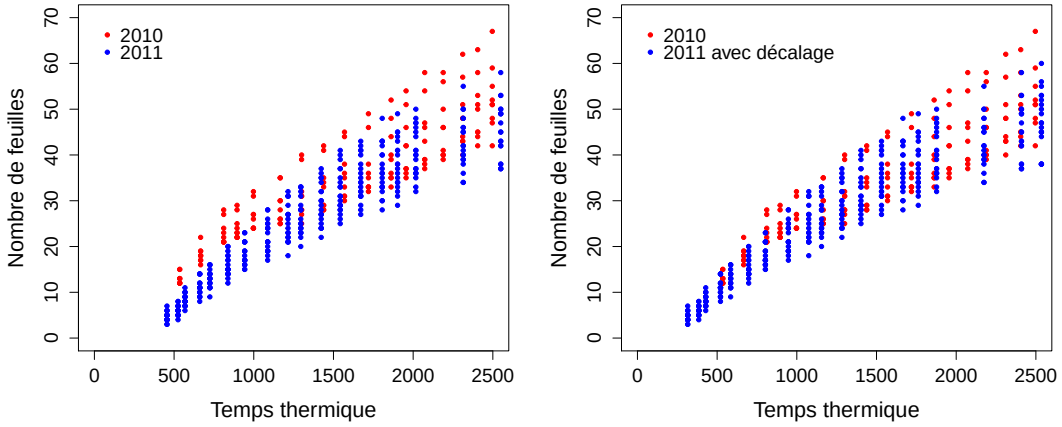
compte, et l'indice de récolte moyen était de 70%. Cette valeur est beaucoup plus faible que celle trouvée par Leviel (2000) dans son étude du modèle CERES pour la betterave, qui se situait autour de 85%.

L'un des paramètres les plus importants pour chaque modèle, comme nous le verrons dans la partie Résultats, et qui fera d'ailleurs l'objet d'une étude plus approfondie au Chapitre 3, est le temps thermique d'initiation, correspondant à la germination de la graine et au déploiement des cotylédons. Une mauvaise estimation de ce paramètre peut entraîner un décalage dans la prise en compte des données environnementales et avoir un effet délétère sur les performances du modèle. Les modèles présentés ici, dans leurs versions parfois simplifiées par rapport à leur formulation originale, ne permettent pas de prédire ce temps d'initiation, qui doit donc faire partie des entrées de chaque modèle.

Dans le cas de la betterave, la forte variabilité qui existe entre les individus nécessite de prélever un nombre suffisant de plantes afin d'obtenir une bonne estimation de la moyenne dans la population. Or, cette variabilité s'observe en particulier au niveau du temps d'initiation de la plante, qui dépend essentiellement des caractéristiques de la graine et des conditions de germination, elles-mêmes très variables au sein de la population ou du champ. L'utilisation d'une plante moyenne pour la calibration (voir Section 2.1), dont le nombre de feuilles correspond à celui de la plante la plus « rapide » nous pousse à utiliser les paramètres d'organogenèse de cette même plante, afin d'obtenir un même nombre de feuilles simulées par le modèle (ceci est valable pour les modèles individus-centrés, CERES et surtout Greenlab). Cette première approche a été utilisée avec succès lors des précédentes calibrations du modèle Greenlab chez la betterave (Lemaire et al., 2008 ; Lemaire, 2010), mais s'est avérée insuffisante lors des premiers tests du modèle en prédiction, à cause de la forte variabilité de ce paramètre dans la population, et à plus forte raison dans des conditions environnementales différentes.

La calibration étant faite sur des données *moyennes* provenant de différentes plantes, la deuxième approche que nous avons développée a consisté en l'utilisation d'une valeur *moyenne* pour le temps d'initiation. Pour cela, nous avons utilisé le modèle développé dans le Chapitre 3, afin d'obtenir une estimation qui puisse prendre en compte la variabilité de ce paramètre dans la population. Il s'agit d'un modèle mixte linéaire par morceaux, caractérisé par quatre paramètres : le temps thermique d'initiation, le temps thermique de rupture, et les deux phyllochrones associés à chacune des deux phases (voir équation (1.7)), implémenté dans le logiciel Monolix. Un premier essai de calibration, sur les 19 plantes pour lesquelles un suivi du nombre de feuilles a été réalisé, ne nous a pas permis d'estimer de façon satisfaisante les paramètres du modèle, en partie à cause de la petite taille d'échantillon, mais surtout parce que les premières observations ont été faites relativement tard et trop proches du temps thermique de rupture. Le peu de points disponibles pour décrire la première phase a rendu l'estimation des paramètres τ_1 et γ_1 peu stable. Or, comme l'ont observé Milford et al. (1985b) ; Lemaire et al. (2008) et Lemaire (2010), le phyllochrone moyen de la première phase reste stable d'une saison à l'autre, pour une même génétique, alors que l'initiation et la durée de cette première phase peuvent varier en fonction des conditions environnementales (elle sera par exemple plus courte lorsque la densité de plantation est élevée). Afin d'améliorer l'estimation, nous avons donc utilisé deux jeux de données provenant de deux années d'expérimentations, 2010 et 2011, réalisées sur la même variété, et dans des conditions environnementales similaires, ce qui nous a permis d'augmenter la taille de l'échantillon (40 plantes) et le nombre de mesures lors de la première phase, car les suivis ont commencé plus tôt en 2011.

En 2011, on observe également un décalage du temps d'initiation par rapport à l'année 2010, ce qui nous a incité tout d'abord à introduire dans le modèle un effet de l'année sur ce paramètre. Cependant, les deux temps d'initiation estimés par Monolix n'étaient pas significativement différents l'un de l'autre, indiquant probablement un problème d'identifiabilité, car les données suggèrent clairement un décalage entre



(a) Sans décalage

(b) Avec décalage de 140°Cj

FIG. 1.8 – Nombre de feuilles en fonction du temps thermique, en 2010 et 2011.

les deux jeux d'expérimentations. Ce décalage peut se voir sur la figure 1.8(a), et s'est également retrouvé sur les premiers essais de prévision sur le jeu de données 2011. En gardant le même temps d'initiation que celui estimé en 2010, les courbes de prévision obtenues laissaient nettement apparaître ce même décalage dû à une initiation trop tardive. Il est possible que le modèle utilisé soit suffisamment flexible pour permettre d'obtenir un temps d'initiation similaire sur les deux jeux de données, en particulier en considérant une variabilité plus forte des paramètres. Pour tenir compte de cette différence de temps d'initiation, qui peut avoir un impact majeur sur les prévisions s'il n'est pas correctement estimé, nous avons alors calibré ce paramètre manuellement. Plus précisément, nous avons introduit un décalage du temps thermique dans les expérimentations 2011, jusqu'à obtenir un nuage de points « homogène » entre les deux années, sur lequel nous avons estimé les quatre paramètres d'organogenèse. Le meilleur modèle au sens des critères AIC et BIC a été obtenu avec un décalage de 140°C jours (voir Figure 1.8(b)). Le temps thermique d'initiation en 2010 a été estimé à 103°C jours.

La liste des paramètres considérés comme fixes pour chaque modèle se trouve en Annexe A.

2.3 Analyse de sensibilité

2.3.1 Principes généraux

L'analyse de sensibilité d'un modèle permet d'évaluer la *sensibilité* de la variable réponse à des perturbations dans les entrées du modèle et permet ainsi d'identifier les paramètres ayant le plus d'influence sur les résultats du modèle. Elle fait partie des étapes indispensables dans l'élaboration et l'évaluation d'un modèle, aux côtés de l'analyse d'incertitude (Wallach et al., 2006, Chapitre 3). Saltelli et al. (2004) en donnent la définition suivante :

> « *[It is] the study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input.* »

Nous nous proposons de présenter dans ce paragraphe les principes généraux de l'analyse de sensibilité, ainsi que la méthode qui a été utilisée dans notre étude. Les différentes définitions et notions abordées sont largement inspirées des deux ouvrages de référence sur le sujet (Saltelli et al., 2000, 2008). Pour simplifier

la présentation, nous supposons que le modèle qui nous intéresse peut se représenter de la façon suivante :

$$y = f(z_1, \ldots, z_p)$$

où $z_1, \ldots, z_p$ sont les facteurs d'entrée (pour nous, ce sont les paramètres dont on veut étudier l'influence sur y).

On distingue classiquement deux types d'approches, d'un côté l'approche *qualitative*, qui permet de fournir une hiérarchie des facteurs selon leur influence sur la variabilité de la réponse, et de l'autre l'approche *quantitative*, qui fournit en plus une mesure de la proportion de variabilité imputable à chaque facteur. Deux grandes familles de méthodes quantitatives existent (Cariboni et al., 2007) :

- les méthodes *locales*, où l'on étudie l'effet de la variation locale d'un facteur z_i autour d'une valeur nominale, souvent à partir de dérivées partielles, tous les autres facteurs étant fixés à des valeurs moyennes ; ces méthodes dépendent du modèle, et il n'est pas possible d'affecter une densité de probabilité à chaque facteur pour mesurer sa variation, mais elles sont en général peu coûteuses en terme de temps de calcul
- les méthodes *globales*, où l'on autorise tous les paramètres à varier simultanément selon des lois de probabilités données dans tout leur intervalle de variation ; ces méthodes permettent l'identification d'éventuelles interactions entre les paramètres, et ne sont pas modèles-dépendantes, mais peuvent être particulièrement gourmandes en temps de calcul.

Le choix de la méthode dépend du contexte de l'étude et des ressources informatiques dont on dispose. Dans certains cas, une analyse locale basée sur les dérivées partielles de la variable réponse pourra suffire, par exemple si le modèle est linéaire. Dans le cadre des modèles écologiques, les méthodes globales sont en général plus appropriées à cause de la non-linéarité des modèles et aux fortes interactions qui peuvent exister entre les paramètres (Cariboni et al., 2007). Parmi celles-ci, les méthodes basées sur la décomposition de la variance de y permettent de s'affranchir de l'hypothèse de linéarité, de monotonie et d'additivité souvent requises par les autres méthodes. Les plus utilisées sont la méthode FAST (Fourier Amplitude Sensitive Test, Cukier et al. (1973)) basée sur la décomposition spectrale de la fonction f, et la méthode de Sobol (Sobol, 1993), basée sur des techniques de type Monte Carlo pour approcher la variance de y, et permettant d'approcher les indices de sensibilité d'ordre supérieur à 1, c'est-à-dire ceux prenant en compte les interactions entre paramètres.

Dans le cadre des modèles de croissance de plante, Wu (2012) propose dans ses travaux de thèse une amélioration de la méthode de Homma-Saltelli pour le calcul des indices de sensibilité de Sobol, permettant de calculer les indices de sensibilité d'ordre supérieur à 1 sans simulations supplémentaires. Dans le cadre de sa thèse, elle a également implémenté ces méthodes dans la plateforme de modélisation de l'équipe, les rendant facilement utilisables pour chaque modèle développé dans PyGMAlion. À la faveur de ses multiples avantages, nous avons donc opté pour la méthode de Sobol, malgré son coût en ressources informatiques, qui a pu être partiellement réduit grâce à l'utilisation d'un mésocentre de calcul.

La méthode de Sobol est basée sur une décomposition de la variance V du modèle en termes de dimension croissante :

$$V = \sum_{i=1}^{p} V_i + \sum_{1 \leq i,j \leq p} V_{ij} + \ldots + V_{1\ldots p}$$

où

$$V_i = \mathrm{Var}(\mathbb{E}(y \mid Z_i))$$
$$V_{ij} = \mathrm{Var}(\mathbb{E}(y \mid Z_i, Z_j)) - V_i - V_j$$

$$\ldots$$

$$V_{1\ldots p} = V - \sum_{i=1}^{p} V_i - \sum_{1\leq i,j\leq p} V_{ij} - \ldots - \sum_{1\leq i_1<\cdots<i_p-1\leq p} V_{i_1\ldots i_{p-1}}$$

On définit ensuite les indices de sensibilité correspondants : l'indice du premier ordre $S_i = \frac{V_i}{V}$, celui du second ordre $S_{ij} = \frac{V_{ij}}{V}$, et l'on procède de la même façon pour les indices d'ordre supérieur. On définit également l'indice de sensibilité total associé au facteur z_i, S_{T_i}, correspondant à la somme de tous les indices de sensibilité faisant intervenir le facteur i. L'indice du premier ordre permet de quantifier la proportion de la variance de y qui peut être expliquée par chaque facteur pris individuellement, l'indice du deuxième ordre permet de quantifier l'effet de l'interaction entre deux facteurs, ... La différence entre l'indice de sensibilité total et l'indice du premier ordre pour un facteur donné donne également une indication sur l'importance de l'interaction entre ce facteur et les autres. En effet, si ces deux indices sont proches, cela signifie que les interactions mettant en jeu ce facteur ont peu d'influence sur la variance V.

Le calcul de ces indices de sensibilité fait intervenir des intégrales de la fonction f qui sont souvent incalculables analytiquement, et qui sont approchées dans le cas de la méthode de Sobol par des méthodes de Monte Carlo. Nous rappelons brièvement le principe des méthodes de Monte Carlo, qui seront présentées de façon un peu plus détaillée dans le chapitre 2. Supposons que nous cherchons à approcher l'intégrale suivante :

$$I = \int f(x)\mu(x)dx,$$

par rapport à la densité de probabilité μ. À partir d'un échantillon de variables aléatoires iid $X_1, \ldots, X_m$ de loi μ, on peut alors estimer cette intégrale par la quantité suivante :

$$\hat{I} = \frac{1}{m}\sum_{i=1}^{m} f(x_i),$$

qui converge presque sûrement vers I d'après la loi forte des grands nombres. Pour le calcul des indices de sensibilité, un tirage aléatoire des facteurs d'entrée selon leurs lois de probabilité permet d'approcher les quantités nécessaires. Le lecteur pourra se référer à Wu et al. (2011) pour plus de détails sur la méthode utilisée pour le calcul des indices.

Nous avons supposé jusqu'à présent que la variable sortie du modèle y était unidimensionnelle, or, les modèles de croissance de plantes que nous avons décrits ci-dessous peuvent être vus comme des systèmes dynamiques dont les sorties sont multidimensionnelles, impliquant le calcul d'indices de sensibilité $S_i^j(t)$ à chaque pas de temps t considéré pour chaque composante $y_j, j = 1, \ldots, q$ de la variable y (Wu et al., 2011). L'indice de sensibilité généralisé du paramètre i pour la composante y_j est ensuite obtenu de la façon suivante :

$$S_i^j = \frac{\sum_{t=1}^{T} S_i^j(t)\mathrm{Var}(y_j(t))}{\sum_{t=1}^{T}\mathrm{Var}(y_j(t))},$$

et l'indice global du paramètre i est obtenu en pondérant les indices S_i^j par la variance $\mathrm{Var}(y_j)$ de chaque composante au cours du temps, calculée à partir des observations. En effet, la variance par composante calculée par le modèle a tendance à donner autant voire plus de poids aux variables d'indice foliaire qu'aux variables de poids racinaires, qui nous intéressent principalement dans le cas de la betterave sucrière. C'est pourquoi nous avons privilégié la pondération par la variance observée de chaque variable au cours du temps.

Tab. 1.2 – Intervalles de variation des lois uniformes utilisées pour chaque facteur et chaque modèle (voir Brisson et al. (2008) et l'annexe A pour la signification des paramètres du modèle STICS).

STICS	Greenlab	Pilote	LNAS
RUE	μ : [5 ;6]	RUE : [3 ;4]	RUE : [3 ;4]
α	s^{pr} : [0.05 ;0.10]	α : [1 ;2]	e : [50 ;70]
β	p_p : [0.3 ;0.45]	β : [0.5 ;3.5]	μ_a : [400 ;800]
ADENSP	q_p : [1.35 ;1.65]	τ_{max} : [800 ;1200]	σ_a : [200 ;2000]
BDENSP	a_b : [3.38 ;3.55]	LAI_{max} : [4 ;6]	μ_s : [2000 ;3000]
LAICOMP	b_b : [5 ;5.45]		σ_s : [3000 ;6000]
SLAMIN	a_p : [3.7 ;3.8]		γ_o : [0.7 ;1]
SPLAIMIN	b_p : [5.25 ;5.40]		γ_f : [0 ;0.3]
DUREEFRUIT	a_r : [4 ;4.6]		
DURVIEIP	b_r : [2.1 ;2.8]		
DURVIEFV			
PGRAINMAXI			
AFPFP			
BFPFP			
TIGEFEUILLE			

2.3.2 Application

Un des pré-requis nécessaire à l'utilisation des méthodes globales d'analyse de sensibilité, et donc de la méthode de Sobol, est la définition des densités de probabilités censées représenter la variabilité des facteurs d'entrée. S'il est parfois possible de trouver dans la littérature des informations sur la façon dont varient certains facteurs, il est également possible qu'aucune information ne soit disponible, soit parce que le modèle est utilisé hors du cadre dans lequel il était utilisé jusque là (adapté à une nouvelle espèce, par exemple), soit lors de l'élaboration d'un nouveau modèle. Nous avons choisi d'utiliser pour chaque paramètre une distribution uniforme sur son intervalle de variation, défini de la façon suivante :

- pour STICS, comme des valeurs de référence sont disponibles pour chaque paramètre (Brisson et al., 2008), nous avons utilisé un intervalle de variation de 10% autour de ces valeurs de références
- pour Greenlab, les intervalles de variation ont été définis d'après les valeurs de référence de Lemaire et al. (2008) pour la version discrète, et d'après les résultats de différentes calibrations du modèle continu sur plusieurs jeux de données à notre disposition
- pour Pilote, nous avons pris en compte les valeurs des paramètres obtenues pour différentes plantes dont la betterave (Mailhol et al., 1996, 1997 ; Taky, 2008)
- pour LNAS, comme il s'agit d'un nouveau modèle, nous avons utilisé les similitudes entre ce modèle et les modèles SUCROS (Guérif et Duke, 1998) ou Greenlab, et l'interprétation biologique de certains paramètres (Cournède et al., 2013).

Le modèle CERES ne contenant qu'un seul paramètre, il n'a pas été inclus dans l'analyse de sensibilité. Les intervalles de variation pour chaque paramètre de chaque modèle sont donnés dans le Tableau 1.2.

Les modèles ayant chacun leurs propres spécificités, ils n'ont pas tous besoin des mêmes données pour être calibrés. Les variables de sortie du modèle, qui seront utilisées pour calibrer le modèle en étant comparées aux observations, seront donc différentes d'un modèle à l'autre :

- pour STICS, il s'agit de la masse sèche totale, racinaire, et des limbes non sénescents

TAB. 1.3 – Classement des paramètres selon les indices de sensibilité totaux.

Ordre	STICS		Greenlab		Pilote		LNAS	
1	RUE	0.50	s^{pr}	0.42	β	0.51	RUE	0.37
2	β	0.29	μ	0.36	RUE	0.39	e	0.29
3	α	0.065	a_r	0.07	τ_{max}	0.03	γ_0	0.10
4	SLAMIN	0.03	b_r	0.05	α	0.03	γ_f	0.04
5	DUREEFRUIT	0.03	p_p	0.03	LAI_{max}	0.03	σ_a	0.025
6	PGRAINMAXI	0.03	a_b	0.004			μ_a	0.02
7	DURVIEIP	0.012	q_p	0.002			σ_s	0.012
8	TIGEFEUILLE	0.011	a_p	0.002			μ_s	0.007
9	BDENSP	0.003	b_b	$1.3.10^{-3}$				
10	ADENSP	5.10^{-4}	b_p	2.10^{-4}				
11	DURVIEFV	4.10^{-4}						
12	LAICOMP	3.10^{-4}						
13	AFPFP	2.10^{-4}						
14	BFPFP	3.10^{-5}						
15	SPLAIMIN	$5.\ 10^{-6}$						

- pour Greenlab, de la masse sèche racinaire, et de la masse individuelle et totale des limbes et des pétioles, sénescents ou non (cependant les masses individuelles, qui correspondent à des vecteurs d'environ 50 éléments, à chaque temps d'observation, n'ont pas été incluses dans l'analyse de sensibilité, pour alléger la procédure)
- pour Pilote, de la masse sèche totale et de l'indice foliaire
- pour LNAS, de la masse sèche totale, et de celles des feuilles vertes et des feuilles sénescentes
- pour CERES, enfin, seule la masse sèche totale est nécessaire

Les résultats de l'analyse de sensibilité sont présentés dans le Tableau 1.3. Les indices de sensibilité totaux ont été utilisés, pour tenir compte des interactions possibles entre les paramètres. Comme précisé dans le paragraphe précédent, les indices de chaque paramètre ont été obtenus par une pondération des indices calculés pour chacune des composantes de la variable de sortie du modèle, en fonction de la variance observée pour chaque composante. Ceci a permis en particulier de donner plus de poids aux variables de masse sèche racinaire ou totale, qui sont les variables d'intérêt principales dans le cas de la betterave sucrière. Ainsi, les paramètres les plus influents *globalement*, sont également les paramètres qui influent le plus sur les variables de masse sèche totale ou racinaire. Par exemple, dans le cas du modèle Pilote, la RUE est la deuxième variable ayant le plus d'influence sur la variabilité de la masse totale, mais seulement la cinquième plus influente sur la variabilité de l'indice foliaire. La pondération par la variance observée permet de donner plus d'importance à ce paramètre, qui n'est classé qu'en quatrième position si l'on pondère par la variance de chaque sortie simulée par le modèle.

Comme attendu, les paramètres liés à l'efficience de conversion de la lumière en biomasse sont parmi les plus influents. On retrouve ensuite les paramètres intervenant dans la construction de l'indice foliaire, en particulier pour les modèles sans allocation. Pour Greenlab, les paramètres d'allocation de biomasse à la racine sont plus influents que les paramètres d'allocation des parties végétatives.

2.4 Sélection du nombre de paramètres

Une fois que les paramètres ont été ordonnés selon leur influence sur les sorties du modèle, chaque modèle a été calibré avec un nombre croissant de paramètres, introduits dans le modèle dans l'ordre déterminé par l'analyse de sensibilité.

2.4.1 Méthode d'estimation

L'estimation paramétrique dans les modèles dynamiques discrets est décrite dans Goodwin et Payne (1977), et une application dans le cas du modèle Greenlab a été présentée par Zhan et al. (2003), Guo et al. (2006). Notons $(t_k)_{1\leq k\leq n}$ la séquence des temps (en jours depuis semis) auxquels la plante a été observée, et y_k le vecteur d'observations au temps t_k. Le vecteur d'observations est donc implicitement une fonction du vecteur de paramètres θ :

$$y = f(\theta) + \epsilon$$

où f représente le modèle utilisé, $y = (y_1, \ldots, y_n)^t \in \mathbb{R}^N$, et $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t$, avec $\epsilon \sim \mathcal{N}(0, \Sigma)$.
En supposant un modèle d'erreur gaussien, la log-vraisemblance s'écrit :

$$\mathcal{L}(\theta) = \left((2\pi)^N \det \Sigma\right)^{-1/2} \exp\left[-\frac{1}{2}\left(y - f(\theta)\right)^t \Sigma^{-1} \left(y - f(\theta)\right)\right]. \tag{1.19}$$

Lorsque la matrice de covariance Σ est connue, l'estimateur du maximum de vraisemblance coïncide avec celui des moindres carrés généralisés, et s'obtient par minimisation du critère suivant :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left(\left(y - f(\theta)^t \Sigma^{-1} \left(y - f(\theta)\right)\right)\right).$$

Dans la pratique, la matrice Σ est supposée connue, et est obtenue à partir des variances observées de chacune des variables. Plus spécifiquement, supposons que le vecteur d'observations y est ordonné en K sous-groupes de taille $N_k, k = 1, \ldots, K$ correspondant chacun à un type d'organe différent. On suppose ensuite que deux éléments d'un même groupe ont la même variance, et qu'ils sont mutuellement indépendants, ce qui nous donne la matrice de covariance suivante :

$$\Sigma = \begin{pmatrix} \sigma_1^2 I_{N_1} & 0 & \ldots & 0 \\ 0 & \sigma_2^2 I_{N_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \sigma_K^2 I_{N_K} \end{pmatrix}$$

où I_k est la matrice identité de taille N_k, et σ_k^2 la variance empirique du groupe d'organes k. Nous renvoyons à Cournède et al. (2011) pour une description détaillée de l'algorithme d'estimation.

2.4.2 Critères de sélection

Pour chaque modèle, les deux critères d'information AICc et BIC ont été utilisés, où le critère AICc correspond à une correction du AIC dans le cas où la taille de l'échantillon est trop faible. L'utilisation de ce critère permet de minimiser les risques de sur-apprentissage pour les petits échantillons, et converge vers

TAB. 1.4 – Critères AICc et BIC en fonction du nombre de paramètres estimés. Les résultats du modèle STICS pour un nombre de paramètres supérieur à 10 ne sont pas présentés, la vraisemblance étant constante à partir de ce point.

Nombre de paramètres		1	2	3	4	5	6	7	8	9	10
STICS	AICc	**4.63**	6.77	8.29	10.82	13.55	16.48	19.63	23.02	26.68	30.65
	BIC	**6.06**	9.50	12.18	15.70	19.26	22.81	26.37	29.92	33.48	37.03
Greenlab	AICc	166.91	74.40	74.10	59.46	60.77	62.54	**50.83**	51.20	53.25	55.31
	BIC	171.46	83.48	87.71	**77.59**	83.42	89.71	82.51	87.37	93.92	100.47
LNAS	AICc	**5.05**	7.12	9.44	11.77	14.34	17.05	19.92	22.95	-	-
	BIC	**6.71**	10.35	14.11	17.76	21.52	25.29	29.05	32.81		
PILOTE	AICc	47.08	45.50	16.15	17.63	**15.19**	-	-	-	-	-
	BIC	48.26	47.68	19.14	21.22	**19.13**					

la version non corrigée lorsque la taille de l'échantillon tend vers l'infini (Burnham et Anderson, 2002). Nous rappelons les définitions de ces deux critères :

$$\text{AICc} = -2\,(\ln \mathcal{L}(\hat{\theta}) - p) + \frac{2\,p\,(p+1)}{(n-p-1)} \tag{1.20}$$

et

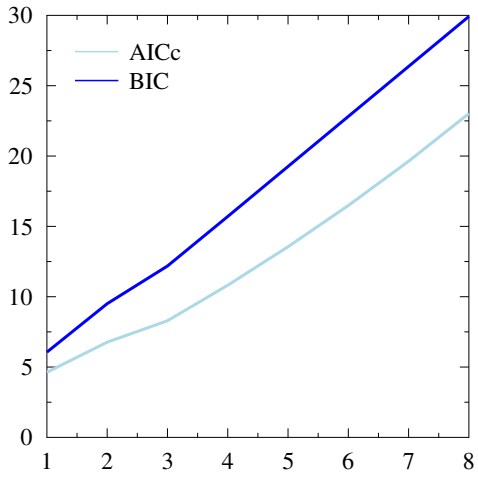
$$\text{BIC} = -2\,\ln \mathcal{L}(\hat{\theta}) + p \ln n, \tag{1.21}$$

où $\mathcal{L}(\theta)$ est la vraisemblance du modèle, $\hat{\theta}$ est l'estimateur du maximum de vraisemblance des paramètres du modèle, p le nombre de paramètres et n la taille de l'échantillon. Lorsque l'on compare différents modèles, on retient celui pour lequel ces critères sont minimaux.

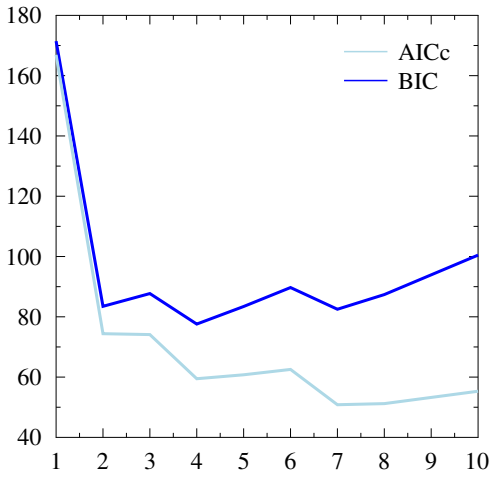
D'après les définitions ci-dessous, le calcul des deux critères nécessite l'utilisation de la méthode du maximum de vraisemblance pour obtenir un estimateur de θ. Cependant, dans le cas gaussien, et lorsque la matrice de variance-covariance des erreurs est supposée connue, le maximum de vraisemblance coïncide avec l'estimateur obtenu par les moindres carrés généralisés, méthode qui a été utilisée ici pour calibrer les différents modèles.

Les résultats de la procédure de sélection du nombre de paramètres sont présentés dans le Tableau 1.4 et en Figure 1.9. Les deux critères AICc et BIC fournissent des résultats similaires, sauf pour le modèle Greenlab, où la version corrigée du AIC préconise l'estimation de 7 paramètres, et le critère BIC seulement 4. Les deux versions seront comparées sur les données tests.

La liste des paramètres sélectionnés pour chaque modèle ainsi que les valeurs estimées correspondantes sont présentées dans le Tableau 1.5. Les autres paramètres ont été fixés à la valeur moyenne de l'intervalle de variation utilisé pour l'analyse de sensibilité (voir Tableau 1.2). Pour les modèles LNAS et STICS, l'estimation d'un seul paramètre suffit à assurer une bonne calibration du modèle, et l'ajout de paramètres supplémentaires ne permet pas d'accroître suffisamment la vraisemblance du modèle. Pour Greenlab et Pilote, il est nécessaire d'inclure un plus grand nombre de paramètres pour calibrer les modèles. L'efficience au niveau du mètre carré pour le modèle Greenlab est égale à 5.93 g.MJ^{-1} pour la version à 4 paramètres, et à 4.03 g.MJ^{-1} pour la version à 7 paramètres (voir Section 1.1 pour le détail du calcul).

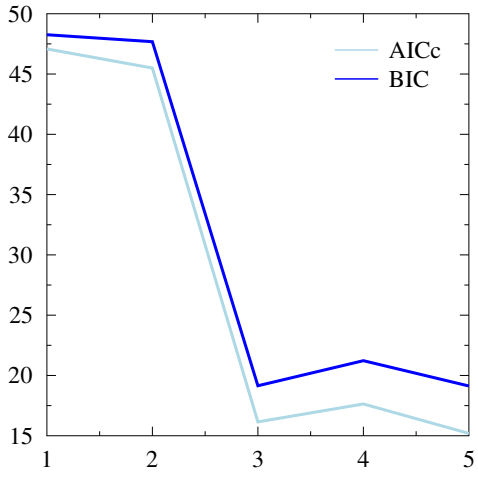


(a) STICS

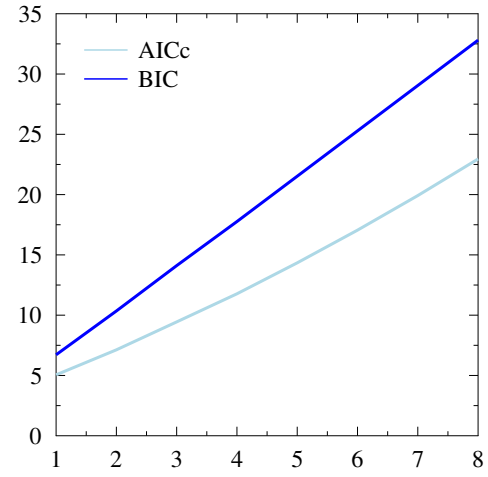


(b) Greenlab

FIG. 1.9 – Évolution du AICc et du BIC en fonction du nombre de paramètres.



(a) Pilote



(b) LNAS

FIG. 1.9 – (suite) Évolution du AICc et du BIC en fonction du nombre de paramètres.

Tab. 1.5 – Données utilisées pour calibrer chaque modèle, et estimation des paramètres.

Modèle	Données de calibration	Estimation
Greenlab 4	Masse de la racine, des limbes et des pétioles Masses individuelles des limbes et pétioles	$\mu = 5.49$ $s^{pr} = 0.0914$ $\rightarrow$ RUE $= 5.93$ $a_r = 4.06$ $b_r = 1.77$
Greenlab 7	Masse de la racine, des limbes et des pétioles Masses individuelles des limbes et pétioles	$\mu = 5.55$ $s^{pr} = 0.0615$ $\rightarrow$ RUE $= 4.03$ $a_r = 3.16$ $b_r = 1.04$ $p_p = 0.0039$ $a_b = 3.08$ $q_p = 1.70$
LNAS	Masse de la racine Masse des feuilles vertes Masse des feuilles sénescentes	RUE = 3.53
PILOTE	Masse totale Indice foliaire LAI	RUE = 4.12 $\alpha = 1.54$ $\beta = 1.92$ $\tau_{max} = 1830$ LAI$_{max} = 3.99$
CERES	Masse totale	RUE = 4.37
STICS	Masse de la racine Masse des limbes verts Masse totale	RUE = 4.76

3 Prévision

Une fois que les modèles ont été calibrés sur le jeu de données d'apprentissage, leurs capacités de prédiction sont testées sur un jeu de données test indépendant. Les prédictions de chaque modèle ont été simulées avec le même jeu de paramètres que celui obtenu à l'étape de calibration, seuls la densité de plantation et le temps thermique d'initiation ont été adaptés au jeu de données test. Nous présentons dans un premier temps le jeu de données test, puis les critères utilisés pour évaluer les capacités prédictives des modèles.

3.1 Données test

Initialement, le jeu de données test devait correspondre aux expérimentations de 2011, conduites sur la même variété Python, à proximité du site d'expérimentation de 2010 (dans la commune de Bourgogne, N49°21'18", E4°4'12") (Didier, 2013). En 2011, le semis a eu lieu le 21 mars, et la densité population a été estimée à 10.89 pl/m^2. Cependant, un épisode de grêle assez violent a eu lieu le 29 juin 2011, environ trois mois après le semis, et a causé de gros dommages aux feuilles, certaines d'entre elles ayant été partiellement ou même totalement détruites.

Par conséquent, même si nous avons comparé les modèles sur ce jeu de données 2011, le véritable jeu de données utilisé pour l'évaluation des qualités prédictives des modèles est celui provenant des ex-

Tab. 1.6 – Jeux de données utilisés pour la calibration et la validation.

	Données d'apprentissage	**Données de validation**	
Année	2010	2011	2008
Site	La Selve	Bourgogne	Bazainville
GPS	N49°34'22" E3°59'24"	N49°21'18" E4°4'123	N48°11'15" E2°5'50"
Variété	Python	Python	Radar
Densité	11.82 pl/m^2	10.89 pl/m^2	10.9 pl/m^2

périmentations de 2008, qui ont été faites sur la variété Radar, dont le génotype est similaire à celui de la variété Radar utilisée en 2010. Le site expérimental se situe à environ 200km au sud-ouest du site de 2010, à Bazainville, dans le Loiret (France, N48°11'15", E2°5'5"). Le semis a eu lieu le 11 mars 2008, et la densité de population finale a été mesurée à 10.9 plantes/m^2. La culture a été irriguée selon ses besoins, calculés à l'aide de l'outil IRRIBET développé à l'Institut Technique de la Betterave, qui considère que la plante a besoin d'un apport en eau dès que la réserve disponible du sol est inférieure à 40% de la réserve utile (Lemaire, 2010). De la même façon, les besoins en azote de la culture ont été calculés par l'outil AZOFERT développé à l'INRA, prenant en compte les besoins de la betterave et les reliquats azotés contenus dans le sol (pouvant provenir par exemple des cultures précédentes sur le même sol). Trente plantes ont été prélevées à 8 dates différentes pour les mesures par compartiments, et 10 plantes supplémentaires ont été pesées de façon détaillée (feuille par feuille) sur 5 des 8 dates de prélèvement. Les données climatiques proviennent d'une station météorologique de Météo France située à Pithiviers, à 8 km du site expérimental.

Pour que les capacités prédictives des modèles puissent être comparées, il faut que les critères d'évaluation présentés dans les sections suivantes soient définis sur les mêmes variables de sortie. Outre la production de biomasse, qui est calculée par chaque modèle, nous avons également comparé les modèles sur leurs capacités à prédire la masse racinaire, cette dernière étant d'un intérêt majeur pour la betterave sucrière. Les données de masses individuelles mesurées en 2008 n'ont donc pas été utilisées.

3.2 Critères

3.2.1 Erreur quadratique moyenne de prédiction (MSEP)

L'erreur quadratique moyenne de prédiction, ou MSEP, est le critère standard pour l'évaluation des capacités prédictives d'un modèle (Wallach et Goffinet, 1987 ; Wallach et al., 2006). Elle mesure la distance entre les observations y et les prédictions du modèle $f(\hat{\theta})$, et est définie de la façon suivante :

$$\mathrm{MSEP}(\hat{\theta}) = \mathbb{E}\left[(y - f(\hat{\theta}))^2 \mid \hat{\theta}\right]. \tag{1.22}$$

Le MSEP ayant pour unité le carré de celle des observations, il est d'usage d'utiliser la racine carrée de l'erreur quadratique :

$$\mathrm{RMSEP}(\hat{\theta}) = \sqrt{\mathrm{MSEP}(\hat{\theta})}. \tag{1.23}$$

Dans la pratique le RMSEP n'est pas calculable explicitement car il dépend de quantités inconnues, et doit donc être estimé.

Lorsque l'on dispose d'un deuxième jeu de données indépendant de celui sur lequel a été conduit l'estimation des paramètres, un estimateur sans biais de (1.22) est donné par (Wallach et al., 2006) :

$$\hat{\text{RMSEP}}(\hat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1.24}$$

avec $(y_1, \ldots, y_n)$ les observations et $(\hat{y}_1, \ldots, \hat{y}_n)$ les prédictions du modèle.

Dans notre cas, le RMSEP a été calculé pour les deux variables de masse totale et racinaire, lorsque ceci a du sens. En particulier, il n'a pas été calculé pour la masse racinaire prédite par Pilote et CERES, ces deux modèles dépendant d'indices de récolte empiriques qui n'ont pas l'ambition d'être valides tout au long de la croissance de la plante, mais uniquement à la fin, lors de la récolte. Un critère spécifique pour la prédiction du rendement a été utilisé à la place pour ces deux modèles (voir Section 3.2.3).

3.2.2 Efficience de modélisation (EF)

L'efficience de modélisation, ou EF, a été définie par Mayer et Butler (1993). C'est une quantité sans dimension qui permet de mesurer la qualité d'ajustement entre prédictions et observations. Il peut s'interpréter de façon similaire au coefficient de détermination de la régression linéaire (voir Section 3.2.4). Elle est définie par :

$$\text{EF} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}, \tag{1.25}$$

où $\overline{y}_i$ correspond à la moyenne des observations.

Ce critère est compris entre $-\infty$ et 1. En cas d'égalité parfaite entre les prédictions et les observations, l'efficience EF vaut 1, et lorsque les prédictions du modèle sont moins bonnes que la moyenne sur l'échantillon, l'efficience devient négative. Un des avantages de l'EF sur le RMSEP réside dans l'absence d'unité associée, ce qui le rend plus facilement comparable pour des variables n'ayant pas les mêmes ordres de grandeurs.

Tout comme pour le RMSEP, ce critère a été évalué pour chaque modèle sur la biomasse totale et sur la masse racinaire, sauf pour Pilote et CERES, pour lesquels l'erreur relative de prédiction du rendement a été calculée à la place.

3.2.3 Erreur relative de prédiction du rendement

Pour les deux modèles qui dépendent d'un indice de récolte empirique et non de processus biologiques d'allocation pour la biomasse racinaire, le calcul des deux critères précédents n'a pas de sens. En effet ces modèles ne prennent pas en compte l'évolution dynamique de la masse racinaire au cours du temps, et de ce fait cette variable ne fait pas partie des sorties du modèle. Un troisième critère a donc été défini pour pouvoir comparer ces deux modèles aux trois autres, sur la prédiction du rendement :

$$ype = \frac{|y_{r,n} - \hat{y}_{r,n}|}{y_{r,n}}, \tag{1.26}$$

où $y_{r,n}$ et $\hat{y}_{r,n}$ sont les masses racinaires respectivement observée et prédites, au moment de la récolte.

3.2.4 Observations *vs.* prédictions

Une premier examen visuel des qualités prédictives d'un modèle peut se faire à l'aide du graphe des observations en fonction des prédictions, accompagné de la droite d'équation $y = x$, qui permet d'identifier rapidement d'éventuels biais de prédiction. Le modèle de régression sous-jacent peut s'écrire

$$y_i = a + b\hat{y}_i + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{1.27}$$

De prime abord, il peut paraître surprenant d'effectuer la régression des données observées sur les données prédites (« OP »), plutôt que la régression des prédictions sur les observations (« PO »). Pourtant, l'aléatoire se situe essentiellement dans les observations, alors que les prédictions sont issues d'un modèle qui est soit déterministe, soit qui peut-être relancé un grand nombre de fois pour en augmenter la précision (Mayer et Butler, 1993). Cette intuition est confirmée par les travaux de Piñeiro et al. (2008), qui ont comparé les deux approches sur un jeu de données simulées. Dans le cas idéal où $y_i = \hat{y}_i + \epsilon_i$, ils ont montré que seule la régression OP permettait d'obtenir une pente égale à 1 et une ordonnée à l'origine nulle, alors que la régression PO sous-estimait la pente et sur-estimait l'ordonnée à l'origine.

Différents paramètres associés à la régression (1.27) peuvent servir d'indicateurs sur la qualité prédictive du modèle : le coefficient de détermination r^2, la pente $\hat{b}$ et l'ordonnée à l'origine $\hat{a}$. Nous rappelons leurs définitions :

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^n (y_i - \overline{y}_i)(\hat{y}_i - \overline{\hat{y}}_i)}{\sum_{i=1}^n (\hat{y}_i - \overline{\hat{y}}_i)^2}, \\ \hat{a} &= \overline{y} - \hat{b}\,\overline{\hat{y}}, \\ r^2 &= 1 - \frac{\sum_{i=1}^n (y_i - (\hat{a} + \hat{b}\,\hat{y}_i))^2}{\sum_{i=1}^n (y_i - \overline{y})^2}. \end{aligned} \tag{1.28}$$

Le r^2 s'interprète comme la proportion de variance expliquée par la droite de régression des y_i sur les $\hat{y}_i$. On voit bien ici la similitude avec l'efficience EF présentée en (1.25), qui peut s'interpréter, de façon similaire, comme la proportion de variance expliquée par la droite d'équation $y = \hat{y}$.

Le r^2 permet de mesurer la variabilité des observations autour de la droite de régression, et les coefficients $\hat{a}$ et $\hat{b}$ permettent de rendre compte du biais. Une bonne qualité de prédiction se traduit donc par un r^2 élevé, une pente $\hat{b}$ proche de 1, et une ordonnée à l'origine $\hat{a}$ proche de 0, ces trois conditions devant être vérifiées simultanément. En effet, même en présence d'un biais systématique, le r^2 peut être élevé (voir Figure 1.10(a)), et inversement, une pente égale à 1 et une ordonnée à l'origine nulle n'impliquent pas nécessairement de bonnes prédictions (voir Figure 1.10(b)).

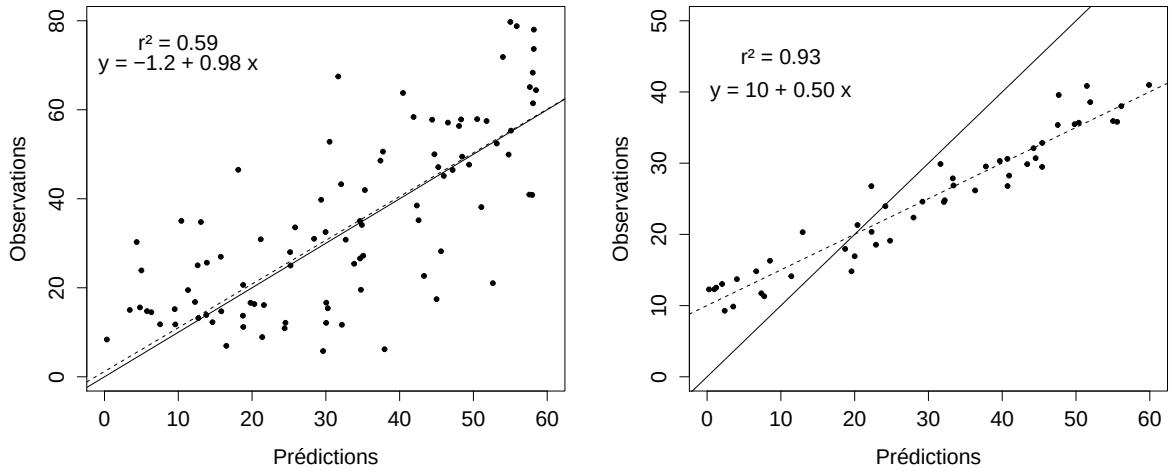
L'absence de biais étant caractérisée par $a = 0$ et $b = 1$, il est possible de tester l'absence de biais à l'aide du test d'hypothèses suivant : $H_0 : \{a = 0 \text{ et } b = 1\}$ vs. $H_1 : \{a \neq 0 \text{ ou } b \neq 1\}$. Le test du rapport de vraisemblance nous donne dans ce cas la statistique de test suivante :

$$T = \frac{\|\hat{\hat{\mathcal{M}}} - \hat{\mathcal{M}}\|^2/2}{\|y - \hat{\mathcal{M}}\|^2/(n-2)}$$

où $\hat{\mathcal{M}}$ est l'estimateur du maximum de vraisemblance du modèle $\mathcal{M}$ sous $H_0 \cup H_1$, et $\hat{\hat{\mathcal{M}}}$ est l'estimateur du maximum de vraisemblance du modèle $\mathcal{M}$ sous H_0.

On a :

$$\|\hat{\hat{\mathcal{M}}} - \hat{\mathcal{M}}\|^2 = \sum_{i=1}^n (\hat{y}_i - \hat{a} - \hat{b}\hat{y}_i)^2$$



(a) Biais nul avec r^2 faible

(b) Biais avec r^2 élevé

FIG. 1.10 – Régressions des observations sur les prédictions, illustrant le cas où la variabilité est trop élevée (r^2 faible) malgré une pente proche de 1 et une ordonnée à l'origine proche de 0 (a), et le cas où une faible variabilité entraine un r^2 élevé malgré un biais évident (b).

$$= n\hat{a}^2 + 2n\hat{a}(\hat{b}-1)\overline{\hat{y}} + (\hat{b}-1)^2 \textstyle\sum_{i=1}^{n} \hat{y}_i^2$$

$$\|y - \hat{\mathcal{M}}\|^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}\hat{y}_i)^2.$$

Sous l'hypothèse nulle, la statistique T suit une loi de Fisher :

$$T \overset{H_0}{\sim} \mathcal{F}_{2,n-2}.$$

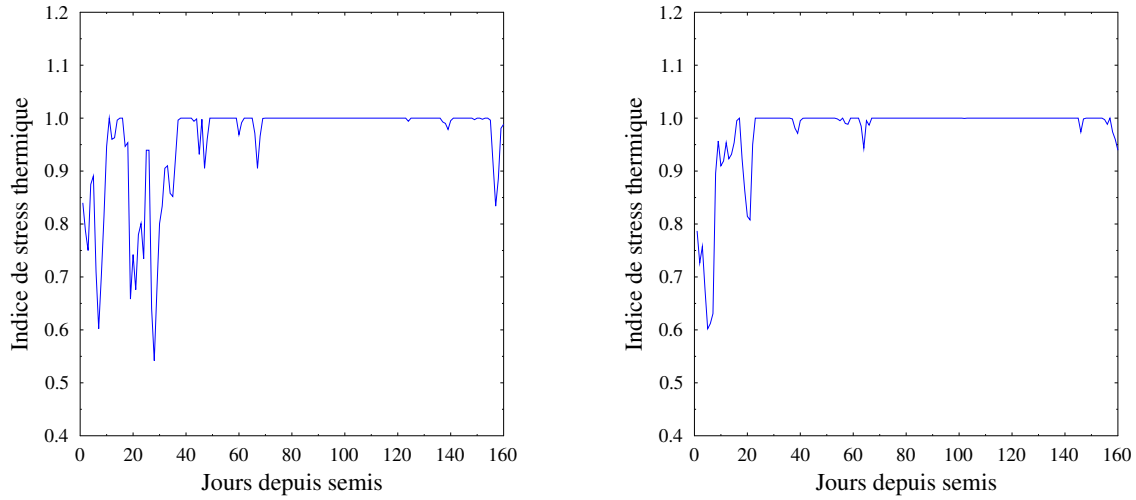
4 Résultats

4.1 Comparaison des différentes versions de STICS

Comme précisé aux Sections 1.3 et 1.6, plusieurs versions du modèle STICS ont été testées, en fonction de la relation entre la production de biomasse et l'interception lumineuse (linéaire ou quadratique), de l'efficience de conversion (constante ou variable selon le stade de développement), et enfin, selon la prise en compte du stress thermique (sur la version linéaire avec efficience constante). Nous présentons ici la comparaison de ces différentes versions sur les données 2008 (voir Tableau 1.7).

TAB. 1.7 – Comparaison des différentes versions de STICS sur les données 2008. Pour le F-test, nous présentons les p-values associées au test.

Version de STICS	Masse sèche totale				Masse sèche racinaire			
	RMSEP	EF	r^2	F-test (p)	RMSEP	EF	r^2	F-test (p)
Initiale	271.24	0.9217	0.999	< 0.0001	59.26	0.9945	0.998	0.051
Linéaire	192.3	0.9606	0.999	< 0.0001	39.05	0.9976	0.998	0.884
Linéaire et RUE cst	168.87	0.9696	0.999	< 0.0001	39.49	0.9976	0.998	0.800
Avec stress thermique	171.73	0.9686	0.999	< 0.0001	38.61	0.9977	0.998	0.845



(a) 2010 (b) 2008

FIG. 1.11 – Indice de stress thermique en 2010 et 2008. La levée de la plante moyenne a eu lieu aux alentours du jour 13 en 2010 et du jour 25 en 2008.

D'après les critères utilisés, les versions modifiées de STICS fournissent de meilleures prédictions que la version initiale avec production de biomasse quadratique et efficience variable. D'un point de vue purement prédictif, et dans les conditions dans lesquelles le modèle a été testé, la version choisie semble plus appropriée. La prise en compte du stress thermique a permis d'améliorer légèrement les performances du modèles pour la prédiction de la masse racinaire, mais pas pour la biomasse totale. Nous remarquons également que, si les quatre versions de STICS permettent d'obtenir des prédictions non biaisées pour la masse de la racine, il n'en est pas de même pour la biomasse totale, qui est moins bien prédite. Nous verrons dans le paragraphe suivant que ce comportement se retrouve dans tous les modèles qui ont été évalués.

Ces résultats nous confortent dans le choix de la version de STICS qui a été comparée aux autres modèles, c'est-à-dire celle avec production de biomasse linéaire, et avec une efficience constante. D'autre part, devant le peu d'effet qu'a eu la prise en compte du stress thermique sur les performances du modèle, nous décidons de ne pas l'inclure dans les autres modèles. L'indice de stress varie en effet très peu en 2008 après l'émergence (voir Figure 1.11). Avant l'émergence, le stress thermique peut avoir un effet sur la germination en la retardant, voire même en empêchant certaines graines de germer. Comme nous travaillons sur une plante « moyenne », l'effet inhibant n'est pas pris en compte, mais le retard potentiel de la germination peut avoir des répercussions sur le temps thermique d'initiation.

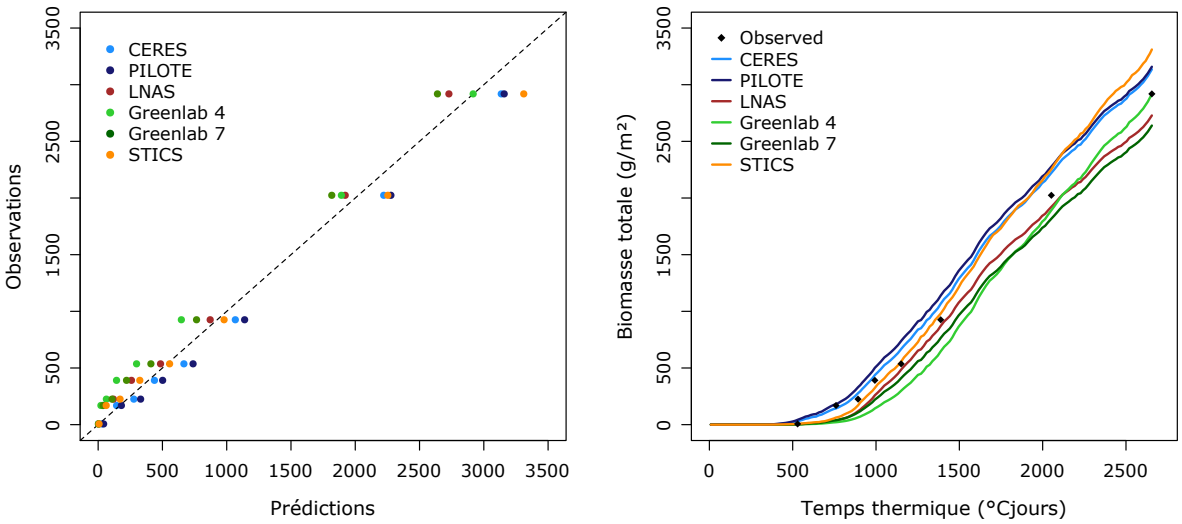
4.2 Comparaison sur les données 2008 : génotype similaire, environnement différent

Nous présentons dans cette section la comparaison des cinq modèles sur les données 2008. Nous rappelons qu'un épisode de grêle a eu lieu environ trois mois après le semis en 2011, ce qui nous a incités à utiliser le jeu de données 2008 comme échantillon test.

Comme lors de la comparaison des quatre versions du modèle STICS, les performances des modèles sont moins bonnes pour la biomasse totale que pour la masse racinaire (voir Tableaux 1.8 et 1.9). En particulier, le critère EF est un peu plus faible pour la masse totale, et surtout, chaque modèle fournit des

prédictions biaisées. Les valeurs très élevées pour les critères EF et r^2 s'expliquent en partie par la petite taille d'échantillon (seulement 8 mesures ont été faites en 2008), ce qui permet d'obtenir une très bonne corrélation linéaire entre les points du fait d'une plus faible variabilité.

4.2.1 Masse totale



(a) Observations vs. prédictions

(b) En fonction du temps thermique

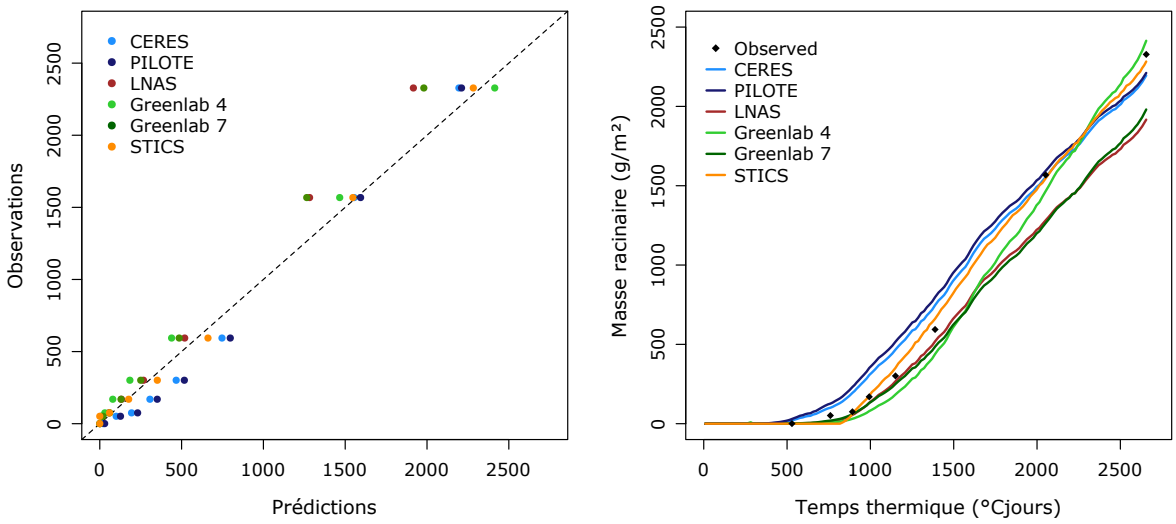
FIG. 1.12 – Prédictions de biomasse totale en 2008.

TAB. 1.8 – Comparaison des modèles sur la prédiction de la masse totale en 2008.

Modèle	Masse sèche totale			
	RMSEP	EF	r^2	F-test (p)
Greenlab 4	180.56	0.965	0.992	0.014
Greenlab 7	166.61	0.970	0.998	0.0002
LNAS	110.85	0.987	0.998	0.005
CERES	127.02	0.983	0.998	0.0005
PILOTE	170.51	0.969	0.997	0.0006
STICS	168.87	0.970	0.999	< 0.0001

En ce qui concerne la biomasse totale, tous les modèles fournissent des estimations biaisées. Le modèle ayant la plus petite erreur de prédiction est le modèle LNAS, même si tous les modèles fournissent globalement de bonnes prédictions, avec une erreur inférieure à 200g/m^2. Comme on peut le voir sur la Figure 1.12, cependant, les performances des modèles varient fortement au cours du temps.

Les modèles Pilote et CERES, par exemple, sont meilleurs que les autres lors de la première phase de croissance, soit jusqu'à environ 1000°C jours. Le temps d'initiation étant le même pour tous les modèles, cela signifie que la croissance de la plante se fait plus rapidement dans ces deux modèles, probablement à cause d'une croissance plus rapide de l'indice foliaire. À partir de 1000°C jours cependant, ils ont tendance à sur-estimer la biomasse totale, à cause d'une valeur élevée de la RUE. Le comportement des autre modèles est différent : malgré une initiation plus lente qui les conduit à sous-estimer la biomasse totale au début



(a) Observations vs. prédictions

(b) En fonction du temps thermique

FIG. 1.13 – Prédictions de masse racinaire en 2008.

de la croissance, une forte RUE leur permet de compenser en partie cet effet. Pour STICS, il semblerait cependant que l'efficience soit trop élevée car le modèle sur-estime largement la biomasse totale lors de la deuxième phase de croissance. Le modèle LNAS, qui possède la plus petite valeur de RUE parmi les cinq modèles, est celui qui parvient le mieux à estimer la biomasse totale, malgré une légère sous-estimation au moment de l'initiation. On observe cependant une sorte de léger « décrochage » à partir d'environ 1500°C jours, conduisant le modèle à sous-estimer la biomasse totale lors des deux dernières mesures.

Sur cet échantillon test, le modèle Greenlab à quatre paramètres donne de meilleures prédictions pour la masse sèche racinaire, mais est moins bon que la version à sept paramètres sur la masse totale, même si la différence est assez faible entre les deux modèles. Sur le modèle à 4 paramètres, l'efficience de conversion a été estimée à 5.93 g.MJ^{-1} (voir Tableau 1.5), ce qui est assez élevé pour la betterave, dont les valeurs de référence trouvées dans la littérature dépassent rarement 4 g.MJ^{-1} : Damay et Le Gouis (1993), entre 2.96 et 3.76 MJ^{-1}, Milford et Riley (1980), entre 3.16 et 4.12 g.MJ^{-1}, Biscoe et Gallagher (1977), 3.5 g.MJ^{-1}, ... De ce fait, la pente de production de biomasse, et par conséquent également celle de la masse racinaire, sont plus élevées que celles des autres modèles, ce qui est encore plus flagrant sur la courbe de masse racinaire (Figure 1.13). La différence entre les deux versions de Greenlab se verra surtout dans la section suivante, sur le jeu de données 2011.

4.2.2 Masse racinaire

En ce qui concerne la masse racinaire, les performances des modèles sont meilleures que pour la biomasse totale, sauf pour le modèle LNAS, suggérant que la masse des feuilles doit être moins bien estimée que celle de la racine (sur-estimée par exemple pour LNAS et Greenlab, et sous-estimée pour STICS). Comme nous l'avons précisé plus haut, les modèles Pilote et CERES ne contiennent pas de module permettant d'allouer la biomasse produite aux racines de façon dynamique. C'est la raison pour laquelle nous n'avons pas utilisé les mêmes critères pour comparer ces deux modèles. En revanche, à titre d'illustration, nous avons tracé la part de biomasse qui serait allouée aux racines, si l'on appliquait tout au long de la croissance de la plante l'indice de récolte empirique HI. La proportion de biomasse allouée aux racines

Tab. 1.9 – Comparaison des modèles sur la prédiction de la masse racinaire en 2008.

Modèle	Masse sèche racinaire				
	RMSEP	EF	ype	r^2	F-test (p)
Greenlab 4	91.8	0.987	3.70%	0.994	0.095
Greenlab 7	169.34	0.955	14.91%	0.999	9.10^{-6}
LNAS	180.5	0.949	17.66%	0.999	2.10^{-6}
CERES	-	-	5.70%	-	-
PILOTE	-	-	5.02%	-	-
STICS	39.5	0.998	1.92%	0.998	0.800

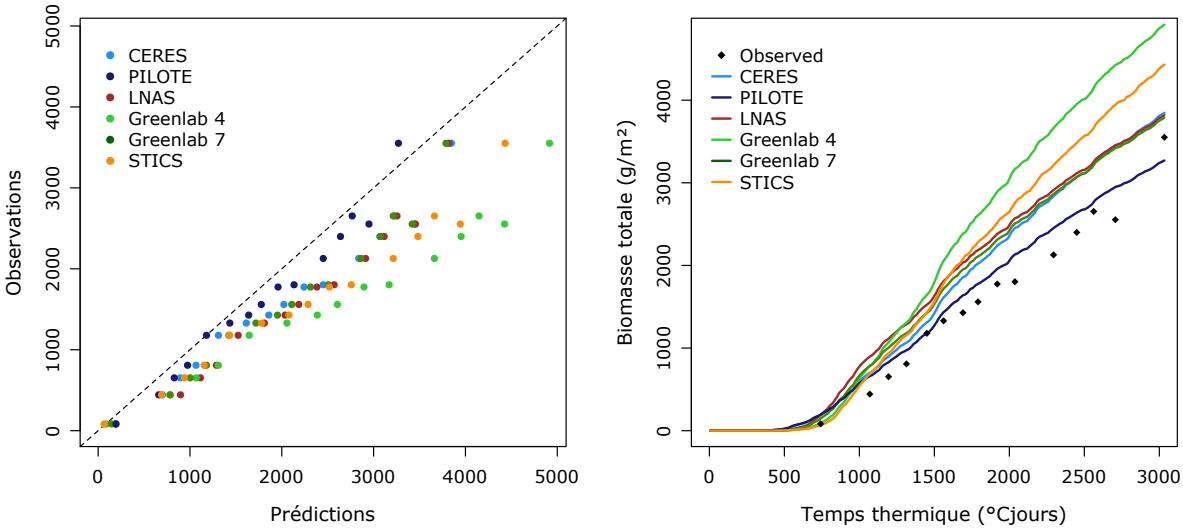
n'étant pas constante au cours du temps (Lemaire, 2010), mais croissante, c'est sans surprise que ces deux modèles sur-estimeraient la masse racinaire lors de la première phase de croissance. En revanche, les deux modèles remplissent correctement leur rôle en ce qui concerne la prédiction du rendement, avec une erreur relative de 5%, ce qui est inférieur à celle du modèle LNAS, par exemple, et comparable aux résultats du modèle Greenlab 4, mais avec une complexité moindre. Notons ici que la version de Greenlab à 4 paramètres est meilleure que celle à 7 paramètres, du fait de la forte RUE qui permet au modèle de « rattraper » son retard en fin de croissance. On peut toutefois se demander quelles auraient été ses performances si l'on avait considéré une période plus longue, et que la récolte avait eu lieu quelques semaines plus tard. Il est probable que le modèle aurait alors sur-estimé la masse racinaire.

Le modèle LNAS, s'il permettait d'obtenir de bonnes prédictions pour la biomasse totale, sous-estime largement la masse racinaire à partir de 1500°Cjour, en même temps que le « décrochage » observé sur la Figure 1.12, ce qui correspond parallèlement à une sur-estimation de la masse foliaire. Rappelons que pour ce modèle, seul un paramètre avait finalement été sélectionné pour être estimé, les autres étant fixés à des valeurs moyennes. Comme il s'agit d'un modèle qui n'a jamais été calibré auparavant, il est possible que ces valeurs moyennes aient permis une bonne calibration du modèle sur les données 2010, mais qu'elles ne correspondent pas à des valeurs correctes pour le jeu de données 2008.

Le modèle STICS fournit d'excellents résultats pour la masse racinaire, avec une erreur de prédiction de seulement 39.5 g m^{-2}, soit environ deux fois moins que la version de Greenlab à 4 paramètres, et plus de quatre fois moins que le Greenlab à 7 paramètres et que le modèle LNAS. Il est également le seul, avec le modèle Greenlab à 4 paramètres, à fournir des estimations non biaisées pour la masse racinaire. Tout ceci avec seulement un paramètre estimé, l'efficience RUE, les autres étant fixés aux valeurs de référence proposées dans Brisson et al. (2008).

4.3 Comparaison sur les données 2011 : même génotype, même site expérimental

Le jeu de données 2008 utilisé dans la section précédente correspondait à un environnement assez différent de celui de l'échantillon d'apprentissage, mais avec un génotype similaire. Dans cette section, nous nous intéressons aux capacités prédictives des modèles dans un contexte comparable à celui sur lequel ils ont été calibrés : même variété, même site expérimental. Cependant, l'épisode de grêle qui a eu lieu quelques mois après le semis et qui a provoqué d'importants dégâts sur les feuilles ne nous a pas permis d'obtenir un jeu de données satisfaisant pour tester les modèles. C'est la raison pour laquelle le jeu de données 2008 a été utilisé comme échantillon test principal, et c'est également la raison pour laquelle on s'attend à obtenir de moins bons résultats sur les données 2011, par rapport aux données 2008. Si cette intuition se révèle



(a) Observations vs. prédictions

(b) En fonction du temps thermique

FIG. 1.14 – Prédictions de biomasse totale en 2011.

TAB. 1.10 – Comparaison des modèles sur la prédiction de la masse totale en 2011.

Modèle	Masse sèche totale			
	RMSEP	EF	r^2	F-test (p)
Greenlab 4	1119.05	-0.5332	0.969	8.10^{-12}
Greenlab 7	520.8	0.67	0.969	2.10^{-7}
LNAS	583.19	0.58	0.967	7.10^{-8}
CERES	483.5	0.71	0.973	2.10^{-7}
PILOTE	227.2	0.94	0.971	6.10^{-3}
STICS	775.3	0.26	0.973	4.10^{-10}

juste dans le cas de la masse totale, les modèles n'ayant pas pu prévoir la perte de biomasse liée à la perte de masse foliaire, les prédictions de masse racinaire s'avèrent tout à fait correctes.

4.3.1 Masse totale

L'effet de la grêle se voit assez nettement sur la Figure 1.14, à partir de 1500°C jours environ, et pour une période d'environ 1000°C jours, pendant laquelle tous les modèles sur-estiment, parfois largement, la biomasse totale. Tous les modèles fournissent également des prédictions fortement biaisées. Ce résultat était attendu, car les modèles n'ont pas pu prévoir cet évènement climatique, et ont donc fourni des prédictions pour la biomasse totale qui aurait été observée en l'absence de grêle. On s'aperçoit par ailleurs que les plantes ont probablement compensé la perte de feuilles et le déficit photosynthétique qui a suivi en remobilisant une partie des assimilats de la racine vers les feuilles, car la biomasse totale observée au moment de la récolte se rapproche des prédictions fournies par les modèles.

Le modèle ayant la plus petite erreur de prédiction est le modèle Pilote, mais si les prédictions de ce modèle sont en accord avec les observations, il faut garder en tête que les données observées ne sont justement pas représentatives des conditions « normales » de croissance, et ne correspondent donc pas à ce que l'on aurait dû observer en l'absence de grêle. De ce fait, dans des conditions « normales » de

croissance, le modèle Pilote aurait sous-estimé la masse sèche totale. L'initiation est plus homogène d'un modèle à l'autre qu'en 2008, mais cela peut être dû au fait que les paramètres d'organogenèse ont été estimés conjointement sur les données 2010 et 2011, et seraient donc plus adaptés aux données 2011. Cependant, si l'initiation semble approximativement bien estimée, la croissance en biomasse totale se fait rapidement après l'initiation, ce qui conduit les modèles à sur-estimer cette variable, et ce même avant l'épisode de grêle.

Les paramètres d'efficience du modèle STICS et de la version de Greenlab à 4 paramètres sont les plus élevés, et conduisent à une sur-estimation de biomasse plus importante que pour les autres modèles. L'effet de cette RUE élevée est d'autant plus visible en 2011, probablement à cause d'un rayonnement plus important dans les conditions d'expérimentations de 2011 par rapport à celles de 2010 ou 2008. La même comparaison qu'en 2008 a été faite sur les différentes versions du modèle STICS, et nous a fait aboutir à la même conclusion : le modèle avec formulation linéaire de la production de biomasse et efficience constante fournit les meilleurs prédictions, et l'ajout du stress thermique ne permet pas d'améliorer significativement les résultats.

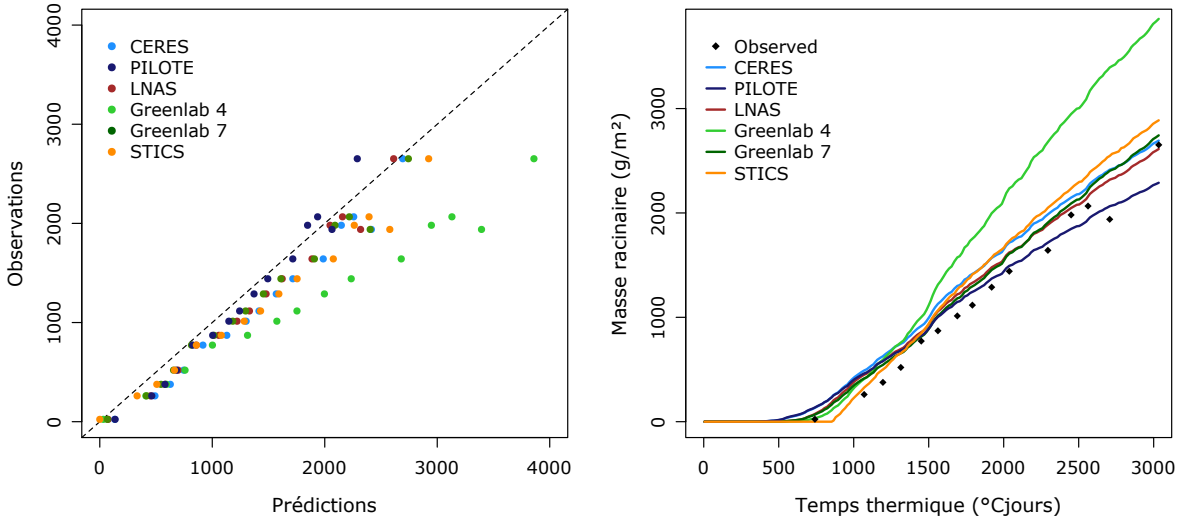
L'un des résultats importants de cette section concerne la différence entre les deux versions du modèle Greenlab, qui se comportaient de façon plus comparable sur les données 2008. Ici, le modèle où seulement 4 paramètres ont été estimés se comporte beaucoup moins bien que la version à 7 paramètres. Les trois paramètres supplémentaires qui ont été estimés dans cette version, par rapport à la première, sont les paramètres de forces de puits des pétioles p_p et q_p, et le paramètre a_b d'allocation pour les limbes. L'ajout de ces trois paramètres dans le processus de calibration a modifié l'estimation des 4 premiers, et a également permis de modifier substantiellement les valeurs qu'avaient ces trois paramètres lorsqu'ils étaient fixés à des valeurs de référence. En particulier, comme nous l'avons déjà vu, l'efficience est plus élevée dans le modèle à 4 paramètres.

4.3.2 Masse racinaire

Concernant la masse racinaire, comme elle n'a pas été impactée directement par l'épisode de grêle, les modèles parviennent à la prédire plus correctement, même s'ils fournissent également des prédictions biaisées, car sur-estimées.

Si les différences entre les modèles sont moins visibles que sur la biomasse totale, en particulier pour le modèle STICS dont les prédictions sont plus consistantes sur la masse racinaire, le modèle Greenlab à 4 paramètres continue à se distinguer des autres. Le modèle LNAS et le modèle Greenlab à 7 paramètres ont un comportement similaire et l'erreur de prédiction associée aux deux modèles est quasiment identique, même si le modèle LNAS prédit le rendement de la culture avec le plus petit taux d'erreur relative (1.5%). Il est intéressant de noter que le modèle CERES, malgré l'utilisation d'un indice de récolte empirique pour la répartition de la biomasse aux racines, qui n'est pas censé être valide tout au long de la croissance de la plante, se comporte de façon similaire aux deux modèles pré-cités, LNAS et Greenlab 7. Il parvient également à estimer correctement le rendement de la culture, avec une erreur relative de prédiction d'un peu plus de 1.5%.

Le modèle Pilote, comme pour la prédiction de biomasse totale, sous-estimerait la masse racinaire si une proportion constante et égale à HI était allouée à la racine. La sous-estimation de biomasse totale induit une mauvaise prédiction de rendement par ce modèle, avec une erreur relative de prédiction d'environ 14%.



(a) Observations vs. prédictions

(b) En fonction du temps thermique

FIG. 1.15 – Prédictions de masse racinaire en 2011.

TAB. 1.11 – Comparaison des modèles sur la prédiction de la masse racinaire en 2011.

Modèle	Masse sèche racinaire				
	RMSEP	EF	ype	r^2	F-test (p)
Greenlab 4	772.2	-0.116	45.5%	0.985	7.10^{-13}
Greenlab 7	190.16	0.932	3.45%	0.986	4.10^{-5}
LNAS	187.98	0.934	1.49%	0.983	1.10^{-4}
CERES	-	-	1.54%	-	-
PILOTE	-	-	13.68%	-	-
STICS	275.75	0.857	8.88%	0.984	6.10^{-7}

Enfin, les résultats du modèle STICS sont moins bons en 2011 qu'en 2008 sur la masse racinaire. Cet effet est probablement dû à la sur-estimation de biomasse totale due à une trop forte efficience, qui a eu moins de répercussions sur les données 2008. En effet, on s'aperçoit que les premiers points sont correctement prédits par STICS, de par une initiation plus lente que pour les autres modèles ; mais, la valeur élevée de la RUE entraîne une pente trop forte et une sur-estimation de la masse racinaire à partir de 1000°C jours environ.

5 Conclusion et perspectives

Le travail présenté dans ce chapitre constitue une première tentative de « benchmarking » dans un domaine où de nombreux modèles co-existent, et propose également une méthodologie pour l'évaluation et l'utilisation de modèles dans un cadre prédictif, et s'inscrit dans la démarche de « bonnes pratiques » de modélisation défendue par Vos et al. (2007).

Il est en effet essentiel, lorsque l'on utilise un modèle dans un objectif donné, de s'assurer qu'il peut remplir cet objectif. Ainsi, un modèle purement descriptif sera souvent moins robuste qu'un modèle spécifiquement dédié à la prédiction. Dans le cadre de notre travail, nous avons donc proposé une approche

permettant de construire, à partir d'un modèle de croissance de plantes, une version robuste et adaptée à la prédiction. Dans un premier temps, les paramètres les plus influents du modèle sont identifiés à l'aide d'une analyse de sensibilité, puis dans un second temps, le nombre de paramètres à estimer est déterminé à l'aide de critères de sélection du type AIC ou BIC. Si cette méthode permet de réduire la variabilité des modèles, elle possède toutefois certains inconvénients inhérents aux méthodes utilisées :

- l'utilisation d'une analyse de sensibilité requiert le choix de lois de probabilités pour modéliser la variabilité de chaque paramètre, or ces informations peuvent être difficiles à obtenir, en particulier lors de la construction d'un nouveau modèle. Dans notre cas par exemple, le modèle LNAS donne de bonnes prédictions pour la biomasse totale en 2008, mais pas pour la masse racinaire sur la même période. Cela peut-être dû à une mauvaise spécification des distributions d'entrée des paramètres.
- la méthode d'analyse de sensibilité présentée dans ce chapitre repose sur des simulations de type Monte-Carlo, qui nécessitent un temps et une puissance de calcul qui peuvent être rédhibitoires, notamment lorsque le modèle est complexe. L'utilisation d'un mésocentre de calcul nous a permis d'obtenir des résultats pour STICS ou Greenlab en une vingtaine de minutes, mais les temps de calcul peuvent être beaucoup plus longs sur un ordinateur moins puissant, même si la méthode développée par Wu et al. (2011) permet de diminuer les coûts. Notons tout de même que des méthodes moins gourmandes peuvent être utilisées si le nombre de paramètres est trop grand, ou le modèle trop complexe (Cariboni et al., 2007), comme la méthode de Morris (1991).

En ce qui concerne l'évaluation des qualités prédictives des modèles, celle-ci doit se faire à l'aide de critères adaptés, sur un échantillon de données *indépendant* de celui sur lequel les modèles ont été paramétrés. Si un tel jeu de données n'est pas disponible, des méthodes de ré-échantillonnage devront être envisagées. Il convient également de bien choisir les critères d'évaluation. Les critères présentés ici sont assez utilisés dans la littérature, et sont faciles à obtenir et à interpréter.

L'analyse de sensibilité que nous avons réalisée ici ne tient pas compte des éventuelles corrélations qui peuvent exister entre les paramètres. Des phénomènes de compensation peuvent également intervenir, par exemple, dans le cas du modèle Greenlab, entre les paramètres de la loi bêta impliquée dans le processus d'allocation de biomasse. Dans le cas du maïs notamment, Ma et al. (2008) fixent l'un des deux paramètres et estiment la quantité a_o/a_o+b_o, qui correspond à l'espérance d'une loi bêta de paramètres (a_o, b_o), et qui semble plus stable. Nous n'avons pas rencontré ce problème sur les données betterave dont nous disposions, mais il peut toutefois se poser lors de l'extrapolation à d'autres plantes. D'une façon plus générale, les corrélations entre paramètres peuvent influencer les résultats de l'analyse de sensibilité si elles ne sont pas correctement prises en compte. Des travaux sont menés en ce sens par Wu et al. (2013) et pourraient permettre d'améliorer les résultats.

Enfin, ce travail est une modeste contribution à l'évaluation des capacités prédictives de différents modèles de croissance de plantes pour la betterave sucrière, et mériterait d'être approfondi sur un plus large échantillon de jeux de données test, en comparant les modèles sur des génotypes différents, dans des conditions de stress environnementaux, ... Malgré tout, plusieurs résultats intéressants émergent de ce travail. Par exemple, la formulation linéaire de la production de biomasse dans le modèle STICS permet d'obtenir de meilleurs résultats que la version originale du modèle. Le modèle STICS s'avère également très robuste et a fourni de très bonnes prédictions sur le jeu de données 2008.

Les modèles avec répartition empirique de la biomasse à l'aide de l'indice de récolte HI sont globalement corrects pour la prédiction de rendement, sur les données 2008 et 2011 pour CERES, mais seulement sur les données 2008 pour Pilote. Cependant, cet indice de récolte a été calculé sur l'échantillon d'apprentissage, et s'est avéré plus faible que celui que l'on peut trouver dans la littérature, et également plus faible

que celui calculé sur les données tests, suggérant que ce paramètre n'est pas très robuste. Dürr et al. (2003) et Lemaire (2010) suggèrent en effet que la répartition de la biomasse entre les différents compartiments d'organes varient en fonction des conditions environnementales et des traitements (engrais) utilisés. Les performances de ces modèles sur la prédiction du rendement peuvent donc être affectées en cas de mauvaise estimation de l'indice de récolte. En revanche, leurs performances sur la masse sèche totale sont comparables à celles des autres modèles. Si ces modèles s'avèrent corrects en conditions non limitantes et sur les variables étudiées ici, ils peuvent toutefois s'avérer limités si l'on s'intéresse par exemple à l'évolution du profil foliaire, ou plus généralement à des variables résultant de l'interaction entre la structure et le fonctionnement de la plante, qui sont par contre accessibles par les modèles de type structure-fonction.

Pour le modèle Pilote, nous avons utilisé une approximation du LAI à l'aide des mesures de masses foliaires comme précisé en Section 2.1, alors que cet indice est normalement obtenu à l'aide d'appareils appelés LAI-mètres, comme par exemple le Licor LAI-2000. Il est probable que les performances du modèle sur la biomasse totale auraient pu être améliorées avec ce type de mesures. De la même façon, les surfaces foliaires maximales nécessaires au modèle CERES ont été obtenues à partir des masses individuelles des limbes, en divisant par la masse surfacique associée. Ces données peuvent bien sûr être très variables d'une année sur l'autre, et d'une variété à l'autre, et même, peuvent ne pas être disponibles, ce qui peut rendre impossible l'utilisation de ce modèle.

Les résultats obtenus, et les difficultés auxquelles nous avons dû faire face montrent également l'importance cruciale du temps thermique d'initiation. Lors des premiers essais de prédiction, où nous avions utilisé les mêmes temps d'initiation que sur l'échantillon d'apprentissage, les performances des modèles étaient médiocres : malgré une pente qui semblait correcte, un décalage au niveau de l'initiation entraînait soit une sous-estimation soit une sur-estimation de la biomasse totale ou racinaire. C'est la raison pour laquelle nous avons décidé d'adapter le temps d'initiation au jeu de données utilisé. D'un point de vue purement prédictif, ce n'est bien sûr pas satisfaisant, car nous aimerions pouvoir utiliser les modèles sans avoir besoin de retoucher à certains paramètres. De plus, la méthodologie que nous avons utilisée n'est pas entièrement satisfaisante, car elle a nécessité un recalage manuel du nuage de points de 2011 afin d'obtenir des résultats cohérents avec les observations. Deux observations peuvent être faites à ce sujet : d'une part, l'utilisation d'une plante moyenne n'est en soi pas satisfaisante, car la variabilité entre plantes n'est pas prise en compte, et que cela pose justement les problèmes auxquels nous avons dû faire face avec les modèles individus-centrés, et d'autre part, certains mécanismes pouvant retarder la germination ou l'émergence de la plante ne sont pas pris en compte dans le modèle d'organogenèse que nous avons utilisé. Une piste à explorer pour améliorer la prise en compte de l'initiation serait d'utiliser des modèles spécifiques permettant d'estimer ce paramètre en fonction des conditions climatiques et des caractéristiques du sol. Dürr et al. (2001) ont notamment proposé un modèle permettant de prédire l'émergence des graines de betterave sucrière en fonction de la texture du sol, de sa température, des précipitations, des caractéristiques de la graine, et des caractéristiques de semis, et ont obtenu des résultats satisfaisants, avec moins de 10% d'écart entre prédictions et observations, ce qui correspond à des écarts compris entre 15 et 30°C jours. Des modèles empiriques, plus simples à utiliser, ont également été développés (Forcella et al., 2000). Un module dédié à l'émergence et à l'initiation existe également dans STICS, prenant en compte l'hydratation des semences, la germination, la vitesse d'élongation de la racine et de la jeune pousse, et la texture du sol. Ces deux approches pourraient, soit être mises en place en amont de l'étape de calibration et fournir le temps d'initiation comme nous l'avons fait ici, soit être intégrées directement dans les modèles. Quoiqu'il en soit, une attention particulière est requise pour les premières phases de croissance de la plante, que ce soit en termes de germination, d'émergence, ou même pour le développement de la jeune plante lors des

premiers cycles de croissance, puisque une mauvaise prise en compte de cette période peut déterminer de bonnes ou de mauvaises performances du modèle pour la suite.

L'importance du temps d'initiation a également été observée par Feng et al. (2014), dans une étude récente sur la variabilité individuelle des populations de maïs, basée sur la comparaison d'un modèle agronomique (Pilote) et d'un modèle structure-fonction (Greenlab). Un décalage entre les courbes produites par les deux modèles a été observé par les auteurs, ce qui les a conduits à étudier la variabilité du temps d'émergence. L'hypothèse des auteurs est la suivante : si la population de l'étude était composée de plantes homogènes, les deux modèles convenablement mis à la même échelle [2] fourniraient des simulations équivalentes. Les différences observées entre simulations seraient donc imputables à l'hétérogénéité de la population. Ils ont notamment montré qu'en supposant une distribution de type binomiale négative pour la date d'émergence (exprimée en cycle de croissance), et en intégrant cette variabilité dans le modèle Greenlab, il était possible de neutraliser l'écart initialement observé entre les deux courbes.

Enfin, les modèles ont été calibrés sur des données individuelles provenant de plusieurs plantes, et résumées en une « plante moyenne ». La grande variabilité qui existe entre les plantes n'est pas prise en compte dans ces modèles, alors qu'elle peut être source d'une grande variabilité sur les prédictions du modèle. L'extrapolation du modèle à un contexte différent de celui sur lequel il a été construit peut ainsi s'avérer délicate, en particulier si la population sur laquelle on souhaite évaluer le modèle est trop éloignée de celle sur laquelle il a été construit. Une façon d'améliorer les performances des modèles peut être de prendre en compte cette variabilité entre plantes, afin d'obtenir non pas une seule valeur en sortie du modèle, mais plutôt la distribution des sorties du modèle. Ceci fera l'objet des chapitres suivants, dans lesquels nous présenterons le modèle d'organogenèse qui a été utilisé dans ce chapitre, ainsi qu'une extension du modèle Greenlab à l'échelle de la population.

2. les auteurs adoptent une démarche différente de celle que nous avons proposée, puisqu'ils adaptent l'échelle et le pas de temps du modèle Pilote à ceux du modèle Greenlab

Chapitre 2

Généralités sur les modèles non linéaires mixtes

"Every plant is an individual.
Wrong again. We are not individuals at all, we are all connected.
We are individuals the way each blossom on an apple tree is an individual."

Dale Pendell, *Pharmako/Poeia : Plant Powers, Poisons, and Herbcraft.*

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. Brouwer et al. (1993) ont notamment montré comment cette double variabilité pouvait contribuer à améliorer le rendement d'une culture en cas de fortes sécheresses, certaines portions du champ étant alors plus résistantes au stress hydrique et pouvant compenser les mauvaises performances obtenues dans d'autres parties du champ. De même, une population de plantes dont la période de floraison n'est pas synchronisée sera moins susceptible de subir les effets d'un stress hydrique ponctuel ou d'une attaque d'insectes (Renno et Winkel, 1996).

Ces deux exemples simples montrent bien, d'une part, l'importance d'une telle variabilité, et d'autre part, la nécessité d'en tenir compte dans les modèles de croissance de plantes à cause de l'impact qu'elle peut avoir à l'échelle agronomique. Cependant, l'extrapolation des modèles individus-centrés à l'échelle de la population n'est pas si immédiate. Les premières tentatives dans cette direction concernent principalement la compétition pour la lumière (voir par exemple Fournier et Andrieu (1999) ; Cournède et al. (2008)), et la croissance de chaque plante y est simulée individuellement, en tenant compte d'un indice de compétition entre plantes. Si l'intérêt de ce type d'approches est clair d'un point de vue théorique, pour mieux comprendre et mieux décrire le fonctionnement d'une plante au sein d'une population, leur application s'avère délicate dans la pratique car il est en général impossible de décrire de façon exhaustive tous les individus d'une population.

L'utilisation de modèles de population *stochastiques*, qui s'attachent à décrire la distribution de caractéristiques individuelles dans la population étudiée apparaît alors comme une bonne alternative pour contourner ces difficultés. L'une des approches possibles est basée sur les modèles à effets mixtes, qui permettent de prendre en compte les variabilités intra- et inter-individuelles. Ils sont largement répandus dans certaines disciplines comme la pharmacodynamique (Comets et al., 2007 ; Beal et Sheiner, 1982), l'épidémiologie (Lavielle et al., 2010 ; Morrell et al., 1995), ou l'écologie (Bolker et al., 2009). On les retrouve

également en agronomie ou en foresterie, essentiellement à des fins descriptives, pour étudier par exemple la variabilité des relations allométriques (Dietze et al., 2008 ; Vieilledent et al., 2010 ; Courbaud et al., 2012) ou de la hauteur des arbres (Hall et Bailey, 2001 ; Nothdurft et al., 2006) ou pour étudier l'effet de l'azote sur le rendement (Makowski et Lavielle, 2006), ou la teneur finale en protéines de la plante (Casagrande et al., 2009). En revanche, aucune application aux modèles de croissance de plantes n'a encore été réalisée, à notre connaissance.

Avant d'appliquer cette approche à la modélisation de la variabilité inter-plantes, nous présentons dans ce chapitre quelques éléments bibliographiques nécessaires à la compréhension du chapitre suivant. En section 1 nous définissons le modèle mixte qui servira de base aux applications du chapitre 3, puis en section 2 les différentes méthodes d'estimation disponibles. En particulier, nous définirons plus en détails les algorithmes MCMC-EM (section 2.3) et SAEM (section 2.4).

N.B. *Nous utiliserons dans les sections suivantes la notation générique $f(\,\cdot\,;\theta)$ pour désigner chaque densité de probabilité considérée, paramétrée par θ. Les variables aléatoires associées seront distinguées par le premier argument de la fonction.*

1 Formulation du modèle

Les modèles à effets mixtes sont une extension des modèles à effets fixes, dans lesquels certains paramètres sont considérés comme aléatoires. Ils ont d'abord été introduits dans le cas linéaire par Laird et Ware (1982), qui proposent d'utiliser l'algorithme Espérance-Maximisation (EM) introduit un peu plus tôt par Dempster et al. (1977) pour estimer les paramètres du modèle. Depuis, ils ont connu un développement rapide grâce aux progrès constants en matière de puissance informatique, rendant possible l'utilisation d'algorithmes d'estimation de type Monte Carlo. L'extension aux modèles linéaires mixtes généralisés puis aux modèles non linéaires a été proposée respectivement par Breslow et Clayton (1993) et Lindstrom et Bates (1990).

Ils sont particulièrement adaptés aux cas où l'on dispose de mesures répétées chez plusieurs individus d'une même population, car ils permettent de prendre en compte les deux sources de variabilité provenant de ce type d'observations : la variabilité intra-individuelle, c'est-à-dire la façon dont varient les mesures d'un même individu, et la variabilité inter-individuelle. La motivation principale derrière l'utilisation de ces modèles répond au principe selon lequel chaque profil individuel est le résultat d'une variation aléatoire autour d'une courbe moyenne. Plus précisément, un même modèle permet d'expliquer les variations intra-individuelles, mais les valeurs de certains paramètres varient d'un individu à l'autre.

Nous notons y_{ij} l'observation faite sur l'individu i sous la condition t_{ij}, avec $i = 1, \ldots, s$ et $j = 1, \ldots, n_i$, où s représente donc la taille de l'échantillon et n_i le nombre d'observations de l'individu i. Dans le cas de données longitudinales, t_{ij} peut représenter le temps auquel a été faite l'observation y_{ij}. Le modèle non linéaire mixte peut être vu comme un modèle hiérarchique à deux niveaux (Davidian et Giltinan, 1995). Dans un premier temps, on mesure la variabilité *intra-individuelle* :

$$y_{ij} = g(t_{ij}, \phi_i) + \varepsilon_{ij}, \tag{2.1}$$

où les termes d'erreur (ε_{ij}) sont i.i.d., de moyenne nulle et de variance σ^2, et où ϕ_i est un vecteur de paramètres spécifiques à l'individu i. La fonction g correspond à la relation *non linéaire* entre les paramètres individuels et les observations, par l'intermédiaire des conditions t_{ij}. Elle représente la façon dont évoluent

les mesures d'un même individu, en moyenne. Dans le cadre général proposé par Davidian et Giltinan (1995), la distribution des termes d'erreur ε_{ij} peut être spécifiée (par exemple, une loi normale), mais il est également possible de ne faire aucune hypothèse particulière sur cette loi, et d'utiliser des méthodes non paramétriques.

Dans une deuxième étape, on s'intéresse à la variabilité *inter-individuelle*, en supposant que les vecteurs de paramètres ϕ_i définis en (2.1) sont des vecteurs aléatoires i.i.d. dont on cherche à caractériser la loi :

$$\phi_i = A_i\beta + \xi_i, \quad \text{Var } \xi_i = \Gamma, \tag{2.2}$$

où β est le vecteur des effets fixes, A_i est une matrice de design supposée connue, et où Γ permet de rendre compte des corrélations entre les paramètres d'un même individu. La matrice A_i dépend de caractéristiques connues de l'individu i, par exemple, la variété, la dose d'azote reçue, ... et permet par exemple d'introduire un effet moyen différent en fonction des caractéristiques de l'individu. Nous verrons plus loin (chapitre 3, section 1) la façon dont cette matrice peut être spécifiée.

Sous les hypothèses précédentes, et en notant $y = (y_{i,j}, 1 \leq i \leq s, 1 \leq j \leq n_i)$, $\phi = (\phi_1, \ldots, \phi_s)^t$ et $\theta = (\beta, \Gamma, \sigma^2)$ le vecteur contenant tous les paramètres du modèle, on a

$$f(y \mid \phi\,; \theta) = \prod_{i=1}^{s} \prod_{j=1}^{n_i} f(y_{ij} \mid \phi_i\,; \theta), \tag{2.3}$$

et

$$f(\phi; \theta) = \prod_{i=1}^{s} f(\phi_i; \theta), \tag{2.4}$$

où $f(y_{ij} \mid \phi_i\,; \theta)$ est la densité de probabilité des y_{ij}, et $f(\phi_i; \theta)$ la densité de probabilité des ϕ_i.

Dans la suite, nous supposerons que les termes d'erreurs ε_{ij} suivent une loi normale centrée et de variance σ^2, et que les effets aléatoires ξ_i suivent également une loi normale, de moyenne nulle et de matrice de covariance Γ.

2 Estimation dans le modèle non linéaire mixte

Le vecteur de paramètres est $\theta = (\beta, \Gamma, \sigma^2)$, et peut être estimé par la méthode du maximum de vraisemblance, qui fournit sous certaines conditions de régularité, des estimateurs convergents, asymptotiquement normaux et asymptotiquement efficaces. Nous reviendrons en section 2.2.2 sur ces conditions. La vraisemblance des observations $L(\theta) := f(y; \theta)$ peut s'écrire en fonction de la vraisemblance complète $L_c(\theta; \phi) := f(y, \phi; \theta)$:

$$L(\theta) = \int_{\mathbb{R}^{P \times s}} L_c(\theta; \phi) d\phi = \int_{\mathbb{R}^{P \times s}} f(y \mid \phi; \theta) f(\phi; \theta) d\phi. \tag{2.5}$$

Dans le cadre du modèle défini en (2.1) et (2.2), la non linéarité de la fonction g rend en général le calcul de cette intégrale impossible analytiquement. Différentes approches ont alors été développées depuis les années 1990, et sont présentées dans les sections suivantes. Nous renvoyons le lecteur à Davidian et Giltinan (1995) pour une revue plus détaillée des méthodes existantes.

2.1 Les différentes approches

2.1.1 Méthodes basées sur l'estimation des paramètres individuels

La première approche est basée sur l'estimation individuelle de chaque paramètre ϕ_i (Davidian et Giltinan, 1993, 1995). La première étape consiste à estimer pour chaque individu i, le vecteur de paramètres ϕ_i, les méthodes classiques de type moindres carrés ordinaires ou généralisés pouvant être utilisées. On obtient ainsi s vecteurs de paramètres estimés $\hat{\phi}_1, \ldots, \hat{\phi}_s$, qui vérifient donc $\hat{\phi}_i = \phi_i + e_i, i = 1, \ldots, s$. Dans le cas où la dépendance entre les ϕ_i et les ξ_i est linéaire, la deuxième étape correspond alors au modèle linéaire mixte $\hat{\phi}_i = A_i\beta + \xi_i + e_i, i = 1, \ldots, s$, pour lequel des méthodes d'estimation fiables sont disponibles dans la plupart des logiciels de traitement statistique. Cependant, les estimateurs obtenus au cours de la première étape étant asymptotiquement consistants, le nombre d'observations par sujet n_i doit être suffisamment grand.

2.1.2 Méthodes basées sur une approximation de la vraisemblance

Lorsque le nombre d'observations par individu est trop faible, et comme c'est le cas dans la plupart des problèmes non linéaires, une autre approche consiste à linéariser la fonction g, et à appliquer ensuite les méthodes disponibles dans le cas linéaire. Plusieurs méthodes ont été proposées, parmi lesquelles la méthode FO (First-Order approximation) introduite par Beal et Sheiner (1982), et la méthode FOCE (First-Order Conditional Estimation) introduite par Lindstrom et Bates (1990). Dans la méthode FO, Beal et Sheiner (1982) proposent un développement de Taylor d'ordre 1 autour du point $\xi_i = 0$, en supposant le terme $\xi_i\varepsilon_{ij}$ négligeable . La fonction g ainsi linéarisée, la densité conditionnelle $f(y \mid \phi_i; \theta)$ peut donc être approchée par la densité d'un vecteur gaussien dont l'espérance et la matrice de covariance dépendent linéairement de ϕ_i. Lorsque la densité $f(\phi_i; \theta)$ est également gaussienne, l'intégrale (2.5) devient calculable analytiquement. Cette méthode est notamment implémentée dans le logiciel NONMEM®, principalement utilisé dans le domaine de la pharmacodynamique et de la pharmacocinétique, et dans la procédure NLMIXED de SAS®. Néanmoins, l'approximation FO peut s'avérer médiocre, et fournir des résultats biaisés (Vonesh, 1992 ; Davidian et Giltinan, 1995).

Dans la méthode FOCE, proposée par Lindstrom et Bates (1990) comme une amélioration de la méthode FO, la linéarisation du modèle ne se fait plus autour du point $\xi_i = 0$, mais autour du mode a posteriori de ξ_i, correspondant au meilleur prédicteur linéaire non biaisé dans le cas du modèle linéaire mixte. Elle est également implémentée dans NONMEM® et dans SAS®, ainsi que dans la fonction `nlme` de R. Si cette méthode fournit de meilleurs estimateurs que la méthode FO (Vonesh, 1992), l'approximation sur laquelle elle repose peut s'avérer mauvaise, en particulier lorsque l'hypothèse de normalité n'est pas vérifiée.

2.1.3 Méthodes « exactes »

Face aux différents inconvénients des méthodes précédentes, des méthodes dites « exactes », qui ne nécessitent pas d'approximation du modèle, ont été développées et rendues accessibles grâce aux progrès informatiques. L'avantage de ce type d'approches est qu'elles reposent sur des approximations numériques dont la précision peut être choisie aussi fine qu'on le souhaite, en fonction des ressources informatiques disponibles, contrairement aux approximations analytiques dont la précision dépend du modèle sous-jacent.

Des méthodes déterministes de type quadrature de Gauss peuvent par exemple être utilisées (Davidian et Gallant (1993) ; Pinheiro et Bates (1995) (section 2.4)) pour approcher l'intégrale (2.5), mais s'avèrent vite coûteuses en temps de calcul et leur complexité augmente avec le nombre de paramètres aléatoires. De la même façon, des méthodes stochastiques de type Monte-Carlo, comme l'échantillonnage d'importance (« importance sampling », introduit par Marshall (1956)) (voir également Pinheiro et Bates (1995), section 2.3. pour plus de détails sur les méthodes exactes) peuvent également être utilisées.

2.1.4 Méthodes exactes basées sur l'utilisation de l'algorithme EM

Une autre approche consiste à voir le modèle mixte défini en (2.1) et (2.2) comme un problème de *données incomplètes*, en considérant les effets aléatoires ϕ_i comme des données « manquantes », ou non observées. Ainsi défini, le modèle mixte peut être traité à l'aide d'une variante appropriée de l'algorithme d'Espérance-Maximisation (algorithme EM, « Expectation-Maximization ») introduit par Dempster et al. (1977) dans sa formulation générale. L'avantage de cette approche est que l'on ne travaille plus avec la densité marginale des observations $f(y;\theta)$ mais avec la densité jointe $f(y,\phi;\theta)$, dont on connaît l'expression analytique. Nous présentons dans la section 2.2 les principes généraux de l'algorithme EM, puis les extensions de l'algorithme qui seront utilisées dans le cadre du modèle non linéaire mixte dans les sections 2.3 et 2.4.

N.B. *Nous noterons dans la suite $x = (y,\phi)$, le vecteur des données complètes, y correspondant aux données observées et ϕ aux données manquantes.*

2.2 L'algorithme EM

L'algorithme EM est un algorithme itératif dont l'objectif est d'obtenir le maximum de vraisemblance d'un modèle dans lequel certaines données ne sont pas observées et sont donc considérées comme manquantes. Il est particulièrement adapté aux cas où la vraisemblance des données complètes s'écrit plus simplement que la vraisemblance des données observées, et repose sur l'idée suivante : lorsque l'on se trouve en présence de données manquantes, une première intuition est d'estimer ou de remplacer ces données manquantes, puis d'estimer les paramètres du modèle à l'aide des données « augmentées ». Chaque itération de l'algorithme se divise alors en deux étapes, l'une dite « Espérance » qui consiste à calculer l'espérance conditionnelle de la log-vraisemblance des données complètes sous la loi des données non observées sachant les observations à l'itération courante, et une seconde étape de « Maximisation » dans laquelle on maximise l'espérance conditionnelle obtenue lors de la première étape. Ces deux étapes seront détaillées plus loin.

L'article fondateur est celui de Dempster, Laird, et Rubin (1977), dans lequel sont énoncés les principes généraux, et qui a donné son nom à l'algorithme, même si d'autres auteurs ont développé avant eux des algorithmes similaires, mais dans des cas particuliers. Ainsi, la plus ancienne référence à un algorithme de type EM revient à Newcomb (1886) pour l'estimation des paramètres d'un modèle de mélange gaussien. Plus tard, Orchard et Woodbury (1972) définissent le « missing information principle », et établissent le lien entre la vraisemblance complète et la vraisemblance incomplète. Sundberg (1974) propose également une étude détaillée de l'algorithme dans le cas particulier des modèles appartenant à la famille exponentielle. La convergence de l'algorithme sous des conditions générales de régularité a été démontrée par Dempster et al. (1977) ; Wu (1983) ; Boyles (1983). De nombreuses extensions ont été proposées depuis, comme par exemple l'algorithme ECM (Meng et Rubin, 1993), pour le cas où l'étape de maximisation ne peut

pas se résoudre explicitement, et est remplacée par une succession de maximisations conditionnelles. Wei et Tanner (1990) proposent également une extension de l'algorithme dans le cas où l'étape E n'est pas explicite et est remplacée par une approximation de type Monte-Carlo. Le lecteur intéressé pourra se référer par exemple à l'ouvrage de McLachlan et Krishnan (2007). Nous discuterons plus loin deux extensions de l'algorithme dans le cas où l'étape E n'est pas explicite, à l'aide de l'algorithme MCMC-EM (section 2.3) et de l'algorithme SAEM (section 2.4).

La convergence de la séquence produite par l'algorithme EM vers le maximum de vraisemblance n'est pas garantie, et en général, dans la plupart des applications, la convergence a lieu vers un point stationnaire de la vraisemblance, qui peut être un maximum local ou global, ou un point-selle. Sous certaines conditions supplémentaires de régularité, Wu (1983) a montré que l'on peut s'assurer de la convergence vers un maximum local. Cependant, ces conditions peuvent être difficiles à vérifier dans la pratique, et l'algorithme peut alors se retrouver bloqué à un point stationnaire de la vraisemblance qui ne soit ni un maximum global, ni même un maximum local. Dans ces cas-là, une perturbation aléatoire du vecteur de paramètres peut permettre à l'algorithme de s'en éloigner McLachlan et Krishnan (2007). Il s'agit d'un des avantages des versions stochastiques de l'algorithme EM.

Nous présentons ici la formulation générale de l'algorithme EM. En partant d'une valeur initiale $\theta^{(0)}$, l'itération $k+1$ de l'algorithme consiste à réaliser successivement les deux étapes suivantes :

Étape E : on évalue l'espérance conditionnelle de la log-vraisemblance des données complètes sous la distribution des données manquantes ou cachées sachant les observations, et sous l'estimation courante de θ^k (appelée Q ou Q-function en anglais) :

$$Q(\theta;\theta^k) = \mathbb{E}\left(\log f(x;\theta) \mid y;\theta^k\right). \tag{2.6}$$

Étape M : on maximise la fonction Q par rapport à θ, et on met à jour le vecteur de paramètres de la façon suivante :

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} Q(\theta;\theta^k). \tag{2.7}$$

Justification de l'algorithme

Notons $\ell(\theta) := \log L(\theta)$ la log-vraisemblance des données incomplètes, et Θ l'espace des paramètres du modèle. On suppose pour simplifier que $\forall\theta \in \Theta$, les densités considérées sont strictement positives. On a alors :

$$\ell(\theta) = \log f(y;\theta) = \log f(x;\theta) - \log f(\phi \mid y;\theta). \tag{2.8}$$

On intègre ensuite de chaque côté de l'équation par rapport à la loi conditionnelle de ϕ sachant y sous l'itération courante θ^k. Le terme de gauche ne dépend pas de ϕ et correspond donc à une variable aléatoire constante, et on a :

$$\int \log f(y;\theta) f(\phi \mid y;\theta^k) d\phi = \int \log f(x;\theta) f(\phi \mid y;\theta^k) d\phi - \int \log f(\phi \mid y;\theta) f(\phi \mid y;\theta^k) d\phi$$

$$\log f(y;\theta) = \mathbb{E}_{\theta^k}(\log f(x;\theta) \mid y) - \mathbb{E}_{\theta^k}(\log f(\phi \mid y;\theta) \mid y)$$

$$\ell(\theta) = Q(\theta;\theta^k) + H(\theta;\theta^k),$$

où Q est la fonction intervenant dans l'algorithme EM et où H vérifie l'inégalité suivante :

$$
\begin{aligned}
H(\theta;\theta^k) - H(\theta^k;\theta^k) &= -\mathbb{E}_{\theta^k}(\log f(\phi \mid y;\theta) - \log f(\phi \mid y;\theta^k) \mid y) \\
&= -\mathbb{E}_{\theta^k}\left(\log \frac{f(\phi \mid y;\theta)}{f(\phi \mid y;\theta^k)} \mid y\right) \\
&\geq -\log \mathbb{E}_{\theta^k}\left(\frac{f(\phi \mid y;\theta)}{f(\phi \mid y;\theta^k)} \mid y\right) \\
&= -\log \int \frac{f(\phi \mid y;\theta)}{f(\phi \mid y;\theta^k)} f(\phi \mid y;\theta^k) d\phi \\
&= -\log \int f(\phi \mid y;\theta) d\phi \\
&= 0
\end{aligned}
$$

où le passage de la deuxième à la troisième ligne se fait grâce à l'inégalité de Jensen et à la concavité de la fonction logarithmique. En particulier, on a $H(\theta;\theta^k) = H(\theta^k;\theta^k)$ si et seulement si $\theta = \theta^k$.

À chaque itération k de l'algorithme EM, on est assuré d'augmenter la fonction H, et grâce à l'étape M, on augmente également la valeur de Q. De ce fait, la valeur de la log-vraisemblance et donc de L est augmentée à chaque itération.

Dans le cas du modèle hiérarchique simple présenté ci-dessus, on peut décomposer le vecteur de paramètres en deux sous-vecteurs $\theta_1 = (\beta, \Gamma)$ et $\theta_2 = \sigma^2$. On obtient alors la décomposition suivante pour la fonction Q :

$$
\begin{aligned}
Q(\theta;\theta^k) &= \mathbb{E}\left(\log f(\phi;\theta_1) \mid y;\theta^k\right) + \mathbb{E}\left(\log f(y \mid \phi;\theta_2) \mid y;\theta^k\right) && (2.9) \\
&= Q_1(\theta_1;\theta^k) + Q_2(\theta_2;\theta^k), && (2.10)
\end{aligned}
$$

ce qui nous permet de décomposer les étapes E et M de l'algorithme en deux sous-étapes, en particulier l'étape de maximisation. Toujours dans le cas du modèle hiérarchique simple, et grâce aux hypothèses de normalité, cette étape peut se faire de manière explicite, lorsque tous les paramètres sont aléatoires, c'est-à-dire, lorsque la variance de ϕ_{ij} est non nulle, $\forall i, j$. Il est cependant possible de considérer que certains paramètres ne possèdent pas de composante aléatoire mais, dans ce cas, la matrice de covariance Γ n'est pas inversible et cela peut poser des problèmes au moment de l'estimation. Une façon de prendre en compte les paramètres fixes est de les incorporer dans la fonction g représentant la variabilité intra-individuelle (Duval, 2008). La matrice de covariance sera ainsi toujours de plein rang. Dans ce cas, θ_2 comporte en plus de σ^2 les paramètres non aléatoires du modèle. Lorsque l'étape de maximisation n'est pas explicite, on peut avoir recours à des généralisations de l'algorithme EM, comme le ECM (Meng et Rubin, 1993), ou à d'autres méthodes de maximisation de la fonction Q_2, comme par exemple des approches de type quasi-Newton.

2.2.1 Le cas du modèle exponentiel

Un cas particulier de l'algorithme, déjà mis en évidence dans l'article de Dempster et al. (1977), est celui de la famille exponentielle, pour laquelle les deux étapes précédentes s'écrivent simplement en fonction de statistiques exhaustives.

Une famille de densités $\{f(\cdot;\theta)\}$ est dite exponentielle de dimension P (Brown, 1986) si $\forall \theta \in \Theta$ la densité peut s'écrire sous la forme suivante, par rapport à une mesure dominante μ :

$$
f(x;\theta) = h(x)\ \exp(\langle s(\theta), t(x)\rangle - a(\theta)), \tag{2.11}
$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire dans $\mathbb{R}^p$, et $s(\theta) \in \mathbb{R}^p$. La fonction $h : \mathbb{R}^n \mapsto \mathbb{R}$ ne dépend que des données x, et la fonction $a : \mathbb{R}^p \mapsto \mathbb{R}$ ne dépend que du paramètre θ. Il est souvent plus commode de travailler avec le modèle *canonique* ou *naturel* associé à (2.11), défini par :

$$f(x;\eta) = h(x)\ \exp(\langle \eta, t(x) \rangle - b(\eta)), \tag{2.12}$$

où $\eta = s(\theta)$. L'ensemble $\Omega^* = \{\eta \in \mathbb{R}^p : \int h(x) e^{\langle \eta, t(x) \rangle} dx < \infty\}$ est appelé *l'espace paramétrique naturel*, et forme un sous-ensemble convexe de $\mathbb{R}^p$. On suppose alors que η appartient à un sous-ensemble ouvert de Ω^*. Lorsque l'ensemble Ω^* est un ouvert non vide de $\mathbb{R}^p$, le modèle est dit *régulier*. Si ce n'est pas le cas, on peut se restreindre à l'intérieur de Ω^*.

La statistique $t(x)$ est dite *exhaustive*, au sens où elle contient toute l'information nécessaire à l'estimation des paramètres η ou θ. Le modèle exponentiel est dit *courbe* lorsque la dimension de $\eta = s(\theta)$ est inférieure à la dimension de θ, et *non courbe* sinon. Les modèles appartenant à la famille exponentielle possèdent un certain nombre de propriétés utiles pour le calcul du maximum de vraisemblance et de la matrice d'information de Fisher. En particulier, on a :

$$e^{b(\eta)} = \int e^{\langle \eta, t(x) \rangle} h(x)\, dx, \tag{2.13}$$

la fonction $b(\eta)$ est donc le logarithme de la transformée de Laplace de la mesure $\nu \circ t^{-1}$, où $\nu(dx) := h(x)\lambda(dx)$, et où λ est la mesure de Lebesgue sur $\mathbb{R}^p$. Vue comme une fonction des variables complexes $\eta_k = u_k + ib_k$, avec $k = 1, \ldots, p$ et $(u_1, \ldots, u_p) \in \Omega^*$, elle est donc analytique en tout point η tel que $(u_1, \ldots, u_p)$ appartienne à l'intérieur de Ω^*, et l'on peut intervertir intégration et dérivation, ce qui conduit aux égalités suivantes :

$$\mathbb{E}_\eta(t(x)) = \nabla b(\eta) \tag{2.14}$$

$$\mathrm{Cov}_\eta(t(x)) = \nabla^2_\eta(b(\eta)) \tag{2.15}$$

où $\nabla^2_\eta(b(\eta))$ désigne la matrice Hessienne de l'application b évaluée en η, contenant les dérivées partielles secondes de b par rapport aux composantes de η.

Pour une famille de densités $\{f(\cdot;\theta)\}$ on peut définir deux matrices d'information de Fisher : la matrice d'information *observée* notée $I(\theta;x)$, et la matrice d'information *attendue* notée $\mathcal{I}(\theta)$, cette dernière étant définie comme l'espérance de la matrice observée. Efron et Hinkley (1978) ont montré que la matrice d'information observée permettait d'obtenir un meilleur estimateur de la covariance de θ que la matrice d'information attendue, dans le cas d'un vecteur de paramètres unidimensionnel. Elle est également beaucoup plus simple à déterminer car elle n'implique pas le calcul d'une espérance. Lorsqu'elles existent (lorsque la densité f admet des dérivées partielles secondes), ces deux matrices sont définies de la façon suivante :

$$I(\theta;x) = -\nabla^2_\theta \log f(x;\theta) \tag{2.16}$$

$$\mathcal{I}(\theta) = \mathbb{E}_\theta(I(\theta;x)).$$

Dans le cas du modèle exponentiel défini ci-dessus, ces deux quantités sont toujours définies grâce aux propriétés d'analycité de la fonction b, et l'on a :

$$I(\eta; x) = I(\eta) = \nabla^2_\eta(b(\eta)) = \mathrm{Cov}_\eta(t(x)). \tag{2.17}$$

Cette matrice ne dépend pas des données, et les matrices d'information de Fisher observée et attendue sont donc égales :

$$\mathcal{I}(\eta) = I(\eta). \tag{2.18}$$

La matrice d'information de Fisher associée au paramètre $\theta = s^{-1}(\eta)$ s'obtient ensuite à l'aide de la méthode Delta de la façon suivante :

$$\mathcal{I}(\theta) = J_s(\theta)^t \, \nabla^2_\eta(b(s(\theta))) \, J_s(\theta), \tag{2.19}$$

où $J_s(\theta)$ est la matrice jacobienne associée au changement de variable $\eta = s(\theta)$, et $\nabla^2_\eta(b(s(\theta)))$ est la matrice hessienne définie plus haut, pour laquelle on a remplacé η par $s(\theta)$.

Lorsque la densité des données complètes $x = (y, \phi)$ appartient à la famille exponentielle, la fonction Q ne dépend des données qu'en la statistique exhaustive, et les étapes E et M peuvent alors se simplifier. De plus, un certain nombre de résultats de convergence ont été obtenus pour les modèles de la famille exponentielle (voir par exemple Delyon et al. (1999) pour l'algorithme SAEM, ou Fort et Moulines (2003) pour l'algorithme MCMC-EM).

Étape E : on évalue la statistique exhaustive à l'itération k :

$$t^{(k)} = \mathbb{E}(t(x) \mid y; \theta^k) \tag{2.20}$$

Étape M : on actualise la valeur de θ en résolvant l'équation suivante :

$$\mathbb{E}_\theta(t(x)) = t^{(k)}. \tag{2.21}$$

2.2.2 Intervalles de confiance

Notons θ^* la vraie valeur inconnue du paramètre θ, et $\hat{\theta}$ l'estimateur du maximum de vraisemblance. Dans le cas particulier de la famille exponentielle, Sundberg (1974) a montré que sous une condition appelée « $n^{1/2}$-consistance » (qui sera détaillée plus bas), $\hat{\theta}$ est un estimateur consistant de θ^*, et qu'avec une probabilité tendant vers 1 au voisinage de θ^*, cet estimateur est unique. De plus, cet estimateur est asymptotiquement normal et efficace :

$$\sqrt{s}\left(\hat{\theta} - \theta^*\right) \xrightarrow{loi} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}), \tag{2.22}$$

où $\mathcal{I}(\theta^*)$ est la matrice d'information de Fisher attendue (voir paragraphe précédent pour la définition de cette matrice). La condition de $n^{1/2}$-consistance requiert que la matrice $\mathcal{I}(\theta^*)$ soit définie positive, et lorsque la densité des données complètes appartient au modèle exponentiel, Sundberg (1974) a montré que cette hypothèse était suffisante pour assurer la consistance et les propriétés asymptotiques de l'estimateur du maximum de vraisemblance.

La relation 2.22 peut être utilisée pour obtenir des intervalles de confiance asymptotiques pour chacune des composantes de θ, en remplaçant $\mathcal{I}(\theta^*)$ par $\mathcal{I}(\hat{\theta})$. Cependant dans notre cas, la matrice d'information

ne peut pas s'écrire explicitement car elle nécessite le calcul d'une espérance par rapport à la densité des données incomplètes, dont l'expression exacte est inconnue ou difficile à obtenir. On peut alors utiliser le « missing information principle » introduit par Orchard et Woodbury (1972), et appliqué par Louis (1982) à l'estimation de la matrice d'information de Fisher.

Commençons tout d'abord par introduire les notations suivantes pour la fonction score et la matrice d'information de Fisher observée :

$$\begin{aligned} S(y;\theta) &= \nabla_\theta \log f(y;\theta), \\ S_c(x;\theta) &= \nabla_\theta \log f(x;\theta), \\ I(\theta;y) &= -\nabla^2_\theta \log f(y;\theta) = -\nabla_\theta\, S(y;\theta), \\ I_c(\theta;x) &= -\nabla^2_\theta \log f(x;\theta) = -\nabla_\theta\, S(x;\theta), \end{aligned}$$

où $S(y;\theta)$ est le score des données incomplètes et $S_c(x;\theta)$ celui des données complètes, et où $I(\theta;y)$ est la matrice d'information observée des données incomplètes, et $I_c(\theta;x)$ celle des données complètes.

Le « missing information principle » nous permet d'exprimer la matrice d'information de Fisher observée en fonction de deux autres matrices faisant intervenir les données complètes et les données manquantes. En particulier, en utilisant la décomposition $\log f(x;\theta) = \log f(\phi \mid y;\theta) + \log f(y;\theta)$, et en prenant l'espérance conditionnelle par rapport à y, on obtient :

$$I(\theta;y) = I_c(\theta;x) + \nabla^2_\theta \log f(\phi \mid y;\theta), \tag{2.23}$$

soit :

$$I(\theta;y) = \mathcal{I}_c(\theta;y) - \mathcal{I}_m(\theta;y) \tag{2.24}$$

où

$$\mathcal{I}_c(\theta;y) = \mathbb{E}_\theta\left[I_c(\theta;x) \mid y\right] \quad \text{et} \quad \mathcal{I}_m(\theta;y) = \mathbb{E}_\theta\left[-\nabla^2_\theta \log f(\phi \mid y;\theta) \mid y\right],$$

L'information observée correspond donc à l'information complète, moins l'information manquante. Nous renvoyons le lecteur à l'Annexe B pour une démonstration détaillée de cette relation. De la même façon, on a

$$S(y;\theta) = \mathbb{E}_\theta\left[S_c(x;\theta) \mid y\right], \tag{2.25}$$

et la quantité $\mathcal{I}_m(\theta;y)$ peut alors s'écrire (voir Annexe B) :

$$\mathcal{I}_m(\theta;y) = \mathbb{E}_\theta\left[S_c(x;\theta)S_c(x;\theta)^t \mid y\right] - S(y;\theta)S(y;\theta)^t. \tag{2.26}$$

Comme $S(y;\hat{\theta}) = 0$, la matrice d'information $I(\theta;y)$ peut alors être estimée par

$$I(\hat{\theta},y) = \mathbb{E}_{\hat{\theta}}\left[I_c(\hat{\theta},x) \mid y\right] - \mathbb{E}_{\hat{\theta}}\left[S_c(x;\hat{\theta})S_c(x;\hat{\theta})^t \mid y\right]. \tag{2.27}$$

Dans le cas de la famille exponentielle les calculs se simplifient, car la matrice $I_c(\eta;x)$ ne dépend pas des données. En particulier (McLachlan et Krishnan, 2007) :

$$\mathcal{I}_c(\eta;y) = \mathcal{I}_c(\eta) = \mathrm{Cov}_\eta(t(x)), \tag{2.28}$$

$$\mathcal{I}_m(\eta;y) = \mathrm{Cov}_\eta(t(x) \mid y). \tag{2.29}$$

Ces relations étant valables pour le paramètre naturel η, on se ramène à des expressions par rapport à θ en appliquant la méthode Delta (voir 2.19).

Comme pour les équations relatives à l'étape d'espérance, (2.6) et (2.20), le calcul de la matrice d'information de Fisher observée se réduit à un calcul d'espérance conditionnelle sous la loi des données cachées ϕ sachant les observations y. Dans notre cas, cette étape n'étant pas explicite, nous utiliserons deux extensions de l'algorithme EM permettant de contourner cette difficulté en remplaçant l'étape E par une étape de simulation. Ces deux approches seront discutées dans les sections 2.3 et 2.4.

2.2.3 Convergence de l'algorithme

Dans le cas particulier de la famille exponentielle, la convergence de l'algorithme EM a été étudiée par plusieurs auteurs, notamment par Sundberg (1974) et Dempster et al. (1977). Quelques années plus tard, Wu (1983), apporte la preuve définitive en corrigeant une erreur dans la démonstration de Dempster et al. (1977). Elle a également été étudiée par les différents auteurs qui ont proposé des extensions de l'algorithme, notamment Fort et Moulines (2003) ou Delyon et al. (1999). Nous présentons ici les hypothèses données par Fort et Moulines (2003).

Nous notons λ la mesure de Lebesgue sur $\mathbb{R}^n$, et ℓ_c la log-vraisemblance des données complètes, c'est-à-dire telle que $\ell_c(t;\theta) := \langle s(\theta), t\rangle - a(\theta)$. Les hypothèses requises pour la convergence de l'algorithme EM sont les suivantes :

(M1) L'espace des paramètres Θ est un ouvert de $\mathbb{R}^d$, où d est la dimension de θ.

(M2) (a) Les fonctions $a : \Theta \to \mathbb{R}$ et $s : \Theta \to \mathbb{R}^d$, définies dans (2.11), sont continues sur Θ, et $t : \mathbb{R}^{q+l} \to \mathcal{T} \subseteq \mathbb{R}^d$ est continue sur $\mathbb{R}^l$,

(b) pour tout θ appartenant à Θ, $\bar{t}(\theta) := \int t(y,\phi) f(\phi|y;\theta)\lambda(d\phi)$ est finie et continue sur Θ,

(c) il existe une fonction continue $\hat{\theta} : \mathcal{T} \to \Theta$ telle que pour tout t appartenant à $\mathcal{T}$, $\ell_c(t;\hat{\theta}(t)) = \sup_{\theta\in\Theta} \ell_c(t,\theta)$,

(d) la vraisemblance des observations L est positive, finie et continue sur Θ,

(e) quel que soit $M > 0$, l'ensemble $\{\theta \in \Theta, L(\theta) \geq M\}$ est compact.

(M3) Si l'on note l'ensemble des points stationnaires de l'algorithme $\mathcal{L} := \{\theta \in \Theta, \hat{\theta} \circ \bar{t}(\theta) = \theta\}$, alors soit $L(\mathcal{L})$ est compact, soit pour tout compact $\mathcal{K} \subseteq \Theta$, $L(\mathcal{L} \cap \mathcal{K})$ est fini.

Remarques

i) Les hypothèses (M2)(a) et (b) correspondent à l'hypothèse de Wu (1983) selon laquelle la fonction Q doit être continue en θ et en θ^k. En effet, dans le cas du modèle exponentiel régulier, la fonction Q est toujours continue en θ dès que s et a sont continues en θ, ce qui est contrôlé par l'hypothèse (M2)(a), et elle est continue en θ^k lorsque l'hypothèse (M2)(b) est vérifiée.

ii) L'hypothèse (M2)(e) peut paraître restrictive dans la pratique. Cependant, elle peut être remplacée par l'hypothèse plus faible suivante :

(e*) il existe un M_0 strictement positif tel que pour tout $M > M_0$, l'ensemble $\{\theta \in \Theta, L(\theta) \geq M\}$ est compact.

Lorsque $M_0 = 0$ les hypothèses (M2)(e*) et (M2)(e) sont identiques, (M2)(e*) est donc plus faible que (M2)(e). De plus, la monotonie de l'algorithme EM nous garantit que s'il existe un entier $n \in \mathbb{N}$ tel que $L(\theta^{(n)}) > M_0$, alors sous l'hypothèse (M2)(e*) la séquence d'itérations produites par l'algorithme EM finit par se concentrer effectivement dans un sous-ensemble compact de Θ.

Nous montrerons en section 2.2.3 comment ces hypothèses s'appliquent dans la pratique, en prenant l'exemple du modèle Greenlab.

2.3 L'algorithme MCMC-EM

Comme précisé en (2.6) et (2.20), chaque étape E de l'algorithme requiert le calcul d'une espérance conditionnelle selon la loi des données cachées ϕ sachant les observations y. Bien souvent, comme c'est le cas ici, ce calcul ne peut pas se faire explicitement, et il faut alors approcher la quantité d'intérêt, que ce soit la fonction Q dans un cas général, ou la statistique exhaustive t dans le cas exponentiel.

Une première approche consiste à approcher l'intégrale par une approximation de type Monte-Carlo (Wei et Tanner, 1990). Le principe des méthodes de Monte-Carlo est simple, et repose sur la loi forte des grands nombres. Si l'on dispose de n variables aléatoires i.i.d., $X_1, \ldots, X_n$ de loi μ, alors on a :

$$\frac{1}{m}\sum_{i=1}^{m} h(X_i) \overset{m\to\infty}{\longrightarrow} \int h(x)d\mu(x) \quad p.s., \tag{2.30}$$

pour toute fonction $h \in L^1(\mu)$.

Il est alors possible d'approcher cette intégrale en simulant un grand nombre de réalisations indépendantes X_i selon la loi μ. La précision peut être choisie aussi fine que l'on souhaite, en augmentant le nombre de simulations, et la vitesse de convergence peut être obtenue à l'aide du théorème central limite, lorsque la fonction h est dans $L^2(\mu)$.

Cependant, cela nécessite de pouvoir simuler facilement selon la loi d'intérêt μ, ce qui n'est pas toujours le cas dans les problèmes de données incomplètes. Une autre méthode proposée par McCulloch (1994, 1997) consiste alors, non pas à utiliser des échantillons i.i.d. de loi μ, mais les réalisations d'une chaîne de Markov de loi stationnaire μ grâce au théorème ergodique appliqué à la fonctionnelle h.

À l'itération $k+1$ de l'algorithme, on remplace alors l'étape E par une étape de simulation :

Étape S : on génère une chaîne de Markov de taille m_k et de loi stationnaire $f(\phi \mid y; \theta^k)$. En notant $\left(\phi^{k,(1)}, \ldots, \phi^{k,(m_k)}\right)$ les m_k réalisations de la chaîne de Markov, la fonction $Q(\theta; \theta^k)$ peut alors être approchée par

$$\hat{Q}(\theta; \theta^k) = \frac{1}{m_k}\sum_{m=1}^{m_k} \log f(y, \phi^{k,(m)}; \theta). \tag{2.31}$$

Dans le cas exponentiel, il suffit de mettre à jour les statistiques suffisantes du modèle :

$$\hat{t}^{(k)} = \frac{1}{m_k}\sum_{m=1}^{m_k} t\left(y, \phi^{k,(m)}\right). \tag{2.32}$$

Dans le cas du modèle exponentiel, la convergence de l'algorithme MCMC-EM a été étudiée notamment par Fort et Moulines (2003). Nous donnons en section 2.3.6 les hypothèses correspondantes.

Dans la pratique, il est plus efficace de définir une période de burn-in pendant laquelle on laisse la chaîne de Markov parcourir l'espace d'états, et de conserver uniquement les simulations suivantes pour calculer les quantités qui nous intéressent (ici, 2.31 et 2.32). À première vue, le problème ne semble pas plus simple, car il faut maintenant savoir simuler une chaîne de Markov de loi stationnaire μ. Heureusement, plusieurs algorithmes ont été proposés, et permettent à partir de noyaux de transition judicieusement

choisis d'obtenir des chaînes de Markov ayant pour loi stationnaire la loi cible μ. Nous présentons brièvement les pré-requis nécessaires à l'application du théorème ergodique, puis deux exemples d'algorithmes de type MCMC. Les notions présentées sont issues des ouvrages de Robert et Casella (1999) et de Roberts et Tweedie (2008).

2.3.1 Conditions d'application du théorème ergodique

Deux hypothèses sont nécessaires à l'application de la loi forte des grands nombres : d'une part les variables $X_1, \ldots, X_n$ doivent être indépendantes, et d'autre part elles doivent avoir la même loi. Dans le cadre des chaînes de Markov, il est possible d'obtenir des réalisations issues de la même loi lorsque la chaîne a convergé vers sa distribution stationnaire, mais les variables restent corrélées, ce qui nous empêche d'utiliser la loi forte des grands nombres. Cependant, sous réserve que la chaîne possède certaines propriétés supplémentaires, le théorème ergodique nous permet d'obtenir un résultat similaire à celui de la loi des grands nombres. Nous énonçons tout d'abord ce théorème, puis nous détaillons les hypothèses nécessaires à son application.

Théorème 2.1. *Une chaîne de Markov $\{X_i\}$ à valeurs dans $\mathcal{X}$ de loi stationnaire π est dite* ergodique *lorsqu'elle est apériodique, irréductible et récurrente positive. Elle vérifie alors la propriété suivante, pour toute fonction $h \in L^1(\pi)$:*

$$\frac{1}{M}\sum_{i=1}^{M} h(X_i) \overset{M\to\infty}{\longrightarrow} \int h(x)d\pi(x) \quad p.s.. \tag{2.33}$$

Dans le cas où l'espace d'états $\mathcal{X}$ est continu, on note $K(x,\cdot)$ la loi conditionnelle de l'état X_{i+1} sachant $\{X_i = x\}$. Pour appliquer le théorème, la chaîne doit alors vérifier les propriétés suivantes :

- *irréductibilité* : une chaîne est dite *φ-irréductible* s'il existe une mesure non triviale φ telle que pour tout x dans $\mathcal{X}$, il existe un entier $n \in \mathbb{N}^*$ tel que $K^n(x, A) > 0$ dès que $\varphi(A) > 0$. Cette propriété permet notamment de s'affranchir des conditions initiales, car tout sous-ensemble de l'espace d'états est atteignable avec une probabilité non nulle, quel que soit le point de départ de la chaîne. Si la chaîne est φ-irréductible, alors elle admet une mesure irréductible maximale ψ, c'est-à-dire telle que toute mesure irréductible φ soit absolument continue par rapport à cette mesure maximale. La chaîne est alors dite ψ-irréductible.
- *récurrence* : une chaîne ψ-irréductible est dite *récurrente* si pour tout ensemble A de mesure non nulle pour ψ, en partant d'un élément dans A, le nombre moyen de retours dans A est infini. Elle est dite *récurrente positive* lorsqu'elle admet comme mesure invariante une mesure de probabilité.
- *apériodicité* : si une chaîne de loi stationnaire π est ψ-irréductible, alors l'espace d'états admet la décomposition cyclique suivante $\mathcal{X} = \left(\cup_{i=1}^{d}\mathcal{X}_i\right) \cup N$, avec $\psi(N) = 0$. La plus grande valeur de d pour laquelle un tel cycle existe est appelée la *période* de la chaîne. Si $d = 1$, la chaîne est dite *apériodique*.

Une fois que l'on a trouvé une chaîne de Markov dont le noyau de transition K admet comme mesure de probabilité invariante la loi d'intérêt μ, il suffit de vérifier l'apériodicité et l'irréductibilité de la chaîne pour pouvoir appliquer (2.33). Les deux sections suivantes proposent deux algorithmes permettant de construire de telles chaînes : l'algorithme de Metropolis-Hastings (section 2.3.2) et l'échantillonneur de Gibbs (section 2.3.4).

Dans le cadre qui nous intéresse, nous utiliserons ces algorithmes pour obtenir à l'itération k de l'algorithme EM, et pour chaque individu i, une chaîne $(\phi_i^{(0)}, \ldots, \phi_i^{(m_k)})$ de loi stationnaire ayant comme densité $f(\phi_i \mid y_i; \theta^k)$.

2.3.2 Algorithme de Metropolis-Hastings

L'algorithme de Metropolis-Hastings (MH) a d'abord été introduit par Metropolis et al. (1953), pour le calcul d'intégrales faisant intervenir des distributions de Boltzmann, et a été généralisé près de 20 ans plus tard par Hastings (1970). Partant d'une loi cible de densité f, on commence par choisir une densité conditionnelle $q(\cdot|u)$, dont on sait facilement simuler des réalisations aléatoires, et qui est soit connue à une constante multiplicative près, soit symétrique (c'est-à-dire telle que $q(v|u) = q(u|v)$). On suppose également que le rapport $f(v)/q(v|u)$ est connue à une constante multiplicative indépendante de u près [1].

Dans notre cas, chaque itération de l'algorithme MH consiste alors à simuler, à partir de l'état courant de la chaine u, un candidat v qui sera accepté comme le nouvel état de la chaîne avec une probabilité $\alpha(x, y)$. Plus précisément :

On initialise la chaîne avec $U^{(0)}$, puis pour $m = 1, \ldots, M$:

1. on génère un candidat $v \sim q(\cdot \mid U^{(m-1)})$
2. on pose

$$U^{(m)} = \begin{cases} v & \text{avec une probabilité } \alpha(U^{(m-1)}, v) \\ U^{(m-1)} & \text{avec une probabilité } 1 - \alpha(U^{(m-1)}, v) \end{cases} \tag{2.34}$$

où

$$\alpha(u, v) = \min\left(1, \frac{f(v)}{f(u)} \frac{q(u \mid v)}{q(v \mid u)}\right). \tag{2.35}$$

On utilise la loi instrumentale $q(\cdot|u)$ pour générer des réalisations de la loi f donc, plus le rapport f/q est faible pour le candidat par rapport à l'état courant de la chaîne, plus la probabilité de le rejeter est forte. Intuitivement, la probabilité d'acceptation $\alpha(u, v)$ permet de faire un compromis entre les deux conditions suivantes : d'une part, on souhaite que l'algorithme se dirige vers des régions de plus forte probabilité sous f, ce qui est contrôlé par le rapport $f(v)/f(u)$ (plus celui-ci est haut, plus on accepte le candidat), et d'autre part on souhaite éviter que l'algorithme ne reste trop longtemps dans une région spécifique de trop forte probabilité sous q, ce qui est contrôlé par le rapport $q(u \mid v)/q(v \mid u)$. L'un des pré-requis nécessaires pour assurer la convergence de l'algorithme est que le support de $q(\cdot|u)$ doit contenir le support de f, $\forall u$.

Le noyau de transition de la chaîne de Markov générée par l'algorithme MH est donné par $K(u, dv) = \alpha(u, v)q(v|u)dv + (1 - \int \alpha(u, v)q(v|u)dv)\delta_u(dv)$, où δ_u désigne la masse de Dirac au point u. En particulier, il vérifie la condition d'équilibre (« detailed balanced condition »), impliquant que la chaîne admet f comme mesure invariante. Lorsque $q(v|u) > 0 \ \forall u, v$ la chaîne est irréductible, et elle est également périodique, car $\mathbb{P}(\{U^{(m+1)} = U^{(m)}\}) > 0$.

Dans la pratique, le choix de la distribution instrumentale q influence fortement la vitesse de convergence de la chaîne de Markov vers la loi stationnaire f car il dirige l'exploration de l'espace d'états par la chaîne de Markov (Robert et Casella, 1999). Deux types d'approches sont utilisées classiquement pour le choix de la loi q :

1. Ce type d'approche est particulièrement utile dans un cadre Bayésien, pour le calcul des lois *a posteriori* que l'on connaît en général à une constante multiplicative près.

Le cas indépendant

Lorsque la loi instrumentale q est indépendante de la position actuelle de la chaîne $X^{(t)}$, l'algorithme peut être vu comme une extension des méthodes d'acceptation-rejet. Dans ce cas, la probabilité d'acceptation s'écrit :

$$\alpha(u, v) = \min\left(1, \frac{f(v)}{f(u)}\frac{q(u)}{q(v)}\right). \tag{2.36}$$

Dans notre cas, rappelons que l'on souhaite simuler selon la loi cible $f(\phi_i|y_i;\theta^k)$. On peut choisir comme loi instrumentale la loi marginale $f(\phi_i;\theta^k)$, c'est-à-dire la loi normale $\mathcal{N}(\beta^{(k)}, \Gamma^{(k)})$, où $\beta^{(k)}$ et $\Gamma^{(k)}$ correspondent aux estimations courantes des paramètres β et Γ. En notant $\tilde{\phi}_i$ le candidat et $\phi_i^{k,(m)}$ l'état actuel de la chaîne, le rapport $\frac{f(v)}{f(u)}\frac{q(u)}{q(v)}$ se simplifie alors de la façon suivante :

$$\begin{aligned}\frac{f(\tilde{\phi}_i \mid y_i;\theta^k)}{f(\phi_i^{k,(m)} \mid y_i;\theta^k)}\frac{f(\phi_i^{k,(m)};\theta^k)}{f(\tilde{\phi}_i;\theta^k)} &= \frac{f(\tilde{\phi}_i, y_i;\theta^k)f(y_i;\theta^k)}{f(\phi_i^{k,(m)}, y_i;\theta^k)f(y_i;\theta^k)}\frac{f(\phi_i^{k,(m)};\theta^k)}{f(\tilde{\phi}_i;\theta^k)} \\ &= \frac{f(y_i \mid \tilde{\phi}_i;\theta^k)}{f(y_i \mid \phi_i^{k,(m)};\theta^k)}.\end{aligned}$$

La marche aléatoire

Un autre choix usuel pour la loi instrumentale est celui d'une marche aléatoire autour de la valeur actuelle de la chaîne :

$$v = U^{(m-1)} + \varepsilon_m,$$

où ε_m suit une loi de densité q indépendante de $U^{(m-1)}$.

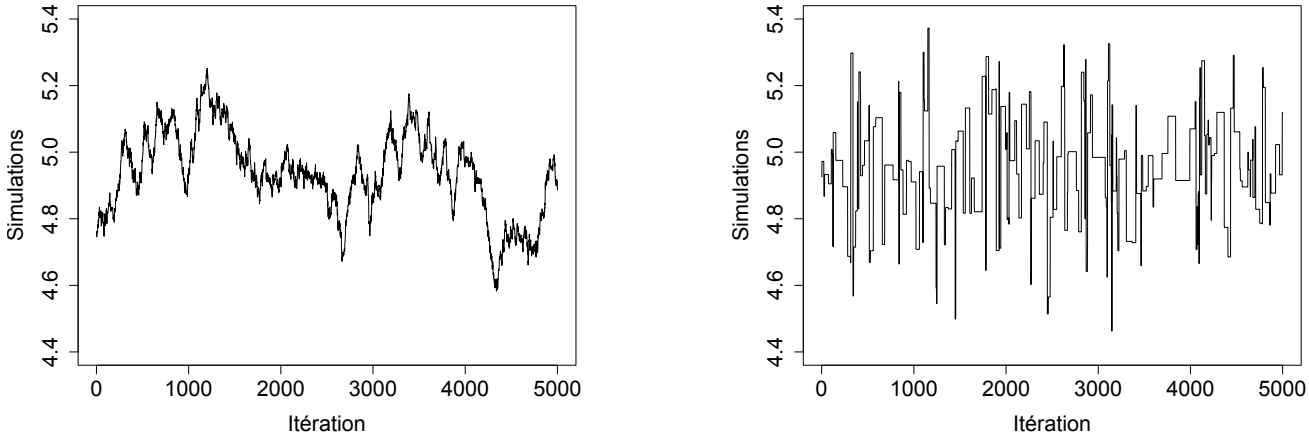
Dans ce cas, la loi $q(v|u)$ peut s'écrire $q(v-u)$. Lorsque, de plus, la loi considérée est symétrique, c'est-à-dire lorsque $q(u) = q(-u)$, la probabilité d'acceptation s'écrit :

$$\alpha(u, v) = \min\left(1, \frac{f(v)}{f(u)}\right). \tag{2.37}$$

Dans notre cas, le rapport $\frac{f(v)}{f(u)}$ se simplifie pour ne faire intervenir que des quantités connues :

$$\begin{aligned}\frac{f(\tilde{\phi}_i \mid y_i;\theta^k)}{f(\phi_i^{k,(m)} \mid y_i;\theta^k)} &= \frac{f(\tilde{\phi}_i, y_i;\theta^k)f(y_i;\theta^k)}{f(\phi_i^{k,(m)}, y_i;\theta^k)f(y_i;\theta^k)} \\ &= \frac{f(y_i \mid \tilde{\phi}_i;\theta^k)f(\tilde{\phi}_i;\theta^k)}{f(y_i \mid \phi_i^{k,(m)};\theta^k)f(\phi_i^{k,(m)};\theta^k)}.\end{aligned}$$

Parmi les choix classiques pour la loi q, on relève la loi uniforme, la loi normale ou la loi de Student, centrées autour de 0. Le choix du support de la loi uniforme, ou de la variance des lois normale et de Student va conditionner la vitesse de convergence de l'algorithme. Si la densité est trop concentrée autour de 0, le candidat généré à chaque étape sera souvent accepté, et l'exploration de l'espace d'états sera lente. Si, au contraire, la densité est trop étendue, les candidats proposés seront souvent rejetés car correspondant à des régions de trop faible probabilité sous f, et la chaîne aura donc tendance à rester trop longtemps au même point (voir figure 2.1).



(a) σ = 0.01 (b) σ = 5

FIG. 2.1 – Algorithme de Metropolis-Hastings à marche aléatoire gaussienne de moyenne 0 et de variance σ^2. Sur la figure de gauche la variance est trop faible, les candidats sont trop souvent acceptés (taux d'acceptation : 0.80) et la chaîne se déplace lentement. Sur la figure de droite, au contraire, la variance étant trop grande les candidats sont trop souvent rejetés (taux d'acceptation : 0.04) et la chaîne reste longtemps au même endroit.

Pour assurer une bonne exploration de l'espace d'états par la chaîne, on peut s'intéresser au taux d'acceptation des candidats, qui doit donc être ni trop élevé ni trop bas. Des valeurs optimales ont été dérivées pour ce taux d'acceptation, selon l'algorithme utilisé. Dans le cas de l'algorithme de Metropolis-Hastings multidimensionnel à marche aléatoire, ce taux optimal a été estimé à 0.234 lorsque la dimension de $\mathcal{X}$ tend vers l'infini (voir par exemple Roberts et al. (1997) ; Roberts et Rosenthal (2001) ; Bédard (2007, 2008)). Dans le cas de l'algorithme de MH à marche aléatoire Gaussienne, Gelman et al. (1996) ont montré que ce taux optimal correspondait à une matrice de covariance optimale égale à $(2.38^2/n_x)\Sigma_\pi$, où Σ_π est la vraie matrice de covariance de la loi cible, et n_x la dimension de $\mathcal{X}$. Plusieurs algorithmes adaptatifs ont alors été développés dans la littérature pour estimer Σ_π, notamment par Haario et al. (1999, 2001) (algorithme AM - « Adaptive Metropolis-Hastings »), qui proposent d'estimer Σ_π à la fin de chaque itération m de la chaîne à l'aide des réalisations précédentes $(X_0, \ldots, X_m)$:

$$\mu_m = \frac{1}{m}\sum_{i=1}^{m} X_i, \qquad \Sigma_m = \frac{1}{m}\sum_{i=1}^{m}(X_i - \mu_m)(X_i - \mu_m)'. \tag{2.38}$$

Andrieu et Thoms (2008) ont quant à eux proposé un cadre général basé sur l'approximation stochastique de Robbin-Monro Robbins et Monro (1951) [2] pour les algorithmes à adaptation *décroissante*, dont fait partie l'algorithme AM. L'atténuation du caractère adaptatif permet notamment de conserver la convergence de la chaîne vers la loi cible malgré la perte du caractère markovien, puisque l'état courant de la chaîne dépend maintenant de tout le passé. Ils proposent notamment la généralisation suivante de l'algorithme de Haario et al. (2001) : à chaque itération $m+1$ de la chaîne, le candidat est généré selon la loi $\mathcal{N}(X_m, \lambda\Sigma_m)$, où λ est une constante strictement positive, et où Σ_m est donnée par :

$$\mu_m = \mu_{m-1} + \frac{1}{\gamma_m}(X_m - \mu_{m-1}), \tag{2.39}$$

2. c'est le même principe d'approximation stochastique qui est utilisé dans l'algorithme SAEM présenté en section 2.4

$$\Sigma_m = \Sigma_{m-1} + \frac{1}{\gamma_m}((X_m - \mu_m)(X_m - \mu_m)' - \Sigma_{m-1}),$$

où la suite (γ_m) vérifie $\sum_{m=1}^{\infty} \gamma_m = \infty$ et $\sum_{m=1}^{\infty} \gamma_m^{1+\nu} < \infty$ pour un certain $\nu > 0$.

Cependant, les mêmes difficultés que celles évoquées plus haut peuvent émerger, en particulier si l'exploration de l'espace d'états est trop lente, car les estimations proposées ci-dessus peuvent alors s'avérer médiocres. Andrieu et Thoms (2008) proposent alors différents schémas adaptatifs où le paramètre λ est également adapté à chaque itération de la chaîne, pour permettre de s'approcher du taux d'acceptation optimal. Ils proposent notamment la mise à jour suivante du paramètre λ_m à l'itération m :

$$\log \lambda_m = \log \lambda_{m-1} + \frac{1}{\gamma_m}(\alpha(X_{m-1}, Y_m) - \alpha_*), \tag{2.40}$$

où Y_m est le candidat généré à l'itération m, $\alpha(X_{m-1}, Y_m)$ est la probabilité d'accepter le déplacement $X_m \hookrightarrow Y_m$ et α_* est le taux d'acceptation que l'on souhaite atteindre, c'est-à-dire dans le cas multidimensionnel, 0.234. En effet, lorsque la quantité $\alpha(X_{m-1}, Y_m) - \alpha_*$ est négative, cela signifie que le taux d'acceptation est trop bas, et que la variance doit être diminuée : dans ce cas la valeur de λ_m diminue. De la même façon, lorsque cette même quantité est négative, cela signifie que le taux d'acceptation est trop élevé, et on augmente alors la variance de la marche aléatoire en augmentant λ_m.

Les auteurs proposent également d'autres algorithmes adaptatifs, et notamment une version où l'adaptation se fait composante par composante. Nous détaillerons cet algorithme dans la section 2.3.4.

2.3.3 Échantillonneur de Gibbs

L'échantillonneur de Gibbs a été introduit par Geman et Geman (1984) puis repris plus tard par Gelfand et al. (1990), et repose sur l'idée selon laquelle on peut simuler selon une loi multidimensionnelle en simulant chaque composante du vecteur selon la loi conditionnelle associée. Plus précisément, si on a $U = (U_1, \ldots, U_p)$ de loi f, et si l'on peut facilement simuler selon les lois conditionnelles $f_1, \ldots, f_p$:

$$U_j \mid u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_p \sim f_j(\cdot \mid u_1, \ldots, u_{j-1}, u_{j+1}, \ldots, u_p), \tag{2.41}$$

il est alors possible de réduire la dimension du problème, car il suffit de savoir simuler selon p lois univariées. L'échantillonneur de Gibbs consiste alors en p simulations séquentielles de chaque composante du vecteur U :

On initialise la chaîne avec $U^{(0)} = (U_1^{(0)}, \ldots, U_p^{(0)})$, puis pour $m = 1, \ldots, M$ on génère :

1. $U_1^{(m)} \sim f_1(\cdot \mid u_2^{(m-1)}, \ldots, u_p^{(m-1)})$,
2. $U_2^{(m)} \sim f_2(\cdot \mid u_1^{(m)}, u_3^{(m-1)}, \ldots, u_p^{(m-1)})$,
3. ...
4. $U_p^{(m)} \sim f_p(\cdot \mid u_1^{(m)}, \ldots, u_{p-1}^{(m)})$.

Les fonctions $f_j, j = 1, \ldots, p$ sont appelées les *lois conditionnelles complètes*. L'un des avantages de cette approche par rapport à celle de l'algorithme MH, est que l'on utilise directement l'information contenue dans la fonction f, sans avoir recours à une loi instrumentale. De plus, il est possible de décomposer un problème multidimensionnel en plusieurs sous-problèmes de dimensions plus faibles. Même si l'échantillonneur de Gibbs est un cas particulier de l'algorithme MH (Robert et Casella, 1999), les conditions d'application du théorème ergodique sont plus difficiles à vérifier. Cependant, dans le cas particulier où le

noyau de transition associé à la chaîne produite par l'algorithme est absolument continu par rapport à la mesure dominante (en général, la mesure de Lebesgue), la chaîne est irréductible, et le théorème ergodique s'applique. Nous renvoyons le lecteur à Robert et Casella (1999) pour une preuve détaillée.

2.3.4 Échantillonneur de Gibbs hybride

Il peut être difficile de simuler directement selon les lois conditionnelles complètes $f_j, j = 1, \ldots, p$. Dans ce cas, on peut utiliser l'échantillonneur de Gibbs hybride, aussi appelé « Metropolis-within-Gibbs », où certaines étapes de simulation selon la loi conditionnelle complète f_j sont remplacées par une étape de type Metropolis-Hastings.

On initialise la chaîne avec $U^{(0)} = (U_1^{(0)}, \ldots, U_p^{(0)})$, puis pour $m = 1, \ldots, M$, et $j = 1, \ldots, p$:

1. on génère un candidat $v_j \sim q_j(\cdot \mid u_1^{(m)}, \ldots, u_{j-1}^{(m)}, u_j^{(m-1)}, u_{j+1}^{(m-1)}, \ldots, u_p^{(m-1)})$
2. on pose

$$U_j^{(m)} = \begin{cases} v_j & \text{avec une probabilité } \alpha_j(U_j^{(m-1)}, v_j) \\ U_j^{(m-1)} & \text{avec une probabilité } 1 - \alpha_j(U_j^{(m-1)}, v_j) \end{cases} \tag{2.42}$$

où

$$\alpha_j(u, v) = \min\left(1, \frac{f_j(v)}{f_j(u)} \frac{q_j(u \mid v)}{q_j(v \mid u)}\right).$$

Dans notre cas, les lois conditionnelles complètes sont les $f(\phi_{ij} \mid y_i; \theta^k)$, et le choix de la distribution q_i se fait comme dans le cas de l'algorithme MH, avec les deux mêmes approches principales. De plus, lorsque chaque étape de simulation se fait à partir d'un algorithme de Metropolis-Hastings (par exemple, lorsqu'il est difficile ou impossible de simuler selon les lois conditionnelles complètes), il est possible de calculer la probabilité d'acceptation une seule fois, c'est-à-dire après avoir simulé chacune des p composantes, et en utilisant la probabilité d'acceptation définie en (2.35) en utilisant le vecteur formé de tous les candidats générés par l'algorithme. L'algorithme peut s'avérer moins efficace dans ce cas car on peut être amené à rejeter l'ensemble des candidats, mais il possède l'avantage de pouvoir s'écrire simplement comme un algorithme de Metropolis-Hastings.

Dans ce contexte, Andrieu et Thoms (2008) proposent deux versions adaptatives correspondant au cas où l'échantillonneur de Gibbs hybride peut s'écrire comme un algorithme MH, où l'adaptation se fait composante par composante. Dans ce cas unidimensionnel, le taux d'acceptation optimal pour une marche aléatoire Gaussienne est de $\alpha_{**} = 0.44$ (Roberts et al., 1997 ; Roberts et Rosenthal, 2001). Ces deux algorithmes sont décrits ci-dessous.

Marche aléatoire et adaptation composante par composante

On initialise la chaîne $U^{(0)} = (U_1^{(0)}, \ldots, U_p^{(0)})$, ainsi que les quantités μ_0, Σ_0 et le vecteur $\lambda_0 := (\lambda_0^1, \ldots, \lambda_0^p)$ puis pour $m = 1, \ldots, M$:

1. on choisit une composante k parmi $1, \ldots, p$

2. on génère $v_k \sim \mathcal{N}(U_k^{(m-1)}, \lambda_0^k\ (\Sigma_{m-1})_{k,k}))$ et on pose $V = (V_1, \ldots, V_p)$, avec $V_k = v_k$ et $V_j = U_j^{(m-1)}$ si $j \neq k$
3. on pose $U^{(m)} = V$ avec probabilité $\alpha(U^{(m-1)}, V)$ et $U^{(m)} = U^{(m-1)}$ avec probabilité $1 - \alpha(U^{(m-1)}, V)$, où $\alpha(u, v)$ est donné en (2.37)
4. on actualise μ_m et Σ_m comme en (2.39), puis le vecteur λ_m en posant $\lambda_m^j = \lambda_{m-1}^j$ si $j \neq k$ et

$$\lambda_m^k = \lambda_{m-1}^k + \gamma_m \left(\alpha(U^{(m-1)}, V) - \alpha_{**}\right). \tag{2.43}$$

Marche aléatoire globale et adaptation composante par composante

On initialise la chaîne $U^{(0)} = (U_1^{(0)}, \ldots, U_p^{(0)})$, μ_0, Σ_0 et $\lambda_0 := (\lambda_0^1, \ldots, \lambda_0^p)$ puis pour $m = 1, \ldots, M$:

1. on génère $V \sim \mathcal{N}_p(0, \Lambda_m^{1/2}\ \Sigma_{m-1}\ \Lambda_m^{1/2})$ où Λ_m : diag$(\lambda_m^1, \ldots, \lambda_m^p)$
2. on pose $U^{(m)} = U^{(m-1)} + V$ avec probabilité $\alpha(U^{(m-1)}, V)$ et $U^{(m)} = U^{(m-1)}$ avec probabilité $1 - \alpha(U^{(m-1)}, V)$ où $\alpha(u, v)$ est donné en (2.37)
3. on actualise μ_m et Σ_m comme précédemment, et pour $j = 1, \ldots, p$ on pose :

$$\log \lambda_m^j = \log \lambda_{m-1}^j + \gamma_m \left(\alpha(U^{(m-1)}, U_k^{(m-1)} + V_k e_k) - \alpha_{**}\right) \tag{2.44}$$

où V_k est la k-ème composante du vecteur V et e_k est le vecteur dont tous les éléments sont nuls sauf le k-ème, qui vaut 1.

Dans ce dernier algorithme, en plus du calcul de la probabilité d'acceptation $\alpha(U^{(m-1)}, V)$, l'actualisation des valeurs de λ_m nécessite le calcul de p probabilités d'acceptation « unidirectionnelles », correspondant à un déplacement unidirectionnel de la chaîne dans chacune des p directions possibles, ce qui peut entraîner un surcoût de temps de calcul.

2.3.5 Taille de la chaîne et critère d'arrêt

Lorsque la fonction Q est approchée par des méthodes de type Monte-Carlo ou Monte-Carlo par chaîne de Markov, il faut tenir compte de l'erreur commise en remplaçant Q par $\hat{Q}$, communément appelée erreur de Monte-Carlo. En particulier, la propriété de monotonie de l'algorithme EM, qui garantissait une augmentation de la vraisemblance à chaque itération, n'est plus nécessairement vérifiée. De plus, l'utilisation d'une taille m_k constante ne permet pas la convergence de l'algorithme, à cause de la persistance de l'erreur de Monte Carlo (voir par exemple Booth et al. (2001)). Ce phénomène est illustré sur la figure 2.2, où nous avons représenté l'estimation d'un paramètre en fonction de l'itération de l'algorithme MCMC-EM, pour une taille constante de la chaine à chaque itération ($m_k = 250$). On remarque que la taille de la chaîne n'est pas suffisante pour obtenir une bonne précision, et que les estimations oscillent avec une amplitude constante autour du maximum de vraisemblance. Une première approche consisterait alors à augmenter la taille de la chaîne afin d'obtenir une meilleure précision. D'un autre côté, il est inutile d'utiliser une chaîne de taille trop importante lorsque l'algorithme est encore loin de la convergence.

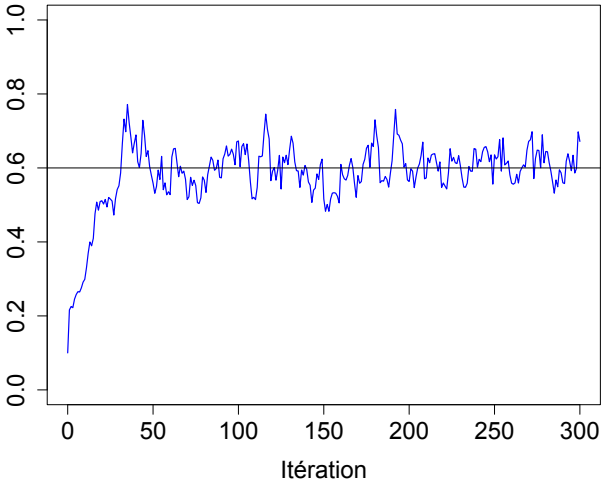


FIG. 2.2 – Exemple d'un algorithme MCMC-EM où la taille de la chaîne est constante à chaque itération de l'algorithme : l'erreur de Monte Carlo empêche la convergence de l'algorithme.

L'idée est donc de démarrer avec une chaîne de taille suffisante sans être trop grande, puis d'augmenter la taille de la chaîne à chaque itération pour obtenir des estimations de plus en plus précises à mesure que l'on s'approche du maximum de vraisemblance. Certains auteurs proposent par exemple une augmentation déterministe de la taille de l'échantillon (McCulloch, 1994, 1997), l'algorithme étant arrêté après un nombre fixé d'itérations. Fort et Moulines (2003) proposent de coupler une augmentation déterministe de la taille de la chaîne à une procédure de moyennisation des estimations basée sur la méthode de Polyak et Juditsky (1992), et ont montré que cette approche permettait d'augmenter la vitesse de convergence de l'algorithme. Parmi les avantages de ce type d'approches, on peut citer notamment la simplicité d'implémentation, qui ne requiert pas de calculs supplémentaires, ce qui peut s'avérer être un avantage certain lorsque le modèle considéré est complexe. De plus, il est possible de contrôler le nombre final d'itérations, ce qui dans certains cas peut s'avérer nécessaire Cappé et al. (2005). Cependant, le choix du nombre d'itérations à partir duquel les estimations sont moyennées, ou de la façon dont on augmente la taille de la chaîne (augmentation géométrique, polynomiale, ...) sont laissés libres à l'utilisateur, et plusieurs réalisations de l'algorithme peuvent être nécessaires avant d'identifier la meilleure paramétrisation.

Parallèlement à ces approches déterministes, d'autres auteurs ont proposé des critères automatiques basés sur l'estimation de l'erreur de Monte-Carlo pour augmenter la taille de la chaîne et arrêter l'algorithme. Booth et Hobert (1999) sont les premiers à proposer un tel algorithme automatique dans le cas où l'échantillon est simulé directement selon la loi d'intérêt, ou par échantillonnage d'importance, c'est-à-dire lorsque les réalisations sont i.i.d. À partir d'un ellipsoïde de confiance construit autour de θ^k, la taille de l'échantillon est augmentée si l'ellipsoïde contient la valeur précédente θ^{k-1}. Les auteurs proposent dans ce cas une augmentation géométrique du type $m_k \leftarrow m_k + m_k/c$, où $c = 2, 3$ ou 4. Cette approche permet notamment d'éviter de simuler des échantillons de taille trop importante lorsque l'on est encore loin du maximum de vraisemblance, et d'augmenter la taille et donc la précision à mesure que l'on s'en approche. De même, ils proposent d'arrêter l'algorithme lorsque la différence entre la valeur courante du paramètre et la valeur obtenue à l'itération précédente est inférieure à un seuil fixé préalablement. Comme ce critère peut être vérifié par le simple fait du hasard, à cause du caractère aléatoire de l'algorithme, les auteurs préconisent d'arrêter l'algorithme lorsque ce critère est vérifié pour trois itérations consécutives.

Levine et Casella (2001) proposent une extension de la méthode de Booth et Hobert (1999) adaptée au cas MCMC, en utilisant la méthode de ré-échantillonnage présentée dans Robert et al. (1999) pour estimer la région de confiance. Levine et Fan (2004) étendent la méthode et donnent une formule explicite pour l'augmentation de la taille de l'échantillon, contrairement aux deux approches précédentes où m_k est augmentée par une quantité arbitraire. Cependant, les deux étapes E et M de l'algorithme doivent être appliquées deux fois à chaque itération, une fois sur l'échantillon complet, et une fois sur le sous-échantillon généré selon la méthode de Robert et al. (1999).

L'approche que nous avons adoptée est celle développée par Caffo et al. (2005), et qui consiste à retrouver la propriété de monotonie de l'algorithme EM. Nous avons vu qu'à cause de l'erreur de Monte-Carlo, les algorithmes MCEM et MCMC-EM ne garantissaient pas une augmentation de la valeur de Q à chaque itération. Caffo et al. (2005) proposent un critère permettant de retrouver cette propriété avec une forte probabilité. Plus précisément, à chaque itération k, on calcule un intervalle de confiance de niveau α pour $\Delta Q_k = Q(\theta^k;\theta^{k-1}) - Q(\theta^{k-1};\theta^{k-1})$, basé sur l'approximation normale suivante (Booth et Hobert, 1999 ; Caffo et al., 2005) :

$$\sqrt{m_k}\left(\Delta\hat{Q}_k - \Delta Q_k\right) \longrightarrow \mathcal{N}(0,\sigma_Q^2), \tag{2.45}$$

où $\Delta\hat{Q}_k = \hat{Q}(\theta^k;\theta^{k-1}) - \hat{Q}(\theta^{k-1};\theta^{k-1})$, et où $\hat{Q}$ est calculée selon l'équation (2.31). Si la borne inférieure ALB de cet intervalle de confiance est supérieure à 0, on accepte la nouvelle estimation θ^k, et sinon, on augmente la taille de l'échantillon. Une augmentation géométrique est suggérée par les auteurs, c'est-à-dire de type $m_k \leftarrow m_k + m_k/c$, où $c = 2, 3, \ldots$ D'un point de vue pratique, cela revient à générer un nouvel échantillon de type MCMC que l'on annexe à l'échantillon courant, puis à ré-évaluer l'intervalle de confiance. La procédure est répétée jusqu'à ce que le candidat soit accepté. Un critère d'arrêt peut également être dérivé, à partir de la borne supérieure AUB d'un intervalle de confiance de niveau γ pour ΔQ_k.

Si l'on dispose d'un estimateur consistant $\hat{\sigma}_Q$ de σ_Q (voir plus bas la méthode d'estimation proposée), on obtient les deux bornes suivantes :

$$ALB = \Delta\hat{Q}_k - z_{1-\alpha}\frac{\hat{\sigma}_Q}{\sqrt{m_k}} \tag{2.46}$$

$$AUB = \Delta\hat{Q}_k + z_{1-\gamma}\frac{\hat{\sigma}_Q}{\sqrt{m_k}}, \tag{2.47}$$

$$\tag{2.48}$$

où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite.

Une fois le candidat θ^k accepté, on détermine la taille d'échantillon à l'itération suivante à l'aide de l'approximation suivante :

$$\Delta\hat{Q}_{k+1} \sim \mathcal{N}\left(\Delta\hat{Q}_k, \frac{\hat{\sigma}_Q^2}{m_{k+1}}\right). \tag{2.49}$$

En imposant que la taille de l'échantillon à l'itération $k+1$ soit au moins égale à celle de l'itération k, on obtient le résultat suivant :

$$m_{k+1} = \max\left(m_k, (z_{1-\alpha} + z_{1-\beta})^2\frac{\hat{\sigma}_Q^2}{\Delta\hat{Q}_k^2}\right), \tag{2.50}$$

où β correspond à l'erreur de deuxième espèce, c'est-à-dire à la probabilité que $ALB < 0$ alors que $\Delta Q > 0$.

Concernant le critère d'arrêt, on considère que la fonction Q s'est suffisamment stabilisée lorsque la quantité AUB est inférieure à une valeur seuil δ fixée par l'utilisateur.

Cette méthode possède l'avantage de reposer sur l'évaluation d'une quantité de dimension 1, plus facile à calculer que les quantités proposées par Booth et Hobert (1999) et Levine et Casella (2001). Elle permet également d'obtenir une taille d'échantillon suffisamment grande lors de la dernière itération, ce qui permet une meilleure précision dans l'estimation des intervalles de confiance des paramètres. Cependant, elle nécessite l'utilisation d'une méthode adéquate pour estimer la variance σ_Q^2 dans le cas MCMC où les réalisations ne sont pas indépendantes et où le théorème central limite classique ne s'applique plus.

Plusieurs approches existent pour estimer la variance σ_Q^2 : les méthodes basées sur la décomposition spectrale de la variance, les méthodes de type « batch means » (BM, Bratley et al. (1987)) et les méthodes de type « regenerative simulation » (RS, Mykland et al. (1995)). Ces deux dernières méthodes consistent à diviser la chaîne en sous-échantillons considérés comme indépendants : dans la méthode BM, la taille des sous-échantillons est fixée à l'avance, alors que dans la méthode RS, les temps de régénération de la chaîne sont identifiés. Jones et al. (2006) ont comparé ces deux approches et ont montré que l'utilisation d'une méthode BM où la taille des sous-échantillons augmente avec la taille totale de la chaîne permet d'obtenir, comme avec la méthode RS, un estimateur consistant de la variance σ_Q^2. Cette méthode a également l'avantage d'être plus simple à implémenter que la méthode RS. Les auteurs suggèrent alors d'utiliser $a_m = \lfloor m_k^{1/2} \rfloor$ pour définir la taille des sous-échantillons. Flegal et Jones (2010), quant à eux, ont comparé les méthodes basées sur la décomposition spectrale et les méthodes de type batch means et ont montré que les deux approches fournissent des résultats similaires. Dans la méthode BM classique, la chaîne est divisée en sous-échantillons distincts, mais il est possible de définir des sous-échantillons qui se chevauchent les uns les autres. Flegal et Jones (2010) ont notamment montré que la méthode OBM (pour « overlapping batch means ») permet d'obtenir un estimateur consistent de l'erreur de Monte-Carlo, tout en étant plus stable que la méthode BM. En effet, si l'on fixe la taille des sous-échantillons à une valeur a_m, avec $m = a_m b_m + r_m$ où m est la taille totale de la chaîne, on obtient avec la méthode BM entre b_m et $b_m + 1$ sous-échantillons selon la valeur de r_m, mais on obtient $m - a_m + 1$ sous-échantillons avec la méthode OBM. Les deux méthodes sont présentées dans l'encadré ci-dessous.

Méthode batch means

Soit $Y_1, \ldots, Y_m$ les réalisations d'une chaîne de Markov de taille m, et soit b_m la taille des sous-échantillons. On note $\bar{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i$. On a $m = a_m b_m + r_m$, et les deux méthodes BM et OBM sont définies de la façon suivante :

- **méthode BM** (sans chevauchement) : on calcule la moyenne des réalisations de la chaîne dans chacun des $a_m + 1$ sous-échantillon, en posant $\bar{Y}_j = \frac{1}{b_m}\sum_{k=1}^{b_m} Y_{jb_m+k}$ si $j = 0, \ldots, a_m - 1$, puis $\bar{Y}_{a_m} = \frac{1}{r_m}\sum_{k=1}^{r_m} Y_{a_m b_m+k}$. La variance est ensuite estimée de la façon suivante :

$$\hat{\sigma}^2_{BM} = \frac{b_m}{a_m}\sum_{j=0}^{a_m}(\bar{Y}_j - \bar{Y}_n)^2, \tag{2.51}$$

- **méthode OBM** (avec chevauchement) : on obtient dans ce cas $m - b_m + 1$ sous-échantillons, sur lesquels on calcule la moyenne $\bar{Y}_j = \frac{1}{b_m}\sum_{k=1}^{b_m} Y_{j+k}$. L'estimateur de la variance est alors défini par :

$$\hat{\sigma}^2_{OBM} = \frac{m b_m}{(n - b_m)(n - b_m + 1)}\sum_{j=0}^{n-b_m}(\bar{Y}_j - \bar{Y}_m)^2. \tag{2.52}$$

La convergence de la chaîne de Markov vers sa loi stationnaire f peut également s'avérer relativement lente, si le point de départ de l'algorithme correspond à une région de faible probabilité sous f. L'usage consiste alors à définir une période de « burn-in », c'est-à-dire un nombre d'itérations K_b pendant lesquelles on laisse la chaîne parcourir l'espace d'états sans mettre à jour l'estimation des paramètres.

2.3.6 Convergence de l'algorithme

Les hypothèses (M1)-(M3) de la section 2.2.3 concernent la convergence de l'algorithme EM. Pour la convergence de l'algorithme MCMC-EM, il est nécessaire d'ajouter des hypothèses sur l'algorithme MCMC. Fort et Moulines (2003) introduisent en ce sens une version *stable* de l'algorithme, qui consiste en la définition d'une suite de sous-ensembles compacts $(\mathcal{K}_n)$ où $\mathcal{K}_n \subsetneq \mathcal{K}_{n+1}$ et $\Theta = \bigcup_n \mathcal{K}_n$, telle qu'à l'itération $k+1$ de l'algorithme MCMC-EM, si la valeur courante du paramètre $\theta^{(k+1)}$ n'appartient pas à l'ensemble $\mathcal{K}_{p_k}$, on le ré-initialise en posant $\theta^{k+1} = \theta^0$, puis on pose $p_{k+1} = p_k + 1$ où p_k est le nombre de réinitialisations.

À partir de cette version stable de l'algorithme, les auteurs ajoutent aux hypothèses (M1)-(M3) :

- une hypothèse permettant de contrôler la norme L^p des fluctuations dues à l'approximation de l'espérance des statistiques exhaustives $\bar{t}(\theta)$ par des méthodes de Monte Carlo par chaîne de Markov (2.31). Dans la pratique, cette hypothèse est toujours valide lorsque le noyau de transition de l'algorithme MCMC produit une chaîne uniformément ergodique, ce qui est le cas par exemple de l'algorithme de Metropolis-Hastings (Robert et Casella, 1999), et de certains algorithmes adaptatifs (Andrieu et Thoms, 2008).
- une hypothèse concernant la suite $\{m_k\}$ des tailles des chaînes de Markov générées à chaque itération de l'algorithme, et qui doit vérifier $\sum_{k=0}^{\infty} m_k^{-p/2} < \infty$, où p correspond à l'exposant intervenant dans la norme L^p étudiée dans l'hypothèse précédente. Cette condition est vérifiée dès lors que la taille de la chaîne est augmentée à chaque itération, en utilisant les méthodes décrites en 2.3.5.

Dans le cas particulier où la vraisemblance des observations admet un unique point stationnaire $\hat{\theta}$, les deux hypothèses ci-dessous, couplées aux hypothèses (M1)-(M3), assurent la convergence presque sûre de la suite (θ^k) vers $\hat{\theta}$ (Théorème 2 de Fort et Moulines (2003)).

2.4 L'algorithme SAEM

Dans l'algorithme MCMC-EM présenté dans la section précédente, les simulations générées ne sont pas conservées d'une itération de l'algorithme à l'autre. De plus, la taille de la chaîne augmente avec le nombre d'itérations, ce qui peut conduire à un allongement du temps de calcul lorsque le modèle est complexe.

2.4.1 Principe général

Une alternative, basée sur la méthode d'approximation stochastique de Robbins et Monro (1951), et proposée par Delyon et al. (1999), consiste à ré-utiliser les réalisations des itérations précédentes en y associant un facteur de pondération qui décroît avec la distance à l'itération courante. Plus formellement, l'étape E est remplacée par une étape de simulation et une étape d'approximation :

Étape S : on génère m_k réalisations $\left(\phi^{k,(1)}, \ldots, \phi^{k,(m_k)}\right)$ sous la loi $f(\phi \mid y; \theta^k)$.
Étape A : à partir d'une séquence décroissante de pas positifs (γ_k), la fonction Q à l'étape k est approchée de la façon suivante :

$$\hat{Q}(\theta; \theta^k) = \hat{Q}(\theta; \theta^{k-1}) + \gamma_k \left[\frac{1}{m_k} \sum_{m=1}^{m_k} \log f(y, \phi^{k,(m)}; \theta) - \hat{Q}(\theta; \theta^{k-1}) \right]. \tag{2.53}$$

Dans le cas de la famille exponentielle, l'étape A peut se réécrire en terme de statistique suffisante :

$$t^{(k)} = t^{(k-1)} + \gamma_k \left[\frac{1}{m_k} \sum_{m=1}^{m_k} t\left(y, \phi^{k,(m)}\right) - t^{(k-1)} \right]. \tag{2.54}$$

La convergence de l'algorithme a été étudiée par plusieurs auteurs dans le cas où le modèle appartient à la famille exponentielle. Delyon et al. (1999) ont montré la convergence de l'algorithme vers un maximum local de la vraisemblance dans le cas où les $\phi^{k,(m)}$ sont indépendants. Kuhn et Lavielle (2004) ont étendu ce résultat au cas où les réalisations proviennent d'une chaîne de Markov ergodique de loi stationnaire $f(\phi|y; \theta^k)$, dans le cas particulier où les données manquantes ϕ appartiennent à un sous-ensemble compact de $\mathbb{R}^P$. Dans la pratique cependant, cette hypothèse peut sembler restrictive, en particulier lorsque l'on considère des variables aléatoires à support dans $\mathbb{R}$ (par exemple, Gaussiennes). Allassonnière et al. (2010) proposent alors une extension basée sur la méthode générale décrite dans Andrieu et al. (2005), qui consiste à rejeter le candidat obtenu à l'étape de simulation, si celui-ci varie de manière trop importante par rapport à la valeur précédente, ce qui permet d'éviter que la chaîne ne dévie trop et n'explore des régions de faible probabilité.

Dans notre cas, comme pour l'algorithme MC-EM, il n'est pas possible de simuler directement selon la loi d'intérêt, et nous allons donc générer une chaîne de Markov de loi stationnaire $f(\phi|y; \theta^k)$. Les mêmes algorithmes que ceux décrits dans la section précédente, à savoir l'algorithme de Metropolis-Hastings (section 2.3.2) et l'échantillonneur de Gibbs hybride (section 2.3.4), peuvent être appliqués.

2.4.2 Convergence de l'algorithme

De même qu'avec l'algorithme MCMC-EM, des hypothèses supplémentaires sont nécessaires pour assurer la convergence de l'algorithme SAEM. Ces hypothèses ont été données par Delyon et al. (1999) dans le cas où l'étape d'espérance E est explicite, et ont été complétées par Kuhn et Lavielle (2004) dans le cas où l'algorithme SAEM est couplé à un algorithme MCMC, lorsque l'étape E n'est pas explicite.

La première condition requise pour assurer la convergence de l'algorithme SAEM porte sur le comportement de la séquence $\{\gamma_k\}$. En particulier, celle-ci doit vérifier :

$$\forall k \in \mathbb{N}, \gamma_k \in [0, 1], \sum_{k=1}^{\infty} \gamma_k = \infty \text{ et } \exists \lambda \in]1/2, 1] \text{ tel que } \sum_{k=1}^{\infty} \gamma_k^{1+\lambda} < \infty.$$

Delyon et al. (1999) reportent les résultats obtenus par Polyak et Juditsky (1992), qui ont montré que la méthode convergeait à une vitesse optimale pour des pas de l'ordre de $\gamma_k \propto k^{-a}$, avec $1/2 < a \leq 1$. Cependant, si des pas de grande taille peuvent permettre à l'algorithme de converger rapidement vers le voisinage de la solution, ils conduisent également à une erreur de Monte Carlo plus importante. De la même façon, choisir un pas de temps trop petit permet de réduire l'erreur, mais peut ralentir la convergence de l'algorithme (Jank, 2006). Kuhn et Lavielle (2005) proposent de commencer par un pas de taille 0, c'est-à-dire $\gamma_k = 1$, pour garantir une convergence rapide vers un voisinage du maximum de vraisemblance, puis

de diminuer le pas de temps une fois que l'on se trouve dans ce voisinage, pour assurer une convergence presque sûre de l'algorithme. On se fixe donc K_1 le nombre d'itérations pendant lesquelles $a = 0$, puis K_2 le nombre d'itérations supplémentaires pendant lesquelles le pas de temps décroît :

$$\gamma_k = \begin{cases} 1 & \text{pour } 1 \leq k \leq K_1 \\ \dfrac{1}{k - K_1 + 1} & \text{pour } k > K_1. \end{cases} \tag{2.55}$$

Pour le moment, K_1 et K_2 sont fixés empiriquement. Kuhn et Lavielle (2004) suggèrent qu'en pratique, choisir $50 < K_1 < 100$ est suffisant. Ce type de recommandations empiriques dépend évidemment beaucoup du contexte de l'étude, et il est certainement plus approprié d'évaluer la valeur de ces paramètres sur nos données, graphiquement par exemple, ou de développer des méthodes automatiques (voir par exemple Jank (2006)).

La deuxième hypothèse requise pour la convergence de l'algorithme SAEM porte sur la log-vraisemblance des observations $\ell : \Theta \mapsto \mathbb{R}$ et sur la fonction $\hat{\theta} : \mathcal{T} \mapsto \Theta$. Plus précisément, ces deux fonctions doivent être d fois dérivables, où d est la dimension de la statistique exhaustive. Dans notre cas, comme L est l'intégrale du produit de deux densités Gaussiennes, qui sont indéfiniment dérivables, et comme l'interversion entre différentiation et intégration est licite, L et donc ℓ sont également indéfiniment dérivables et en particulier, d fois dérivables. Pour la fonction $\hat{\theta}$, cela dépend de l'expression obtenue à l'étape de maximisation. Nous verrons que dans le cadre de nos applications, cette hypothèse sera également vérifiée.

La troisième condition concerne la chaîne de Markov généré par l'algorithme MCMC utilisé dans l'étape d'approximation stochastique. Tout d'abord, la chaîne doit prendre ses valeurs dans un sous-ensemble compact de $\mathbb{R}^P$. Ensuite, la densité de probabilité $\pi_\theta(x, y)$ vu comme une fonction de θ doit être lipschitzienne de constante uniforme sur (x, y). Puis, comme pour l'algorithme MCMC-EM, le noyau de transition doit produire une chaîne de Markov uniformément ergodique. Enfin, la fonction $\bar{t}$ doit être bornée sur le sous-ensemble compact dans lequel la chaîne prend ses valeurs.

Une dernière condition est nécessaire, et porte sur la suite de statistiques exhaustives $t^{(k)}$ qui doit rester dans un sous ensemble compact de $\mathcal{T}$. Cependant cette condition peut être difficile à vérifier en pratique, ou peut même s'avérer fausse, et les auteurs proposent alors d'avoir recours dans ce cas à une version stable de l'algorithme, où l'on réinitialise l'algorithme dès que la suite sort d'un sous ensemble compact fixé.

2.5 Estimation de la vraisemblance

L'algorithme EM et ses extensions présentées ci-dessus nous ont permis d'obtenir des estimateurs du maximum de vraisemblance (ou, qui convergent vers ces estimateurs), sans avoir besoin de calculer cette vraisemblance explicitement. Cependant, il peut s'avérer utile de disposer également d'une estimation de la fonction de vraisemblance en un point θ^* fixé, dans le but de comparer différents modèles entre eux à l'aide des critères AIC et BIC, par exemple, ou pour réaliser des tests du rapport de vraisemblance. Dans cette optique, des méthodes d'approximation de la vraisemblance peuvent être utilisées. Certaines ont été évoquées dans les sections 2.1.2 et 2.1.3 (voir aussi Pinheiro et Bates (1995), pour une revue détaillée des méthodes d'approximation de la vraisemblance). Nous présentons ici une méthode dite « exacte » reposant sur l'échantillonnage d'importance.

Supposons, comme cela sera le cas généralement, que l'on souhaite évaluer la log-vraisemblance du modèle au point θ^* où θ^* est la valeur de θ qui maximise la vraisemblance. On a donc :

$$
\begin{aligned}
l(\theta^*) &= \log f(y;\theta^*) \\
&= \sum_{i=1}^{s} \log f(y_i;\theta^*),
\end{aligned}
\tag{2.56}
$$

où chaque fonction $f(y_i;\theta^*)$ est définie par l'intégrale suivante :

$$
f(y_i;\theta^*) = \int f(y_i,\phi_i;\theta^*)\, d\phi_i = \int f(y_i \mid \phi_i;\theta^*) f(\phi_i;\theta^*)\, d\phi_i. \tag{2.57}
$$

Ces intégrales ne peuvent pas être évaluées analytiquement, mais peuvent cependant être approchées par des méthodes d'échantillonnage d'importance (« importance sampling »), à θ^* fixé. Ces méthodes permettent donc d'approcher la valeur de la vraisemblance en un point donné, mais ne peuvent pas être utilisées dans la pratique pour approcher la fonction de vraisemblance car cela nécessiterait d'approcher l'intégrale (2.57) pour chaque valeur de θ ...

Une première idée pour approcher l'intégrale serait de simuler des réalisations i.i.d. $\phi_i^{(1)},\ldots,\phi_i^{(N)}$ de loi $f(\phi_i;\hat{\theta})$ pour chaque individu i, où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ. Cependant, il est possible de contrôler plus efficacement la variance de l'estimateur de $f(y_i;\hat{\theta})$ en utilisant les méthodes d'échantillonnage d'importance (Robert et Casella, 1999). Cette approche correspond en fait à la méthode de Monte Carlo présentée en section 2.3, mais en utilisant la relation suivante, où q_i est la densité de probabilité d'une loi instrumentale absolument continue par rapport à $f(\phi_i;\theta)$ (Robert et al., 1999) :

$$
\int f(y_i \mid \phi_i;\theta^*) f(\phi_i;\theta^*)\, d\phi_i = \int f(y_i \mid \phi_i;\theta^*)\, \frac{f(\phi_i;\theta^*)}{q_i(\phi_i;\theta^*)}\, q_i(\phi_i;\theta^*)\, d\phi_i. \tag{2.58}
$$

Il suffit alors de simuler pour chaque individu i, N variables aléatoires i.i.d. de loi q_i, $\phi_i^{(1)},\ldots,\phi_i^{(N)}$ puis d'utiliser l'approximation suivante pour chacune des intégrales (2.57) :

$$
\hat{f}(y_i;\theta^*) = \frac{1}{N}\sum_{m=1}^{N} f(y_i \mid \phi_i^{(m)};\theta^*)\, \frac{f(\phi_i^{(m)};\theta^*)}{q_i(\phi_i^{(m)};\theta^*)} \tag{2.59}
$$

Le choix de la loi instrumentale q_i va déterminer les performances de l'estimateur, en particulier si elle est suffisamment proche de la loi conditionnelle $f(\phi_i \mid y_i;\hat{\theta})$.

On peut par exemple utiliser une loi instrumentale Gaussienne de même espérance et même variance que la loi $f(\phi_i \mid y_i;\hat{\theta})$. Pour cela, on estime les quantités $\mathbb{E}(\phi_i \mid y_i;\hat{\theta})$ et $\mathrm{Var}(\phi_i \mid y_i;\hat{\theta})$ pour chaque individu i. En notant $\hat{\mu}_i$ un estimateur de $\mathbb{E}(\phi_i \mid y_i;\hat{\theta})$, et $\hat{\Omega}_i$ un estimateur de $\mathrm{Var}(\phi_i \mid y_i;\hat{\theta})$, on peut alors utiliser la loi normale $\mathcal{N}(\hat{\mu}_i,\hat{\Omega}_i)$ pour simuler $\phi_i^{(1)},\ldots,\phi_i^{(N)}$:

- dans le cas de l'algorithme MCMC-EM, ces estimations peuvent être obtenues à partir de la chaîne de Markov simulée à la dernière itération de l'algorithme, dont la loi cible est $f(\phi_i \mid y_i;\hat{\theta})$,
- dans le cas de l'algorithme SAEM, la taille de la chaîne étant faible à chaque itération de l'algorithme, il est nécessaire de générer une nouvelle chaîne de taille suffisamment importante.

L'expression (2.59) peut alors s'écrire :

$$
\hat{f}(y_i;\theta^*) = \frac{1}{N}\sum_{m=1}^{N}\prod_{n=1}^{n_i} f(y_{in}\phi_i^{(m)};\theta^*)\; \frac{|\Omega^*|^{-1/2}}{|\hat{\Omega}|^{-1/2}}\, \frac{\exp\left[-\frac{1}{2}(\phi_i^{(m)}-\beta^*)^t(\Omega^*)^{-1}(\phi_i^{(m)}-\beta^*)\right]}{\exp\left[-\frac{1}{2}(\phi_i^{(m)}-\hat{\beta})^t\hat{\Omega}^{-1}(\phi_i^{(m)}-\hat{\beta})\right]}.
$$

On obtient ainsi un estimateur sans biais de la vraisemblance au point θ^* :

$$\hat{L}(\theta^*) = \prod_{i=1}^{s} \hat{f}(y_i; \theta^*), \tag{2.60}$$

puis un estimateur (biaisé) de la log-vraisemblance

$$\hat{\ell}(\theta^*) = \log \hat{L}(\theta^*) = \sum_{i=1}^{s} \log \hat{f}(y_i; \theta^*).$$

En effet, l'inégalité de Jensen nous assure que $\mathbb{E}(\hat{\ell}(\theta^*)) < \ell(\theta^*)$. Cependant, ce biais diminue à mesure que N augmente, et si la loi instrumentale choisie est suffisamment proche de la loi conditionnelle des ϕ_i sachant y_i.

3 Évaluation du modèle

3.1 Structure de covariance

Il est possible de tester si la variance d'un effet aléatoire est significativement non nulle, et si l'effet en question doit donc être considéré comme variable au sein de la population, ou s'il peut être considéré comme fixe. Supposons que l'on souhaite tester le caractère aléatoire du $k^{\text{ème}}$ paramètre, c'est-à-dire tester l'hypothèse nulle suivante :

$$H_0 : \{\sigma_k^2 = 0\} \quad \text{vs.} \quad H_1 : \{\sigma_k^2 > 0\}, \tag{2.61}$$

où σ_k est le $k^{\text{ème}}$ élément situé sur la diagonale de la matrice Γ. On peut utiliser un test du rapport de vraisemblance (TRV) en définissant la statistique de test suivante :

$$T = -2(\ell_0(\theta) - \ell_1(\theta)),$$

où ℓ_0 est la log-vraisemblance sous H_0 et ℓ_1 la log-vraisemblance sous H_1. Usuellement dans ce type de test, la statistique T suit sous l'hypothèse nulle une loi du chi-deux dont le degrés de liberté correspond au nombre de paramètres sous H_1 moins le nombre de paramètres sous H_0. Cependant ici, l'hypothèse nulle correspond à la frontière de l'espace du paramètre σ_k^2, ce dernier étant positif ou nul par définition. Self et Liang (1987) ont montré que, dans ce cas, la loi de la statistique de test sous l'hypothèse nulle est un mélange de deux lois du chi-deux : $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_0^2$, où la loi χ_0^2 correspond à la masse de Dirac en 0. D'un point de vue pratique, cela signifie que les p-values associées aux tests du rapport de vraisemblance sont divisées par deux par rapport au cas où l'on supposerait uniquement une loi χ_1^2 pour T sous H_0.

3.2 Erreur de prédiction sur la distribution

L'une des étapes essentielles lors de la construction d'un modèle est celle de sa validation. L'importance de cette étape de validation a notamment été discutée dans le chapitre précédent, où nous précisions en particulier que le choix des critères d'évaluation dépendait de l'objectif pour lequel le modèle avait été construit initialement. Par exemple, lorsque le but est d'obtenir un outil prédictif, les critères présentés au chapitre précédent s'avèrent particulièrement adaptés, alors que si l'objectif est de construire un modèle purement descriptif, d'autres critères seront utilisés. Il convient donc, comme précisé au chapitre précédent, de bien définir en amont l'objectif du modèle.

Dans le cadre des applications de ce chapitre, notre principal objectif sera essentiellement descriptif. Le but est de décrire le mieux possible la population étudiée. Pour cela, Mentré et Escolano (2006) proposent, dans le cadre des modèles non linéaires mixtes, d'utiliser un critère basé sur les *écarts de prédiction* du modèle. À partir des estimations obtenues grâce aux algorithmes décrits précédemment, il est possible d'estimer la distribution des prédictions à l'aide de méthodes de Monte-Carlo. On définit ensuite l'écart de prédiction pour l'individu i sous la condition t_{ij}, e_{ij}, comme la valeur de la fonction de répartition des prédictions au point y_{ij}.

Plus précisément, en notant $Y_{ij}, i = 1, \ldots, s, j = 1, \ldots, n_i$ la prédiction du modèle pour l'individu i et la condition t_{ij}, on a :

$$e_{ij} = \mathbb{P}(Y_{ij} \leq y_{ij}), \tag{2.62}$$

où y_{ij} est l'observation pour l'individu i sous la condition t_{ij} (voir section 1). La loi des Y_{ij} est inconnue, mais la fonction de répartition peut être approchée par des méthodes de Monte Carlo. Pour cela, on simule M jeux de données $(Y_{ij}^m), m = 1, \ldots, M$ selon le modèle (2.1) - (2.2), à partir de l'estimateur $\hat{\theta}$. On approche ensuite e_{ij} par le quantile empirique correspondant, c'est-à-dire par

$$\hat{e}_{ij} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}_{Y_{ij}^m \leq y_{ij}}. \tag{2.63}$$

Sous l'hypothèse nulle $\{H_0$: le modèle décrit bien les données$\}$, ces écarts de prédiction suivent une loi uniforme sur $[0, 1]$. Cependant, les erreurs de prédictions d'un même individu sont corrélées, et les tests usuels du type Kolmogorov-Smirnov, qui requièrent l'indépendance des données, ne peuvent pas être utilisés. Comets et al. (2008) ont alors proposé d'utiliser une version décorrélée de ces erreurs de prédiction. Pour chaque individu, on note Y_i^m le vecteur de prédictions du jeu de données simulé m. On calcule ensuite sa moyenne et sa variance empiriques sur les M échantillons de Monte-Carlo, puis on définit la prédiction décorrélée $Y_i^{m,*}$:

$$Y_i^{m,*} = (\text{Var } Y_i)^{-1/2} \left(Y_i^m - \mathbb{E}(Y_i)\right). \tag{2.64}$$

Les *erreurs de prédiction* $\hat{e}_{ij}^*$ sont ensuite obtenues de la même façon que les écarts de prédiction, mais en utilisant $Y_i^{m,*}$ au lieu de Y_i^m. Pour finir, on définit alors les *erreurs normalisées de prédiction* comme :

$$\text{npde}_{ij} = \Phi^{-1}(\hat{e}_{ij}^*). \tag{2.65}$$

Ainsi normalisées, les erreurs de prédiction suivent, sous l'hypothèse nulle, une loi normale centrée réduite.

Chapitre 3

Modélisation de la variabilité inter-plantes

"La plupart des hommes ont, comme les plantes, des propriétés cachées que le hasard fait découvrir."

François de la Rochefoucauld, *Maximes*.

Dans le chapitre précédent, nous avons présenté une approche basée sur les modèles à effets mixtes, permettant de prendre en compte à la fois la variabilité intra- et inter-individuelle dans une population. La variabilité intra-individuelle est représentée par une fonction non linéaire décrivant l'évolution des observations d'un même individu, et la variabilité inter-individuelle, celle qui nous intéresse plus spécifiquement, est prise en compte en introduisant des effets aléatoires dans le modèle.

Dans ce chapitre, nous présentons deux applications de cette méthodologie à la modélisation de la variabilité inter-plantes. Nous nous sommes intéressés tout d'abord à la variabilité de l'organogenèse chez la betterave sucrière (section 1), à l'aide d'un modèle mixte linéaire par morceaux, permettant de prendre en compte les deux phases de développement observées (voir Chapitre 1, section 1.1.1). Ce modèle a été implémenté sous Monolix (The Monolix Team, 2011), logiciel dédié à l'estimation dans les modèles non linéaires mixtes et basé sur l'algorithme SAEM présenté au chapitre précédent.

Puis, nous proposons en section 2 une extension du modèle Greenlab à l'échelle de la population. Les performances des deux algorithmes d'estimation présentés dans le chapitre précédent ont été comparés sur des jeux de données simulées, puis sur des jeux de données réelles provenant de la betterave et du colza, en supposant un modèle de bruit portant uniquement sur les observations.

1 L'organogenèse chez la betterave

Nous présentons dans cette section une application des méthodes précédentes au modèle d'organogenèse chez la betterave dont la variabilité génétique est assez forte, ne s'agissant pas d'une lignée pure.

Cette variabilité s'observe en particulier sur le nombre de feuilles, qui peut être très différent d'une plante à l'autre, même dans des conditions environnementales identiques. Or, les plantes étant en compétition constante pour la lumière, la capacité pour certaines d'entre elles de produire plus de feuilles, ou plus rapidement que leurs voisines, peut leur garantir un meilleur accès à la lumière et leur permettre de produire plus de biomasse. Liu et al. (2004) ont par exemple observé d'importantes variations de rendement lorsque le rythme d'émission des feuilles ou la vitesse d'émergence différaient d'une plante à l'autre. L'interception lumineuse étant directement reliée à la production de biomasse (voir chapitre 1), chaque

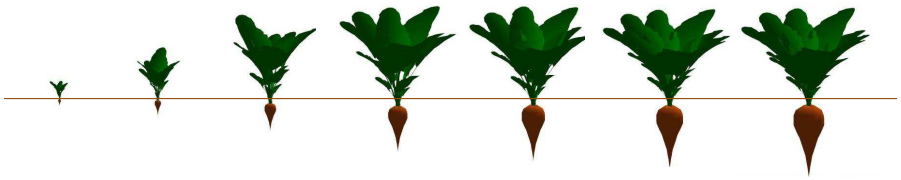


FIG. 3.1 – Croissance de la betterave sucrière simulée par le logiciel DigiPlant (Cournède et al., 2006), aux cycles de croissance 15, 17, 20, 26, 30, 34, 39 et 43.

facteur ayant une influence sur la vitesse d'expansion de la surface foliaire et donc sur la surface foliaire totale, pourra avoir un impact sur le rendement.

Le rythme d'apparition des feuilles est donc un paramètre crucial dans le développement de la plante. Il est en général défini à l'aide de son inverse, le phyllochrone (voir aussi chapitre 1, section 1.1.1). La variabilité du phyllochrone a été étudiée pour plusieurs espèces, et plusieurs facteurs environnementaux ayant une influence sur ce paramètre ont été identifiés. Dans leur étude du sorgho, Clerget et al. (2008) montrent par exemple qu'il existe une corrélation positive entre la température du sol et le phyllochrone, et une corrélation négative entre la photopériode et la longueur du jour d'une part, et le phyllochrone d'autre part. Des résultats similaires ont été observés par Cao et Moss (1989) pour le blé et l'orge, et une courte revue des facteurs ayant une influence sur le phyllochrone a été proposée par Wilhelm et McMaster (1995). Parmi les facteurs pouvant provoquer une hausse du phyllochrone et donc un ralentissement du rythme d'apparition des feuilles, on retrouve la température, un stress hydrique très important, ou même de fortes concentrations en sel. À l'inverse, le phyllochrone décroît lorsque la concentration en CO_2 augmente, ou lorsque la quantité et la qualité de la lumière diminuent.

Dans le cas de la betterave, Milford et al. (1985a,b) ont observé deux phases distinctes dans le développement de la plante, ce qui les conduit à définir deux phyllochrones, un pour chaque phase. Ils ont observé, en comparant plusieurs années d'expérimentations et plusieurs traitements agricoles (avec ou sans irrigation, avec ou sans engrais, en faisant varier la densité de plantation et la date de semis), que le phyllochrone de la première phase restait stable, mais que la durée de cette première phase, ainsi que le phyllochrone de la deuxième phase, étaient plus variables. Lemaire et al. (2008) ont également observé ces deux phases successives dans le développement de la betterave, avec une première phase qui s'étend de l'émergence jusqu'à l'apparition de la vingtième feuille environ, puis une seconde phase de ralentissement du développement foliaire correspondant à un plus grand phyllochrone. Plusieurs hypothèses peuvent être avancées pour expliquer ce phénomène : Milford et al. (1985a) suggèrent par exemple un changement dans la température de base servant au calcul du temps thermique (voir équation 1.6), et une augmentation de la compétition pour les ressources entre les feuilles et la racine. De leur côté, Lemaire et al. (2008, 2009) ont montré que ce changement intervenait au début de la phase linéaire de croissance racinaire, et au moment de la couverture du sol par le feuillage, lorsque la compétition pour la lumière augmente. Cette cassure se retrouve également chez certaines autres plantes, notamment chez le colza, mais dans ce dernier cas on assiste plutôt à une accélération qu'à un ralentissement du rythme d'émission des feuilles.

Néanmoins, si le phyllochrone moyen reste stable, il est fortement variable d'une plante à l'autre (Frank et Bauer, 1995), ce qui, associé à la variabilité du temps de germination ou d'émergence (Dürr et Boiffin, 1995), implique une forte variabilité du nombre total de feuilles (voir figure 3.2).

Malgré cette forte variabilité inter-individus, peu d'études portant sur le phyllochrone la prennent en compte. La plupart d'entre elles sont basées sur des modèles à effets fixes, soit basés sur toute la population

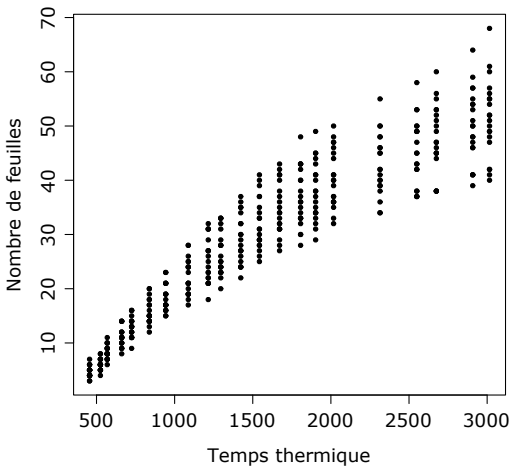


FIG. 3.2 – Nombre de feuilles en fonction du temps thermique pour 20 plantes poussant dans des conditions standard de densité (10.89 pl/m^2) et d'apport en fertilisants (136 kg/ha).

de plantes, et reposant donc sur l'hypothèse (irréaliste) selon laquelle les données provenant d'un même individu sont indépendantes (voir par exemple Xue et al. (2004) ; Frank et Bauer (1995) ; Bauer et al. (1985) ; Streck et al. (2005) ; Juskiw et al. (2005)), soit basés sur des valeurs moyennes, ce qui permet de contourner le problème des corrélations entre mesures, mais implique également une perte importante d'information (Lemaire et al., 2009). En prenant l'exemple de l'engrais, même si la littérature est relativement abondante concernant l'influence que peut avoir l'administration d'azote sur la croissance et le développement de la betterave, son effet sur chacun des paramètres d'organogenèse a rarement été étudié. Lee et Schmehl (1988) ont observé un effet non significatif de l'azote sur le phyllochrone, mais un effet significatif de l'interaction azote-date de récolte, et azote-date de récolte-date de semis. A contrario, Stout (1961) a montré qu'un niveau élevé d'azote pouvait stimuler la croissance de nouvelles feuilles. Quoiqu'il en soit, les résultats de ces études reposent sur des hypothèses d'indépendance entre mesures d'une même plante, ou sur des valeurs moyennes, et doivent donc être interprétés avec précaution. En effet, il n'est pas possible avec ce type d'approches de distinguer la variance imputable à la variabilité inter-individuelle de la variance résiduelle, toutes les sources de variabilité étant alors réunies sous un même terme d'erreur, ce qui peut conduire à sous-estimer ou à sur-estimer la significativité des tests statistiques.

Nous proposons dans cette section un modèle d'organogenèse non linéaire mixte, permettant de modéliser l'évolution du nombre de feuilles en fonction du temps thermique. Le modèle sera appliqué dans un premier temps sur une population standard, puis adapté pour tester l'effet des facteurs environnementaux, en prenant pour exemple l'effet de l'azote. Le modèle a été entièrement implémenté sous le logiciel Monolix (The Monolix Team, 2011), dans lequel l'estimation se fait par l'algorithme SAEM présenté en section 2.4.

1.1 Formulation du modèle

Afin de tenir compte des deux phases de développement observées chez la betterave par Milford et al. (1985a) et Lemaire et al. (2008), nous avons utilisé un modèle linéaire par morceaux, avec quatre paramètres dont l'interprétation biologique est immédiate : le temps d'initiation (correspondant au temps de germination), le rythme d'apparition des feuilles au cours de la première phase, le temps de rupture cor-

respondant au début de la deuxième phase de développement, et le rythme d'apparition des feuilles lors de cette deuxième phase.

Comme détaillé en section 1, le modèle non linéaire mixte peut s'écrire sous forme de modèle hiérarchique à deux niveaux : un premier niveau dans lequel on s'intéresse à la variabilité intra-individuelle, et un second dans lequel on modélise la variabilité inter-indivduelle. Nous notons y_{ij} le nombre de feuilles de la plante i ($i = 1, \ldots, s$) au temps thermique t_j ($j = 1, \ldots, n_i$).

Variabilité intra-individuelle :

$$y_{ij} = f(t_j, \phi_i) + g(t_j, \phi_i)e_{ij} \tag{3.1}$$

où ϕ_i est le vecteur de paramètres spécifiques à la plante i, et e_{ij} un terme d'erreur avec $\mathcal{N}(0, 1)$. Par rapport à l'équation (2.1), nous avons introduit un modèle d'erreur résiduel par l'intermédiaire de la fonction g. Ceci permet notamment de tester des modèles d'erreur additifs en posant $g = \sigma$, ou multiplicatifs en posant par exemple $g = \sigma f$. Des modèles combinés peuvent également être testés, avec $g = a + bf$.

La fonction f représente l'évolution non linéaire du nombre de feuilles en fonction du temps thermique. Dans notre cas, il s'agit d'une fonction linéaire par morceaux définie de la façon suivante :

$$f(t_j, \phi_i) = \phi_{i,1}(t_j - \phi_{i,0})\ \mathbf{1}_{t_j \geq \phi_{i,0}} + \phi_{i,3}(t_j - \phi_{i,2})\ \mathbf{1}_{t_j \geq \phi_{i,2}} \tag{3.2}$$

où $\phi_{i,0}$ est le temps thermique d'initiation, $\phi_{i,1}$ le rythme d'apparition des feuilles lors de la première phase, $\phi_{i,2}$ le temps de rupture et $\phi_{i,3}$ le changement de rythme observé au cours de la deuxième phase, pour la plante i. On modélise ainsi le changement de pentes entre les deux phases de développement, et on force les deux segments de droite à s'intercepter au point d'abscisse le temps de rupture. Les phyllochrones des deux phases de développement se déduisent aisément à partir des paramètres $\phi_{i,1}$ et $\phi_{i,3}$: en notant $\gamma_{i,1}$ et $\gamma_{i,2}$ les phyllochrones des première et deuxième phases de développement, respectivement, on a les relations suivantes :

$$\begin{aligned} \gamma_{i,1} &= \frac{1}{\phi_{i,1}} \\ \gamma_{i,2} &= \frac{1}{\phi_{i,1} + \phi_{i,3}}. \end{aligned} \tag{3.3}$$

Variabilité inter individuelle :

$$\phi_i = A_i\beta + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \Gamma) \tag{3.4}$$

où A_i est une matrice de design connue, β le vecteur des effets fixes, et ξ_i le vecteur d'effets aléatoires associé à la plante i, et Γ la matrice 4×4 de covariance des effets aléatoires.

La matrice A_i permet de tenir compte de l'effet de certaines covariables sur les paramètres. Dans notre cas, nous avons considéré deux formulations différentes pour A_i, selon que l'on souhaite évaluer l'effet de covariables ou non :

- dans le cas où l'on s'intéresse à la population standard, aucune covariable n'est introduite dans le modèle, et on a donc $A_i = I_4$ et $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^t$
- lorsque l'on s'intéresse à l'effet de certaines covariables, et plus précisément ici à l'effet de l'azote, on considère que la valeur moyenne dans la population varie selon la dose reçue. Trois niveaux ont été comparés, et donc, deux covariables ont été introduites dans le modèle sous forme de variables indicatrices : s_i, qui vaut 1 si la plante i a reçu une dose standard d'azote, et 0 sinon, et h_i, qui vaut

1 si la plante i a reçu une dose élevée d'azote et 0 sinon. Par conséquent, les plantes n'ayant pas reçu d'azote sont celles pour lesquelles $s_i = 0$ et $h_i = 0$. Finalement, nous avons :

$$A_i = \begin{pmatrix} 1 & s_i & h_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & s_i & h_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & s_i & h_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & s_i & h_i \end{pmatrix}$$

et $\beta = (\beta_0, \delta_{s,0}, \delta_{h,0}, \beta_1, \delta_{s,1}, \delta_{h,1}, \beta_2, \delta_{s,2}, \delta_{h,2}, \beta_3, \delta_{s,3}, \delta_{h,3})^t$, où $\delta_{s,j}$ représente la différence entre la valeur moyenne du paramètre $\phi_{i,j}$ pour les plantes ayant reçu une dose standard (s) d'azote, et la valeur moyenne de ce même paramètre pour les plantes n'ayant pas reçu d'azote. On peut définir de la même façon le paramètre $\delta_{h,j}$ pour les plantes recevant une forte dose d'azote (h). Finalement, pour les plantes ne recevant pas d'azote, la moyenne du vecteur d'effets aléatoires est $(\beta_0, \beta_1, \beta_2, \beta_3)^t$, pour les plantes recevant une dose standard d'azote, cette moyenne vaut $(\beta_0 + \delta_{s,0}, \beta_1 + \delta_{s,1}, \beta_2 + \delta_{s,2}, \beta_3 + \delta_{s,3})^t$, et enfin, pour les plantes ayant reçu une forte dose d'azote, cette moyenne vaut $(\beta_0 + \delta_{h,0}, \beta_1 + \delta_{h,1}, \beta_2 + \delta_{h,2}, \beta_3 + \delta_{h,3})^t$. La variance des effets aléatoires reste la même quelque soit la dose d'azote reçue.

Le vecteur de paramètres est alors $\theta = (\beta, \Gamma, \sigma)$ si $g = \sigma$ ou $g = \sigma f$, et $\theta = (\beta, \Gamma, a, b)$ si $g = a + bf$.

Il est possible d'approcher la moyenne et la variance des deux phyllochrones (voir équation (3.3)) à l'aide d'un développement de Taylor à l'ordre 1. Si X est une variable aléatoire de moyenne m et de variance σ^2, et si h est une fonction dérivable telle que $h'(m) \neq 0$, on peut utiliser les approximations suivantes :

$$\begin{aligned} \mathbb{E}(h(X)) &\approx h(m) \\ \text{Var}(h(X)) &\approx (h'(m))^2 \, \text{Var}(X). \end{aligned} \tag{3.5}$$

L'utilisation d'une matrice de covariance Γ non diagonale et sans structure particulière augmente le nombre de paramètres à estimer, ce qui, compte tenu du faible échantillon dont nous disposons, peut conduire à des problèmes d'identifiabilité ou de convergence de l'algorithme. C'est la raison pour laquelle nous avons également testé d'autres structures de covariance plus parcimonieuses :

1. on considère que le paramètre $\phi_{i,k}$ n'est corrélé à aucun autre $\phi_{i,l}$ pour $l \neq k$. Par exemple, pour $k = 0$, cela signifie que le temps d'initiation est indépendant des trois autres paramètres, ce qui donne la forme suivante pour Γ :

$$\begin{pmatrix} \sigma_0^2 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ 0 & \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ 0 & \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}, \tag{3.6}$$

2. on considère que les paramètres $\phi_{i,k}$ et $\phi_{i,l}$, où $k \neq l$, ne sont corrélés avec aucun autre paramètre. Par exemple, pour $k = 0$ et $l = 1$on obtient la matrice de covariance suivante :

$$\begin{pmatrix} \sigma_0^2 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 \\ 0 & 0 & \sigma_2^2 & \sigma_{23} \\ 0 & 0 & \sigma_{23} & \sigma_3^2 \end{pmatrix}, \tag{3.7}$$

3. on considère enfin l'indépendance entre les paramètres, ce qui équivaut à une matrice Γ diagonale.

Les différentes structures de covariance ont été comparées à l'aide des critères AIC et BIC.

1.2 Estimation sous Monolix

Le logiciel Monolix (MOdèles NOn LInéaires à effets miXtes) a été initié en 2003 à l'initiative d'un groupe de travail formé de statisticiens et de biostatisticiens, avec pour objectif initial une application en pharmacologie, plusieurs modèles spécifiques à ce domaine étant déjà implémentés par défaut. Cependant, il est possible de créer son propre modèle en langage MLXTRAN. L'estimation des paramètres du modèle se fait par l'algorithme SAEM, dans lequel l'étape de simulation directe est remplacée par une étape MCMC. La chaîne de Markov est générée par un algorithme de Metropolis-Hastings, pour lequel quatre noyaux de transition associés à quatre lois instrumentales sont utilisés successivement à chaque itération k de l'algorithme SAEM :

1. pendant m_1 itérations, on utilise la loi marginale des effets aléatoires. C'est le cas indépendant présenté en (2.36).
2. pendant m_2 itérations, on génère une permutation aléatoire du vecteur obtenu à l'itération précédente.
3. pendant m_3 itérations, on utilise une marche aléatoire gaussienne de matrice de covariance $\tau_k \Gamma^{(k)}$, où τ_k est ajusté à chaque itération pour atteindre un taux d'acceptation optimal α, fixé par défaut à 0.3.
4. pendant m_4 itérations, on utilise une marche aléatoire pour chaque composante de ϕ_i. Cela revient à utiliser un algorithme de Gibbs hybride ou Metropolis-within-Gibbs (voir section 2.3.4).

Par défaut, $m_1 = 2, m_2 = 0, m_3 = 2$, et $m_4 = 2$. En effet, comme précisé dans le manuel d'utilisateur de Monolix (The Monolix Team, 2011), la deuxième loi instrumentale n'est à recommander que dans certains cas très spécifiques liés à la pharmacocinétique.

1.3 Données

Les données que nous avons utilisées proviennent des expérimentations 2011, décrites dans le Chapitre 1, section 3.1. Trois doses différentes d'azote ont été administrées : une dose nulle pour servir de contrôle, une dose correspondant aux recommandations standard de 136 kg/ha, et une dose élevée de 196 kg/ha. Dans chacune de ces trois conditions, la densité de plantation a été estimée à 11.08 pl/m^2, 10.89 pl/m^2 et 9.14 pl/m^2 respectivement. Le nombre de feuilles a été relevé hebdomadairement et de façon non-destructive sur un groupe de 60 plantes (20 plantes pour chaque dose d'azote), prises aléatoirement dans le champ. Une feuille était considérée comme apparue lorsqu'elle avait atteint 10 mm.

TAB. 3.1 – Estimation des paramètres dans la population standard (20 plantes, dose normale d'azote). TRV : test du rapport de vraisemblance, VII : variabilité inter-individuelle. Les corrélations ρ_{ij} sont définies de la façon suivante : $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, où σ_i, σ_j et σ_{ij} sont les composantes de la matrice de covariance Γ.

Paramètre	Estimation	Écart-type	TRV pour VII
β_0	241	11	-
β_1	0.0257	0.001	-
β_2	1580	41	-
β_3	-0.0136	0.0009	-
Variabilité inter-individuelle			
σ_0	45.5	9.1	$p < 0.0001$
σ_1	0.0043	0.0007	$p < 0.0001$
σ_2	172	31	$p < 0.0001$
σ_3	0.0038	0.0006	$p < 0.0001$
Corrélations			
ρ_{01}	0.55	0.18	-
ρ_{03}	-0.52	0.19	-
ρ_{13}	-0.84	0.07	-
a	0.811	0.07	-
b	0.0032	0.0023	

1.4 Résultats

Des valeurs initiales sont requises pour l'algorithme SAEM, et ont été obtenues à l'aide d'une estimation individuelle des paramètres. Suivant la méthode décrite en section 2.1.1, un modèle à effets fixes a été utilisé pour chaque plante, conduisant à une estimation individuelle des paramètres pour chaque plante. Puis, l'estimation de la moyenne et la variance des paramètres dans la population a été utilisée comme valeur initiale de l'algorithme. D'autres valeurs initiales ont également été testées pour vérifier la consistance des résultats.

1.4.1 Population standard

Dans un premier temps, le modèle a été utilisé pour décrire la population standard des 20 plantes ayant reçu une dose d'azote de 136 kg/ha. Parmi les différentes structures de covariance testées, la meilleure au sens des critères AIC et BIC est celle correspondant au cas où le temps thermique de rupture $\phi_{i,2}$ est indépendant des autres paramètres. En ce qui concerne le modèle d'erreur, les versions additive et combinée (c'est-à-dire $g = \sigma$ et $g = a + bf$ respectivement) se sont avérées meilleures que la version multiplicative. Les deux modèles fournissent des résultats similaires en terme d'estimation, le terme correspondant à la partie multiplicative étant très faible. Finalement, nous avons retenu le modèle d'erreur combiné car l'hypothèse de normalité des résidus n'était pas vérifiée avec le modèle purement additif. Les résultats sont présentés dans la Table 3.1.

La Figure 3.3 représente la distribution des observations prédites par le modèle, et montre une bonne prise en compte de la variabilité par le modèle. Le test de normalité des résidus a fourni une p-value de 0.06. La convergence de l'algorithme a été vérifiée en utilisant plusieurs jeux de données initiaux. D'autres formulations non linéaires ont également été testées, comme suggéré par Xue et al. (2004), mais ont conduit à des valeurs de AIC et BIC plus élevées (voir tableau 3.2).

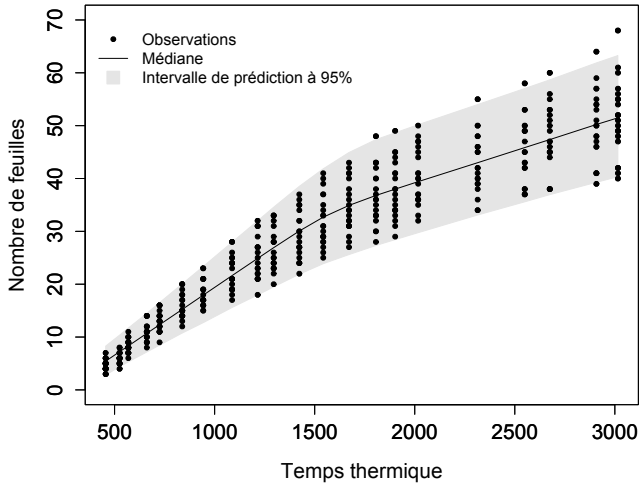


FIG. 3.3 – Prédictions du modèle sur la population standard. La ligne continue correspond à la médiane et les frontières de la zone grisée aux quantiles d'ordre 5% et 95%.

TAB. 3.2 – Comparaison des différentes versions du modèle.

Modèle	TRV (p-value)	AIC	BIC
Linéaire par morceaux et indépendance entre $\phi_{i,2}$ et les autres paramètres	0.32	1377	1390
Linéaire par morceaux et matrice Γ générale	-	1380	1396
Linéaire par morceaux et matrice Γ diagonale	< 0.0001	1399	1409
Modèle de Gompertz	-	1604	1610
Sigmoïde	-	2140	2155

En utilisant l'approximation (3.5), la moyenne et l'écart-type des deux phyllochrones ont été estimées à 39°C jour (écart-type = 6.5°C jour) pour la première phase et 83° C jour (écart-type = 16.0°C jour) pour la deuxième phase. A partir de ces valeurs moyennes, on peut prédire un nombre moyen de 35 feuilles au moment du changement de rythme. Les résultats des tests du rapport de vraisemblance pour les paramètres de variance indiquent que chacun des quatre paramètres du modèle a une variabilité inter-individuelle significative, et doit donc être traité comme un effet aléatoire. Cependant, la variabilité est plus forte lors de la deuxième phase, avec une variance plus élevée du temps de rupture par rapport au temps d'initiation et du second phyllochrone par rapport au premier.

Ceci peut être dû au fait que les différences entre plantes observées lors de la première phase de développement deviennent plus prononcées à mesure que le temps thermique augmente. En effet, les plantes qui ont démarré leur croissance plus tôt que les autres, et avec un phyllochrone plus faible, ont eu tendance à produire plus de feuilles et à pousser plus haut, projetant ainsi de l'ombre sur leurs voisines qui, de ce fait, reçoivent moins de lumière et produisent moins de feuilles. De façon générale, le temps thermique de changement de rythme est fortement influencé par l'écophysiologie de la plante et par les conditions environnementales. Lemaire (2010) a notamment montré que la densité de plantation avait une forte influence sur ce paramètre, alors que les phyllochrones eux restaient stables d'une densité à l'autre. Ce point sera également développé dans la section suivante.

Un autre résultat intéressant de notre modèle est la non corrélation entre le temps thermique de rupture et les autres paramètres, ce qui peut signifier que ce changement de rythme intervient lorsqu'un nombre

TAB. 3.3 – Estimation des paramètres pour la comparaison des doses d'azote. Dans la dernière colonne, les p-value correspondent au test de Wald lorsqu'il s'agit des covariables, et au test du rapport de vraisemblance pour les paramètres de covariance (variabilité inter-individuelle).

Paramètre	**Estimation**	**Écart-type**	p**-value**
Temps thermique d'initiation			
β_0	147	11	-
$\beta_0 + \delta_{n,0}$	242	11	< 0.001
$\beta_0 + \delta_{h,0}$	252	11	< 0.001
Rythme d'apparition des feuilles lors de la première phase			
β_1	0.0219	0.0008	-
$\beta_1 + \delta_{n,1}$	0.0257	0.0008	0.002
$\beta_1 + \delta_{h,1}$	0.0263	0.0008	< 0.001
Temps thermique de rupture			
β_2	1860	43	-
$\beta_2 + \delta_{n,2}$	1580	41	< 0.001
$\beta_2 + \delta_{h,2}$	1640	41	< 0.001
Différence de rythme d'apparition entre les deux phases			
β_3	-0.012	0.0008	-
$\beta_3 + \delta_{n,3}$	-0.0137	0.0008	0.15
$\beta_3 + \delta_{h,3}$	-0.0136	0.0008	0.18
Variabilité inter-individuelle			
σ_0	41.6	5.2	< 0.001
σ_1	0.00373	0.00035	< 0.001
σ_2	170	18	< 0.001
σ_3	0.00357	0.00036	< 0.001
Corrélations			
σ_{01}	0.54	0.11	-
σ_{03}	-0.46	0.12	-
σ_{13}	-0.77	0.06	-
σ	0.98	0.022	-

fixé de feuilles a été atteint, et ne dépend donc pas directement de la vitesse à laquelle les feuilles ont été émises par la plante, ni de la date d'émergence.

1.4.2 Comparaison des doses d'azote

Les résultats de la comparaison des différentes doses d'azote sont présentés dans le tableau 3.3, et la distribution prédite par le modèle est représentée sur la Figure 3.4.

Le temps thermique d'initiation a été plus tardif pour les plantes ayant reçu un traitement azoté ($p < 0.001$ pour les deux doses). Ceci peut s'expliquer par l'effet néfaste que peut avoir l'azote sur la germination, en particulier s'il se trouve concentré en trop grande quantité à proximité de la graine (Draycott et Christensen, 2003). Aucune différence significative n'a été relevée entre les deux niveaux de dose ($p = 0.51$) [1].

1. Les p-values pour la comparaison entre les deux doses d'azote ont été obtenues en relançant le modèle avec pour classe de référence la dose standard d'azote. Cela a permis d'obtenir des tests de Wald portant d'une part sur la différence entre une dose nulle d'azote et la dose

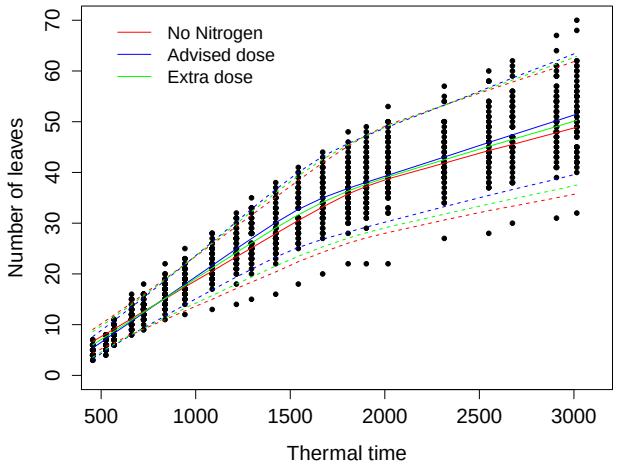


FIG. 3.4 – Prédictions du modèle en fonction de la dose d'azote reçue. La ligne continue représente la médiane et les lignes pointillées les quantiles d'ordre 5% et 95%.

Cet effet négatif a pu être contrebalancé par un rythme d'apparition des feuilles plus rapide chez les plantes fertilisées lors de la première phase de développement, même si l'utilisation d'une forte dose n'a pas permis d'augmenter significativement ce rythme ($p = 0.62$).

Le temps de rupture et donc le changement de phase se fait plus tôt chez les plantes ayant reçu un apport azoté ($p < 0.001$ pour les deux doses), mais, comme pour les autres paramètres, cette différence n'est pas significative entre les deux doses d'azote.

Aucune différence significative n'a été observée sur la différence de rythme entre les deux phases de développement. Le rythme d'apparition des feuilles lors de la première phase étant plus élevé pour les plantes fertilisées, cela signifie donc que le rythme d'apparition des feuilles reste plus élevé pour ces plantes au cours de la deuxième phase de développement. Cela signifie également que ce changement de rythme, qui se traduit par un ralentissement de la vitesse d'apparition des feuilles, et donc un allongement du phyllochrone, a le même effet quel que soit la dose d'azote reçue.

En utilisant les valeurs moyennes de chaque paramètre, le nombre de feuilles émises au moment du changement de phase est estimé à 37 pour les plantes sans azote, à 34 pour celles ayant reçu une dose normale et à 36 pour celles ayant reçu une forte dose. On remarque donc qu'en plus d'être indépendant des autres paramètres, le temps de rupture semble également associé à un nombre de feuilles relativement stable d'une dose à l'autre, et situé autour de 35–36 feuilles.

La dose normale d'azote était probablement suffisante pour permettre aux plantes de ne pas subir de stress azoté, puisqu'aucune différence significative n'a été relevée entre les deux différentes doses.

1.5 Discussion

Le modèle hiérarchique présenté ci-dessus nous a permis de prendre en compte à la fois les variabilités intra et inter-individuelles impliquées dans le phénomène d'organogenèse. Contrairement aux approches classiques basées sur des modèles à effets fixes, nous avons également pu tenir compte de la structure de corrélation des observations. La variabilité inter-individuelle étant significativement non nulle pour chaque

standard, et d'autre part sur la différence entre la dose élevée et la dose standard d'azote. Les résultats ne sont pas présentés dans le tableau pour des raisons de clarté.

paramètre du modèle, cela signifie qu'aucun de ces paramètres ne peut être traité comme un effet fixe, et qu'il est donc réellement nécessaire de prendre en compte sa variabilité dans la population comme nous l'avons fait. Il est alors possible d'estimer cette variabilité, et de comparer différentes populations en tenant compte de cette variabilité.

Nous avons montré que le temps de rupture était non corrélé aux autres paramètres du modèle, et que le changement de phase associé pouvait être relié à un nombre de feuilles relativement stable d'environ 35 feuilles. Lors de la comparaison des différentes doses, nous avons également montré qu'un même ralentissement s'opérait au cours de la deuxième phase, quelle que soit la dose d'azote reçue. Des résultats similaires ont été observés par Lemaire et al. (2009) sur la comparaison de trois densités de plantation. Ils ont montré que le rapport de rythme, c'est-à-dire le ratio entre les rythmes d'apparition des deux phases, étaient identiques pour chaque densité comparée, mais que le temps de rupture variait. Ces différents phénomènes peuvent s'expliquer par la forte compétition pour la lumière qui prévaut avant la couverture du sol par le feuillage. Avant cette date, les plantes cherchent à pousser plus vite et plus haut que leurs voisines et ont donc intérêt à avoir un phyllochrone qui soit le plus faible possible. Par contre, une fois ce point atteint, elles n'ont plus autant d'intérêt à développer plus de feuilles, et pourraient donc adopter un comportement identique, qui pourrait correspondre en quelque sorte au comportement « minimal » requis.

Le même comportement a été observé sur d'autres plantes, comme le navet, le colza, le chou frisé ou le rutabaga (Fletcher et al., 2012). Sibma (1977) a également montré qu'il était possible d'augmenter le rendement d'une culture en lui faisant atteindre le point de couverture plus tôt, à un moment où l'ensoleillement et la production potentielle sont à leurs maximums. Dürr et Boiffin (1995) ont également montré l'importance d'atteindre la couverture foliaire du sol au moment le plus approprié, c'est-à-dire lorsque la radiation est optimale.

Le premier intérêt de notre approche est de mieux comprendre ce phénomène clé de l'organogenèse, et sa variabilité au sein d'une population de plantes, et dans différentes conditions environnementales. Une bonne description de l'organogenèse est également primordiale pour l'application de modèles individus-centrés de type structure-fonction. En effet, une fois que les sources de variabilité du modèle ont été identifiées et quantifiées, il est possible d'étudier leurs propagations dans le système dynamique de croissance de la plante. Ces méthodes, dites d'analyse d'incertitudes, ont été étudiées dans le cadre des modèles de plantes par Monod et al. (2006) : étant donnée une densité de probabilité pour les facteurs d'entrée du modèle, on utilise des méthodes de type Monte Carlo pour évaluer l'incertitude dans les facteurs de sortie. Si les objectifs sont différents, les méthodes employées sont proches de celles utilisées lors d'une analyse de sensibilité (voir Chapitre 1, section 2.3). L'avantage de cette approche est qu'elle permet d'obtenir en sortie du modèle, non pas une valeur unique, mais une distribution de probabilité, ce qui peut notamment s'avérer utile en analyse de risque. Cependant il est nécessaire, comme pour une analyse de sensibilité, de définir au préalable la densité de probabilité des facteurs d'entrée. Une autre approche consiste alors à développer un modèle de croissance de plante à effets mixtes.

C'est ce que nous proposons dans la section suivante, avec une adaptation du modèle structure-fonction Greenlab. Dans le cadre de ce modèle, nous pourrons également utiliser comme paramètres d'organogenèse, ceux estimés individuellement à l'aide du modèle présenté ici.

2 Le modèle Greenlab de population

Dans cette section, nous proposons une extension du modèle individu-centré Greenlab, présenté au Chapitre 1, Section 1.1 dans le cas de la betterave (voir Jullien et al. (2011) pour l'adaptation du modèle Greenlab au colza), à l'échelle de la population. Pour cela, nous utilisons la structure hiérarchique des modèles mixtes : dans un premier temps nous caractérisons la variabilité intra-individuelle, c'est-à-dire la façon dont varient les observations d'une même plante, en fonction de paramètres spécifiques à cette plante, puis dans un second temps nous étudions la variabilité de ces paramètres individuels dans la population.

Si le modèle d'organogenèse présenté dans la section précédente avait été entièrement implémenté dans le logiciel Monolix, nous avons adopté la démarche inverse dans le cas du modèle Greenlab : au lieu d'implémenter le modèle sous Monolix, nous avons choisi d'implémenter les algorithmes MCMC-EM et SAEM dans la plateforme de modélisation de l'équipe, PyGMAlion.

Ce choix a été motivé par plusieurs raisons. Premièrement, le modèle Greenlab était déjà implémenté sous la plateforme, qui est spécifiquement dédiée à la gestion de modèles dynamiques. Les fichiers d'observations, d'environnement, et de paramètres associés au modèle Greenlab étaient également déjà au format requis par PyGMAlion. Deuxièmement, l'un de nos objectifs étant de comparer les deux algorithmes MCMC-EM et SAEM, il nous fallait choisir un environnement permettant d'implémenter les deux méthodes, afin d'assurer une comparaison la plus objective possible. Enfin, dans l'optique de rendre cette méthode accessible aux autres modèles développés dans l'équipe, il nous a paru indispensable de commencer par la mettre en œuvre sous PyGMAlion.

Le modèle d'organogenèse et le modèle Greenlab de population ont été codés séparément, principalement pour des raisons pratiques. En effet, les paramètres d'organogenèse déterminent le nombre d'observations et donc la taille des vecteurs individuels d'observation du modèle Greenlab de population : estimer ces paramètres avec l'algorithme EM impliquerait donc que pour chaque plante i observée (possédant un nombre fixe n_i de feuilles), on obtiendrait à chaque itération k de l'algorithme EM et à chaque étape de la procédure MCMC un vecteur de simulations de taille $n_i^{k,(m)}$, où $n_i^{k,(m)}$ n'a aucune raison d'être égal à n_i. On chercherait donc à comparer des vecteurs de tailles différentes, ce qui poserait des problèmes numériques. Une solution pourrait être d'augmenter artificiellement la taille du plus petit vecteur en ajoutant des zéros, mais cela peut poser des problèmes lorsque l'on utilise un modèle logarithmique, ou augmenter artificiellement l'écart entre simulations et observations. Une autre solution pourrait être de tronquer les vecteurs simulés dont la taille serait supérieure à celle du vecteur d'observation, mais cela ne règle pas le cas où le vecteur de simulations est plus petit.

Nous présentons dans cette section la formulation générale du modèle Greenlab à l'échelle de la population, en considérant deux modèles d'erreurs, additif ou log-additif. Nous proposons également une formulation du modèle permettant la prise en compte de bruits de modélisation au niveau du processus de production de biomasse. Les deux algorithmes MCMC-EM et SAEM sont ensuite comparés sur plusieurs jeux de données virtuels, puis sur des données réelles provenant de la betterave sucrière et du colza.

2.1 Formulation du modèle

Comme les observations dont nous disposons pour chaque plante sont issues de mesures destructives, nous ne disposons pas du suivi au cours du temps de la croissance de la plante. Cependant, en utilisant la structure récursive de formation des biomasses de chaque organe, il est possible d'accéder à l'historique du développement de la plante à partir des observations faites, à un temps donné fixé, sur l'ensemble des

organes de la plante. Plus spécifiquement, la biomasse de l'organe de rang n observé à un temps donné peut également être vue comme la biomasse de l'organe initié au temps t_n. Ceci nous permettra en particulier d'écrire le modèle comme un modèle de Markov caché lorsque l'on introduira des bruits de modélisation.

À un temps d'observation t_{obs} fixé, nous notons $y_i = (y_i(t_1), \dots, y_i(t_{n_i}))$ le vecteur d'observation des biomasses initiées aux temps $t_1, \dots, t_{n_i}$ pour la plante i, et $\mathcal{O}$ l'ensemble des organes de la plante. Le nombre d'observations par plante, n_i, correspond donc au nombre d'organes initiés, c'est-à-dire au nombre d'organes présents sur la plante au moment où elle a été observée.

Or nous avons vu dans le premier chapitre, lors de la description du modèle Greenlab, mais également dans la première partie de ce chapitre, que ce processus d'organogenèse dépendait lui-même de plusieurs paramètres. Dans le cas de la betterave, on observe deux phases de développement, auxquelles sont associés quatre paramètres d'organogenèse : le temps thermique d'initiation, le phyllochrone de la première phase, le temps thermique de rupture qui correspond à la mise en place de la deuxième phase de développement, et le phyllochrone de cette deuxième phase. Dans le cas du colza, une seule phase de développement est observée au cours du stade rosette, et seulement deux paramètres sont nécessaires : le temps d'initiation et le phyllochrone.

2.1.1 Variabilité intra-individuelle

On suppose ensuite que ces mesures correspondent à l'observation du système dynamique de croissance de la plante, que l'on suppose ici défini par le modèle Greenlab, assortie d'un bruit d'observation. Deux approches peuvent être utilisées pour prendre en compte ce bruit d'observation, par l'intermédiaire d'un modèle additif ou d'un modèle log-additif. L'utilisation d'un modèle additif peut cependant poser quelques soucis au niveau de la variance du bruit d'observation, car à la date à laquelle sont faites les mesures, certains organes sont encore en expansion, et n'ont pas encore atteint leur taille maximale. Ainsi, leurs masses pourront être largement inférieures à celles des organes matures, et l'utilisation d'un modèle additif peut conduire à une variance du bruit d'observation trop forte pour ces organes. Au contraire, le modèle multiplicatif tient compte naturellement de cette différence potentielle d'échelle. De plus, l'utilisation d'un modèle additif avec bruit Gaussien possède également l'inconvénient théorique de fournir des observations à support dans $\mathbb{R}$, alors que nous travaillons avec des quantités strictement positives.

Nous considérons dans la suite les deux modèles suivants, log-additif ($\mathcal{M}_1$) ou additif ($\mathcal{M}_2$) :

$$\begin{aligned} y_i(t_n) &= G_n(\phi_i) \circ e^{\varepsilon_{i,n}}, \\ \varepsilon_{i,n} &\sim \mathcal{N}_{d_n}(0, \Sigma_n) \end{aligned} \qquad (\mathcal{M}_1)$$

$$\begin{aligned} y_i(t_n) &= G_n(\phi_i) + \varepsilon_{i,n}, \\ \varepsilon_{i,n} &\sim \mathcal{N}_{d_n}(0, \Sigma_n) \end{aligned} \qquad (\mathcal{M}_2)$$

où $\circ$ représente le produit de Hadamard (multiplication terme à terme de deux matrices), d_n est le nombre d'organes de rang n, $\epsilon_{i,n}$ est un vecteur de taille d_n, $G_n(\phi_i)$ est le vecteur des biomasses théoriques des organes de rang n pour la plante i, et ϕ_i est un vecteur de paramètres spécifiques à la plante i. En reprenant les notations du Chapitre 1, les biomasses des organes s'écrivent en fonction de la séquence de biomasses produites depuis l'initiation de chaque organe jusqu'au temps t_{obs} auquel ont été faites les

observations :

$$G_n(\phi_i) = \left(\sum_{u=t_n}^{t_n^e \wedge t_{obs}} \frac{s_{o,n}(u, p_i^{al})}{d(u, p_i^{al})}\, q_i(u) \right)_{o \in \mathcal{O}}, \tag{3.8}$$

où nous rappelons que :

- t_{obs} est le temps auquel les observations sont faites,
- t_n est le jour auquel sont initiés les organes de rang n,
- t_n^e est le temps de fin d'expansion (en jours) des organes de rang n,
- $s_{o,n}(u, p_i^{al})$ est la fonction puits de l'organe de type o et de rang n de la plante i au temps u,
- $d(u, p_i^{al})$ est la demande totale de la plante i au temps u,
- $q_i(u)$ est la production de biomasse de la plante i au temps u.

Nous adoptons la convention $q_i(t_0) := q_0(i)$, où $q_0(i)$ est la masse de la graine de la plante i.

Betterave

Dans le cas de la betterave, trois types d'organes sont considérés : les limbes, les pétioles et la racine. L'ensemble des organes est donc $\mathcal{O} = \{l, p, r\}$. La racine étant initiée en même temps que la feuille de rang 1 (= les cotylédons), elle est incluse dans l'observation $y_{i,1}$, et n'apparaît plus dans les observations ultérieures. On a donc les modèles suivants :

$$y_i(t_n) = G_n(\phi_i) \circ e^{\varepsilon_{i,n}}, \tag{$\mathcal{M}_1'$}$$

$$\text{ou} \qquad y_i(t_n) = G_n(\phi_i) + \varepsilon_{i,n}, \tag{$\mathcal{M}_2'$}$$

avec

$$\varepsilon_{i,n} \sim \begin{cases} \mathcal{N}_3(0, \Sigma_{l,p,r}) & \text{si } n = 1, \\ \mathcal{N}_2(0, \Sigma_{l,p}) & \text{si } n > 1, \end{cases}$$

$$\Sigma_{l,p,r} = \begin{pmatrix} \Sigma_{l,p} & 0 \\ 0 & \sigma_r^2 \end{pmatrix},$$

$$\Sigma_{l,p} = \begin{pmatrix} \sigma_l^2 & \rho\sigma_l\sigma_p \\ \rho\sigma_l\sigma_p & \sigma_p^2 \end{pmatrix}.$$

Colza

Dans le cas du colza, au stade rosette seules les feuilles sont prises en compte dans le modèle, on a donc $\mathcal{O} = \{l\}$. Il n'y a donc qu'une seule fonction puits, celle correspond aux feuilles. Le modèle s'écrit alors simplement sous la forme suivante :

$$y_i(t_n) = G_n(\phi_i)\, e^{\varepsilon_{i,n}}, \tag{$\mathcal{M}_1''$}$$

$$\text{ou} \qquad y_i(t_n) = G_n(\phi_i) + \varepsilon_{i,n}, \tag{$\mathcal{M}_2''$}$$

$$\varepsilon_{i,n} \sim \mathcal{N}(0, \sigma_l^2).$$

2.1.2 Variabilité inter-individuelle

On note $\phi_i = (\phi_{i,1}, \ldots, \phi_{i,P})^t$ le vecteur contenant les paramètres du modèle spécifiques à la plante i. En supposant que la matrice de covariance Γ est diagonale, on a :

$$\begin{aligned} \phi_i &= \beta + \xi_i, \\ \xi_i &\sim \mathcal{N}_P(0, \Gamma). \end{aligned} \tag{3.9}$$

Pour s'assurer que les paramètres considérés ont bien pour support $\mathbb{R}$, on pourra considérer si nécessaire une transformation de la forme $\phi_{i,j} = h_j(\psi_{i,j}), j = 1, \ldots, P$, où les $\psi_i := (\psi_{i,1}, \ldots, \psi_{i,P})$ sont les paramètres individuels initiaux du modèle.

2.2 Estimation

L'ensemble des paramètres θ de notre modèle peut se décomposer en deux sous-vecteurs de paramètres θ_1 et θ_2, comme indiqué au Chapitre 2, Section 2.2, où θ_1 correspond aux paramètres associés aux effets aléatoires, et θ_2 aux paramètres de bruits. On a donc $\theta_1 = (\beta_1, \ldots, \beta_P, \sigma_1^2, \ldots, \sigma_P^2)$ et $\theta_2 = (\sigma_b^2, \sigma_p^2, \rho, \sigma_r^2)$ dans le cas de la betterave, et $\theta_2 = \sigma_l^2$ dans le cas du colza.

Dans la suite, et pour simplifier et unifier les notations, on pose $\tilde{y} = \log y$ et $\tilde{G} = \log G$ dans le cas du modèle $\mathcal{M}_1$, $\tilde{y} = y$ et $\tilde{G} = G$ dans le cas du modèle $\mathcal{M}_2$, et $N = \sum_{i=1}^{s} n_i$. En supposant que chaque effet aléatoire a une variance σ_i^2 non nulle, pour $i = 1, \ldots, P$, la vraisemblance complète du modèle Greenlab de population appartient bien à la famille exponentielle. De plus, comme précisé dans le chapitre précédent, le vecteur de statistiques exhaustives peut se décomposer en deux sous-vecteurs t_1 et t_2 correspondant à la décomposition de θ en θ_1, et θ_2.

Finalement, on obtient :

$$f(\tilde{y}, \phi; \theta) = h(\tilde{y}, \phi) \exp\left\{\langle s_1(\theta_1), t_1(\phi)\rangle - a_1(\theta_1)\right\} \exp\left\{\langle s_2(\theta_2), t_2(\tilde{y}, \phi)\rangle - a_2(\theta_2)\right\}, \tag{3.10}$$

avec :

$$s_1(\theta_1) = \begin{pmatrix} \Gamma^{-1}\beta \\ \Gamma^{-1} \end{pmatrix} \quad t_1(\phi) = \begin{pmatrix} \sum_{i=1}^{s} \phi_i \\ -\frac{1}{2}\sum_{i=1}^{s} \phi_i \phi_i^t \end{pmatrix}, \tag{3.11}$$

$$a_1(\theta_1) = \frac{sP}{2}\log 2\pi + \frac{s}{2}\log|\Gamma| + \frac{s}{2}\beta^t\Gamma^{-1}\beta, \tag{3.12}$$

$$h(\tilde{y}, \phi) = 1, \tag{3.13}$$

puis, dans le cas de la betterave :

$$s_2(\theta_2) = \begin{pmatrix} \Sigma^{-1} \\ \sigma_r^{-2} \end{pmatrix} \quad t_2(\tilde{y}, \phi) = \begin{pmatrix} -\frac{1}{2}\sum_{i=1}^{s}\sum_{n=1}^{n_i}\left(\tilde{y}_i(t_n) - \tilde{G}_n(\phi_i)\right)\left(\tilde{y}_i(t_n) - \tilde{G}_n(\phi_i)\right)^t \\ -\frac{1}{2}\sum_{i=1}^{s}(\tilde{y}_i^r(t_0) - \tilde{G}_0^r(\phi_i))^2 \end{pmatrix}, \tag{3.14}$$

$$a_2(\theta_2) = \frac{s+2N}{2}\log 2\pi + \frac{N}{2}\log|\Sigma_{l,p}| + \frac{s}{2}\log\sigma_r^2,$$

et dans le cas du colza :

$$s_2(\theta_2) = \sigma_l^{-2} \qquad t_2(\tilde{y}, \phi) = -\frac{1}{2} \sum_{i=1}^{s} \sum_{n=1}^{n_i} \left(\tilde{y}_i(t_n) - \tilde{G}_n(\phi_i) \right)^2, \tag{3.15}$$

$$a_2(\theta_2) = \frac{N}{2} \log 2\pi + \frac{N}{2} \log \sigma_l^2. \tag{3.16}$$

2.2.1 Étape E

Le modèle appartenant à la famille exponentielle, on a vu (Chapitre 2, équations 2.20, 2.32, 2.54) que l'étape d'espérance dans l'algorithme EM s'écrit simplement en fonction des statistiques exhaustives. En effet, il s'agit de calculer l'espérance conditionnelle du vecteur des statistiques exhaustives sous la loi des données manquantes ϕ sachant les observations y. Dans notre cas, l'étape E n'est pas explicite, car la loi $f(\phi \mid y; \theta)$ est inconnue, et nous remplaçons donc cette étape par une étape de simulation, dans le cas de l'algorithme MCMC-EM, ou d'approximation stochastique, dans le cas de 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 utiliserons les algorithmes présentés au chapitre précédent (sections 2.3.2 et 2.3.4).

Le tableau 3.4 présente la simplification du ratio $(f(u)q(v|u))/(f(v)q(u|v))$ intervenant dans le calcul de la probabilité d'acceptation, en fonction de l'algorithme (Metropolis-Hastings ou échantillonneur de Gibbs hybride), et de la loi instrumentale (marginale ou marche aléatoire) utilisés (voir Chapitre 2, section 2.3.2 pour le détail des calculs).

Tab. 3.4 – Ratio $(f(u)q(v|u))/(f(v)q(u|v))$ intervenant dans la probabilité d'acceptation en fonction de l'algorithme et de la loi instrumentale. On se place ici à l'itération k de l'algorithme MCMC-EM ou SAEM, et l'on simule une chaîne de Markov de taille m, de loi cible $f(\phi_i \mid y; \theta^k)$, où θ^k est l'estimation courante de θ.

	Metropolis-Hastings	Échantillonneur de Gibbs hybride
Loi marginale	$\dfrac{f(y_i \mid \tilde{\phi}_i; \theta^k)}{f(y_i \mid \phi_i^{k,(m)}; \theta^k)}$	$\dfrac{f(y_i \mid \tilde{\phi}_{i,j}; \theta^k)}{f(y_i \mid \phi_{i,j}^{k,(m)}; \theta^k)}$
Marche aléatoire	$\dfrac{f(y_i \mid \tilde{\phi}_i; \theta^k) f(\tilde{\phi}_i; \theta^k)}{f(y_i \mid \phi_i^{k,(m)}; \theta^k) f(\phi_i^{k,(m)}; \theta^k)}$	$\dfrac{f(y_i \mid \tilde{\phi}_{i,j}; \theta^k) f(\tilde{\phi}_{i,j}; \theta^k)}{f(y_i \mid \phi_{i,j}^{k,(m)}; \theta^k) f(\phi_{i,j}^{k,(m)}; \theta^k)}$

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,

$$t_1^{(k)} = \frac{1}{m_k} \sum_{m=1}^{m_k} t_1(\phi^{k,(m)}) \tag{3.17}$$

$$t_2^{(k)} = \frac{1}{m_k} \sum_{m=1}^{m_k} t_2(\tilde{y}, \phi^{k,(m)}) \tag{3.18}$$

– pour l'algorithme SAEM,

$$t_1^{(k)} = t_1^{(k-1)} + \gamma_k \left[\frac{1}{m_k} \sum_{m=1}^{m_k} t_1(\phi^{k,(m)}) - t_1^{(k-1)} \right] \tag{3.19}$$

$$t_2^{(k)} = t_2^{(k-1)} + \gamma_k \left[\frac{1}{m_k} \sum_{m=1}^{m_k} t_2(\tilde{y}, \phi^{k,(m)}) - t_2^{(k-1)} \right]. \tag{3.20}$$

2.2.2 Étape M

Lorsque tous les éléments du vecteur ϕ_i sont considérés comme aléatoires (à chaque effet fixe est associé un effet aléatoire), 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 :

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

Maximisation par rapport à θ_1

On obtient les équations suivantes pour $\theta_1 = (\beta, \Gamma)$ à l'itération k de l'algorithme MCMC-EM ou SAEM :

$$\hat{\beta}_j = \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}_{\theta^k}(\phi_{i,j} \mid \tilde{y}_i), \quad j = 1, \ldots, P \tag{3.21}$$

$$\hat{\sigma}_j^2 = \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}_{\theta^k}(\phi_{i,j}^2 \mid \tilde{y}_i) - \hat{\beta}_j^2, \quad j = 1, \ldots, P. \tag{3.22}$$

Maximisation par rapport à θ_2

Le vecteur de paramètres θ_2 correspond aux bruits d'observation, et il convient donc de distinguer le cas de la betterave et du colza. On a :

Betterave

$$\hat{\Sigma}_{l,p} = \frac{1}{N} \sum_{i=1}^{s} \sum_{n=1}^{n_i} \mathbb{E}_{\theta^k} \left[\left(\tilde{y}_{i,n} - \tilde{G}_n(\phi_i) \right) \left(\tilde{y}_{i,n} - \tilde{G}_n(\phi_i) \right)^t \mid \tilde{y}_i \right], \tag{3.23}$$

$$\hat{\sigma}_r^2 = \frac{1}{s} \sum_{i=1}^{s} \mathbb{E}_{\theta^k} \left[(\tilde{y}_{i,0}^r - \tilde{G}_0^r(\phi_i))^2 \mid \tilde{y}_i \right], \tag{3.24}$$

où l'on ne considère que les biomasses des limbes et des pétioles pour $y_{i,0}$ dans l'équation 3.23.

Colza

$$\hat{\sigma}_l^2 = \frac{1}{N} \sum_{i=1}^{s} \sum_{n=1}^{n_i} \mathbb{E}_{\theta^k} \left[(\tilde{y}_{i,n} - \tilde{G}_n(\phi_i))^2 \mid \tilde{y}_i \right]. \tag{3.25}$$

2.2.3 Convergence de l'algorithme

Dans le cas du modèle exponentiel, la convergence de l'algorithme est assurée lorsque les hypothèses décrites au chapitre 2, section 2.2.3, sont vérifiées. Nous montrons dans cette section comment ces hypothèses s'appliquent dans le cas du modèle Greenlab pour la betterave, les résultats s'appliquant de façon similaire dans le cas du colza.

Le vecteur de paramètres θ peut s'écrire $\theta = (\beta, \sigma^2, \sigma_b^2, \sigma_p^2, \rho, \sigma_r^2)$, où σ^2 est un vecteur de taille P contenant les variances des effets aléatoires (c'est-à-dire les éléments diagonaux de Γ). Les composantes de θ appartiennent alors aux sous-espaces suivants :

- $\beta \in \mathbb{R}^P$ pour $j = 1, \ldots, P$, (si le paramètre initial n'a pas pour support $\mathbb{R}$, il est toujours possible d'utiliser une transformation, par exemple de type logarithmique, pour que cette relation reste valable),
- $\sigma^2 \in (\mathbb{R}_+^*)^P$, σ_b^2, σ_p^2, $\sigma_r^2 \in \mathbb{R}_+^*$,
- $\rho \in]-1, 1[$, ce qui correspond à l'hypothèse selon laquelle la matrice $\Sigma_{b,p}$ est non singulière, sachant que σ_b^2, $\sigma_p^2 \in \mathbb{R}_+^*$.

D'où $\Theta = \mathbb{R}^P \times (\mathbb{R}_+^*)^{P+3} \times (-1, 1)$, qui est bien un sous-ensemble ouvert de $\mathbb{R}^{2P+4}$: l'hypothèse (M1) est donc bien vérifiée.

L'hypothèse (M2)(a) est trivialement vérifiée, et l'hypothèse (M2)(c) est également vérifiée en utilisant les expressions explicites obtenues à l'étape de maximisation (voir équations 3.21 et 3.23-3.25). En effet, la fonction $\theta \mapsto \ell_c(t; \theta)$ étant strictement concave, la fonction $\hat{\theta}$ est bien définie, et elle est également continue grâce au théorème des fonctions implicites.

L'hypothèse (M2)(b) se déduit facilement des propriétés de la famille exponentielle (Sundberg, 1974). En utilisant la paramétrisation canonique ou naturelle $\eta = s(\theta)$, la densité conditionnelle de ϕ sachant $\tilde{y}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^l$ peut s'écrire :

$$f(\phi|\tilde{y}; \eta) = \exp\{\langle \eta, t(\tilde{y}, \phi)\rangle - b_y(\eta)\}, \tag{3.26}$$

où

$$e^{b_y(\eta)} = \int e^{\langle \eta, t(\tilde{y}, \phi)\rangle} \, \lambda(d\phi). \tag{3.27}$$

De même que la fonction b définie en (2.13), b_y est le logarithme de la transformée de Laplace associée à la mesure image de λ par l'application t, et est donc indéfiniment dérivable sur Ω. Comme on a :

$$\nabla b_y(\eta) = \mathbb{E}_\eta(t(\tilde{y}, \phi) \mid \tilde{y}) = \bar{t}(\eta), \tag{3.28}$$

l'application $\bar{t}$ est continue et finie sur Ω. Puis, comme la transformation inverse $s^{-1}(\eta)$ est continue, $\bar{t}$ est également continue et finie sur Θ et l'hypothèse (M2)(b) est donc vérifiée.

La condition (M)(d) est également vérifiée. En effet, d'après l'équation (2.12) où $h = 1$ (voir aussi (3.10) and (3.26)), on a $L(\eta) = \exp\{b_y(\eta) - b(\eta)\}$, où $b_y(\eta)$ et $b(\eta)$ sont toutes les deux analytiques, ce qui entraîne que la fonction L est positive, finie et continue.

Une autre démonstration de la continuité de la fonction L ne reposant pas sur l'utilisation de transformées de Laplace, et qui peut s'appliquer de façon plus générale aux cas où la densité complète n'appartient pas à la famille exponentielle, peut également être donnée. Cette deuxième preuve utilise une représentation différente de la vraisemblance, particulièrement adaptée aux modèles à effets aléatoires et qui permet également d'illustrer le fait que la condition (M2)(e*) proposée ici est plus appropriée que la condition initiale (M2)(e) qui n'est généralement pas satisfaite par les modèles à effets aléatoires.

Montrons tout d'abord que sous une condition très générale portant sur les paramètres, la vraisemblance des observations $L(\theta)$ est bornée (ce qui est un des pré-requis nécessaires à l'existence du maximum de vraisemblance). Nous définissons les quantités suivantes :

$$\Sigma(\beta) := \left(\sum_{i=1}^{s} n_i\right)^{-1} \sum_{i=1}^{s}\sum_{n=1}^{n_i} \left(\tilde{y}_{i,n} - \tilde{G}_n(\beta)\right)\left(\tilde{y}_{i,n} - \tilde{G}_n(\beta)\right)^t, \tag{3.29}$$

$$\sigma_r^2(\beta) := \frac{1}{s}\sum_{i=1}^{s}(\tilde{y}_{i,1}^r - \tilde{G}_1^r(\beta))^2. \tag{3.30}$$

Proposition 3.1. *La vraisemblance $L(\theta)$ est continue. Si de plus, on suppose* $\inf_{\beta\in\mathbb{R}^P}(\det\Sigma(\beta)) > 0$ *et* $\inf_{\beta\in\mathbb{R}^P}(\sigma_r^2(\beta)) > 0$, *alors $L(\theta)$ est également bornée.*

Démonstration. La vraisemblance des observations peut s'écrire sous la forme suivante :

$$L(\theta) = \int_{\mathbb{R}^{P\times s}} f(\tilde{y} \mid \phi;\theta_2) f(\phi;\theta_1) d\phi = \mathbb{E}_{\theta_1}\left[f(\tilde{y} \mid \Phi;\theta_2)\right], \tag{3.31}$$

où $\theta_1 = (\beta, \sigma^2)$, $\theta_2 = (\sigma_b^2, \sigma_p^2, \sigma_r^2, \rho)$ et $\Phi \sim f(\phi;\theta_1)$.

Montrons d'abord la continuité de la fonction L. Soit $\theta_n = (\theta_{1,n}, \theta_{2,n})$ une suite de $\mathbb{R}^{2P+4}$ convergeant vers $\theta = (\theta_1, \theta_2)$, et soient $h_n(\phi) := f(\tilde{y} \mid \phi;\theta_{2,n})$, $h(\phi) := f(\tilde{y} \mid \phi;\theta_2)$, $\Phi_n \sim f(\phi;\theta_{1,n})$ et $\Phi \sim f(\phi;\theta_1)$. On a :

$$L(\theta_n) = \mathbb{E}_{\theta_{1,n}}\left[f(\tilde{y} \mid \Phi;\theta_{2,n})\right] = \mathbb{E}\left[f(\tilde{y} \mid \Phi_n;\theta_{2,n})\right] = \mathbb{E}\left[h_n(\Phi_n)\right]. \tag{3.32}$$

Les conditions du théorème de Mann-Wald généralisé (ou « generalized continuous mapping theorem ») sont alors vérifiées :

i) h_n et h sont des fonctions mesurables à valeurs dans $(\mathbb{R}, |\cdot|)$ qui est un espace métrique séparable,

ii) par continuité de $f(\tilde{y} \mid \phi;\theta_2)$ en (ϕ, θ_2), pour tout ϕ et pour toute suite $\{\phi_n\}$ convergeant vers ϕ, on a $h_n(\phi_n) = f(\tilde{y} \mid \phi_n;\theta_{2,n}) \to f(\tilde{y} \mid \phi;\theta_2) = h(\phi)$,

iii) par continuité de $f(\phi;\theta_1)$ en θ_1, on a $f(\phi;\theta_{1,n}) \to f(\phi;\theta_1)$, et donc par le lemme de Scheffé, $\Phi_n \xrightarrow{loi} \Phi \sim f(\phi;\theta_1)$,

et l'on a donc $h_n(\Phi_n) \xrightarrow{loi} h(\Phi)$.

La convergence en loi d'une suite de variables aléatoires (X_n) vers une variable aléatoire X n'implique pas en général la convergence des moments (dont l'existence même n'est pas assurée). Cependant, si l'on suppose que la suite de variables aléatoires est également uniformément intégrable, on peut toutefois en déduire certains résultats sur la convergence des moments (Billingsley, 2012). En particulier, si $|X_n|^k$ est uniformément intégrable pour un certain $k \geq 1$, alors $\mathbb{E}(X_n^r) \to \mathbb{E}(X^r)$ pour $1 \leq r \leq k$. Une condition suffisante pour assurer l'uniforme intégrabilité d'une suite de variables aléatoires est que la suite soit uniformément bornée, c'est-à-dire qu'il existe une constante C telle que $|X_n| < C$ quel que soit n. Or dans notre cas, on a :

$$h_n(\phi) \leq \frac{(2\pi)^{-\sum_{i=1}^s n_i - s/2}}{(\det\Sigma_{b,p;n})^{\sum_{i=1}^s n_i/2}\,(\sigma_{r;n}^2)^{s/2}}, \qquad n \geq 1. \tag{3.33}$$

Comme $\theta_{2,n} \to \theta_2$, la suite des majorants de h_n converge également, et est donc par conséquent bornée par une certaine constante $C > 0$. On en déduit que $\{h_n(\Phi_n)\}$ est uniformément bornée par C, et on a donc bien $L(\theta_n) = \mathbb{E}\left[h_n(\Phi_n)\right] \to \mathbb{E}\left[h(\Phi)\right] = L(\theta)$, ce qui montre la continuité de la fonction L.

Montrons maintenant que la vraisemblance L est bornée. Considérons le modèle non linéaire mais sans effets aléatoires suivant :

$$\tilde{y}_{i,n} = \tilde{G}_n(\beta) + \varepsilon_{i,n}, \qquad \varepsilon_{i,n} \sim \mathcal{N}_{d_n}(0, \Sigma_n). \tag{3.34}$$

Notons $L(\mu)$ la vraisemblance associée à ce modèle, où $\mu := (\beta, \theta_2)$. On a :

$$\begin{aligned} L(\mu) &= \frac{(2\pi)^{-\sum_{i=1}^s n_i - s/2}}{(\det \Sigma_{b,p})^{\sum_{i=1}^s n_i/2} (\sigma_r^2)^{s/2}} \times \exp\left\{ -\frac{1}{2} \sum_{i,n} \left(\tilde{y}_{i,n} - \tilde{G}_n(\beta)\right)^t \Sigma_{b,p}^{-1} \left(\tilde{y}_{i,n} - \tilde{G}_n(\beta)\right) \right\} \\ &\quad \times \exp\left\{ -\frac{1}{2\sigma_r^2} \sum_{i=1}^s (\tilde{y}_{i,1}^r - \tilde{G}_1^r(\beta))^2 \right\}. \end{aligned} \tag{3.35}$$

En général le modèle correspondant n'appartient pas à la famille exponentielle, à cause de la dépendance non linéaire de G_n en β. Cependant la fonction L est tout de même continue en μ, par continuité de $\tilde{G}_n(\beta)$ en β et de L en θ_2. De plus, si l'on fixe β, $L(\mu)$ possède un unique maximum $\theta_2(\beta) := (\sigma_b^2(\beta), \sigma_p^2(\beta), \rho(\beta), \sigma_r^2(\beta))$, où $\Sigma(\beta)$ et $\sigma_r^2(\beta)$ sont données par (3.29) et (3.30) respectivement. En particulier, on obtient facilement qu'il existe une constante $c > 0$ telle que :

$$L(\mu) \leq \sup_{\theta_2} L(\beta, \theta_2) = \frac{c}{(\det \Sigma(\beta))^{\sum_{i=1}^s n_i/2} (\sigma_r^2(\beta))^{s/2}}. \tag{3.36}$$

On en déduit que :

$$\sup_{\mu} L(\mu) = \sup_{\beta, \theta_2} L(\beta, \theta_2) \leq \frac{c}{\inf_\beta \left[\det \Sigma(\beta)\right]^{\sum_{i=1}^s n_i/2} \inf_\beta \left[\sigma_r^2(\beta)\right]^{s/2}} =: M < \infty, \tag{3.37}$$

puisque le dénominateur est strictement positif par hypothèse, et $L(\mu)$ est donc bornée.

Il reste maintenant à étendre ce résultat à $L(\theta)$. En effet on peut montrer facilement, avec des arguments similaires à ceux utilisés pour montrer la continuité de $L(\theta)$, que lorsque des effets aléatoires sont ajoutés dans le modèle initial de vraisemblance $L(\mu)$, la vraisemblance du modèle ainsi étendu est également continue. On remarque également que lorsque la variance d'un effet aléatoire $\Phi_j \sim \mathcal{N}(\beta_j, \sigma_j^2)$ converge vers 0, c'est-à-dire vers la frontière de son domaine de définition, alors $L(\theta_{-j}, \beta_j, \sigma_j^2)$ converge vers $L(\theta_{-j}, \beta_j) := \mathbb{E}_{\theta_{1,-j}}\left[f(\tilde{y} \mid \Phi_{-j}; \theta_2, \beta_j)\right]$, où $\theta_{-j} = \theta \setminus (\beta_j, \sigma_j^2)$, $\theta_{1,-j} = \theta_1 \setminus (\beta_j, \sigma_j^2)$ et $\Phi_{-j} = \Phi \setminus \Phi_j$, où $\Phi_j = (\Phi_{1j}, \ldots, \Phi_{sj})$. La preuve de ce résultat est donnée par Chen et al. (2013a) (Proposition 1). On en déduit que la vraisemblance du modèle étendu est bornée lorsque la variance de l'effet aléatoire correspondant s'approche de la frontière de l'espace, et comme c'est également le cas sur le reste de son domaine de définition, la fonction $L(\theta)$ est donc bornée. □

Pour montrer que l'ensemble $\{\theta \in \Theta, L(\theta) \geq M\}$ est compact, la fonction L étant continue et bornée, il suffit de montrer que lorsque θ tend vers la frontière de son domaine de définition, alors $L(\theta) \to 0$. Dans notre cas, ceci est vrai pour toutes les composantes de θ, sauf pour les variances des effets aléatoires. En effet, lorsque la variance d'un effet aléatoire tend vers zero, le modèle ainsi obtenu correspond à celui où ce paramètre est traité comme un effet fixe et n'appartient plus à la famille exponentielle. L'ensemble $\{\theta \in \Theta, L(\theta) \geq M\}$ ne peut donc pas être compact. Cependant, l'hypothèse (M2)(e*) peut tout de même

être vérifiée. Par exemple, dans le cas où tous les paramètres, exceptés les variances des effets aléatoires, tendent vers la frontière de leurs domaines de variation, on a $L(\theta) \to 0$ et $\sup_j L(\theta_{-j}, \beta_j) < \sup_\theta L(\theta)$.

Dans la pratique cependant, si les variances de certains paramètres tendent vers 0, les paramètres correspondant seront traités comme des effets fixes et incorporés à θ_2. Plusieurs cas peuvent alors se produire :

- si l'on peut obtenir une maximisation explicite des effets fixes associés, on procède comme précédemment
- sinon, on peut chercher si une maximisation explicite de ces paramètres peut être obtenue conditionnellement aux autres paramètres, auquel cas une généralisation de l'algorithme EM de type ECM pourra être utilisée
- si aucune des situations ci-dessus n'est possible, on peut avoir recours à des alternatives de type quasi-Newton pour maximiser les paramètres (voir par exemple Trevezas et al. (2012) ; Trevezas et Cournède (2013) pour une application de ces méthodes dans le cas du modèle Greenlab).

Remarque 3.1. Il est également possible, comme proposé par Racine-Poon (1985) ou Kuhn et Lavielle (2005), de considérer une formulation Bayésienne pour les effets fixes, c'est-à-dire de supposer une loi a priori pour ces paramètres, et de prendre par exemple la moyenne ou le mode a posteriori comme estimateur de β. L'utilisation d'une formulation Bayésienne permet en effet de rester dans la famille exponentielle, mais d'un autre côté nous quittons le cadre de l'estimation par maximum de vraisemblance. De plus, le choix de la distribution a priori doit également être fait avec soin.

La dernière hypothèse (M3) est vérifiée dans notre cas, $\mathcal{L}$ étant ici clairement compact.

2.2.4 Intervalles de confiance

À l'aide de la matrice d'information de Fisher

On rappelle que la matrice d'information de Fisher s'écrit $I(\theta; \tilde{y}) = \mathcal{I}_c(\theta; \tilde{y}) - \mathcal{I}_m(\theta; \tilde{y})$. En utilisant l'écriture sous forme de modèle exponentiel, et la décomposition du vecteur $t(x)$ en deux sous-vecteurs indépendants, on obtient pour la première matrice $\mathcal{I}_c(\theta; \tilde{y})$ une décomposition en deux blocs diagonaux, l'un correspondant à θ_1 et l'autre à θ_2 :

$$\mathcal{I}_c(\theta) = \begin{pmatrix} \mathcal{I}_c(\theta_1) & 0 \\ 0 & \mathcal{I}_c(\theta_2) \end{pmatrix} \tag{3.38}$$

où la matrice $\mathcal{I}_c(\theta_1)$ est de taille $2P \times 2P$, et est diagonale par blocs, avec chaque bloc défini de la façon suivante :

$$\mathcal{I}_c(\theta_{1,j}) = \begin{pmatrix} \dfrac{s}{\sigma_j^2} & 0 \\ 0 & \dfrac{s}{2\sigma_j^2} \end{pmatrix}, \quad j = 1, \ldots, P \tag{3.39}$$

et où la matrice $\mathcal{I}_c(\theta_2)$ dépend de la plante considérée :

- pour la betterave, $\mathcal{I}_c(\theta_2) = \text{diag}\left\{\mathcal{I}_c(\theta_{2,1}), \dfrac{s}{2\sigma_r^2}\right\}$, avec :

$$\mathcal{I}_c(\theta_{2,1}) = \frac{N}{1-\rho^2}\begin{pmatrix} \frac{(2-\rho^2)}{4\sigma_b^4} & -\frac{\rho^2}{4\sigma_b^2\sigma_p^2} & -\frac{\rho}{2\sigma_b^2} \\ -\frac{\rho^2}{4\sigma_b^2\sigma_p^2} & \frac{(2-\rho^2)}{4\sigma_p^4} & -\frac{\rho}{2\sigma_p^2} \\ -\frac{\rho}{2\sigma_b^2} & -\frac{\rho}{2\sigma_p^2} & \frac{1+\rho^2}{1-\rho^2} \end{pmatrix} \tag{3.40}$$

– pour le colza : $\dfrac{N}{2\sigma_l^2}$.

La matrice $\mathcal{I}_c(\theta)$ peut être estimée par la matrice $\mathcal{I}_c(\hat{\theta})$, qui peut être calculée lorsque l'algorithme a convergé et que l'on dispose de l'estimateur du maximum de vraisemblance de θ.

En revanche, le calcul de la matrice $\mathcal{I}_m(\theta, y)$ ne peut pas se faire de façon explicite (voir équation (2.29)), et l'on propose alors d'estimer cette matrice à l'aide de la covariance empirique des statistiques exhaustives générées par l'algorithme. En effet,

– dans le cas de l'algorithme MCMC-EM, on a :

$$\hat{\mathcal{I}}_m^{(k)}(\eta;\tilde{y}) = \frac{1}{m_k}\sum_{m=1}^{m_k} t_m^{(k)} t_m^{(k)\prime} - \left(\frac{1}{m_k}\sum_{m=1}^{m_k} t_m^{(k)}\right)\left(\frac{1}{m_k}\sum_{m=1}^{m_k} t_m^{(k)}\right)', \tag{3.41}$$

le calcul étant fait lorsque l'algorithme a convergé, en utilisant la chaîne générée à la dernière itération.

– dans le cas de l'algorithme SAEM, on a :

$$\hat{\mathcal{I}}_m^{(k)}(\eta;\tilde{y}) = \hat{\mathcal{I}}_m^{(k-1)}(\eta;\tilde{y}) + \gamma_k\left[\frac{1}{m_k}\sum_{m=1}^{m_k} t_m^{(k)} t_m^{(k)\prime} - \left(\frac{1}{m_k}\sum_{m=1}^{m_k} t_m^{(k)}\right)\left(\frac{1}{m_k}\sum_{m=1}^{m_k} t_m^{(k)}\right)' - \hat{\mathcal{I}}_m^{(k-1)}(\eta;\tilde{y})\right], \tag{3.42}$$

où $t_m^{(k)} := t^{(k)}(\tilde{y}, \phi^{k,(m)})$.

On en déduit la matrice $\hat{\mathcal{I}}_m(\theta, y)$ à l'aide de la méthode Delta (voir aussi équation (2.19)) :

$$\hat{\mathcal{I}}_m(\theta,\tilde{y}) = J_s(\theta)^t\, \hat{\mathcal{I}}_m(s(\theta),\tilde{y})\, J_s(\theta). \tag{3.43}$$

Par bootstrap paramétrique

Les intervalles de confiance ainsi obtenus peuvent être comparés aux intervalles de confiance obtenus par bootstrap paramétrique. À partir d'un estimateur $\hat{\theta}$ du vecteur de paramètre obtenu par l'un des deux algorithmes MCMC-EM ou SAEM, on génère aléatoirement B échantillons boostrap de taille s, $\mathbf{y}_{b,i}^* = (y_{b,i}^*(t_1), \ldots, y_{b,i}^*(t_{n_i}))_{i=1,\ldots,s}$, $b = 1, \ldots, B$. Puis, pour chacun de ces échantillons bootstrap, on relance la procédure d'estimation. On obtient ainsi B estimateurs de $\hat{\theta}$, $\hat{\theta}^1, \ldots, \hat{\theta}^B$, à partir desquels on peut estimer la loi de l'estimateur $\hat{\theta}$. En particulier, la moyenne et l'écart-type de chaque composante de l'estimateur $\hat{\theta}$ s'obtiennent de la façon suivante :

$$\hat{\theta}_j^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_j^b$$

$$\hat{\mathrm{Var}}(\theta_j^*) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_j^b - \hat{\theta}_j^*).$$

Un intervalle de confiance de niveau $1-\alpha$ du bootstrap s'obtient grâce aux quantiles empiriques de la distribution bootstrap :

$$IC^*_{1-\alpha}(\theta_j) = \left[\hat{\theta}^*_j - (\hat{\theta}^b_j)_{(\frac{\alpha}{2}B)}; \hat{\theta}^*_j + (\hat{\theta}^b_j)_{((1-\frac{\alpha}{2})B)}\right], \tag{3.44}$$

où $(\hat{\theta}^b_j)_{(\frac{\alpha}{2}B)}$ est la statistique d'ordre $\frac{\alpha}{2}B$ des estimateurs bootstrap. Lorsque la taille de l'échantillon bootstrap B est suffisamment grande, l'intervalle de confiance ainsi construit converge vers l'intervalle de confiance asymptotique théorique. Dans la pratique, B est choisi suffisamment grand pour pouvoir obtenir une bonne approximation de la distribution de l'estimateur $\hat{\theta}$, en général $B = 1000$.

2.3 Simulations

Nous présentons dans cette section une première comparaison des algorithmes sur données simulées. Pour des raisons de clarté, nous présentons les résultats obtenus dans le cas de la betterave, puisque le modèle utilisé est plus complexe que pour le colza. Plusieurs jeux de données virtuelles ont été générés, en considérant un nombre croissant d'effets aléatoires, puis en introduisant la variabilité provenant du processus d'organogenèse (voir Section 1 de ce chapitre). Nous présentons seulement ici les résultats du modèle log-additif (modèle $\mathcal{M}_1$). Le modèle additif sera comparé au modèle log-additif sur les données réelles (voir section 2.4.2).

Les effets aléatoires ont été introduits dans le modèle en suivant l'ordre fourni par l'analyse de sensibilité du modèle Greenlab, dont on trouvera les résultats au Chapitre 1. Un premier test des méthodes a été réalisé en prenant en compte deux effets aléatoires : l'efficience de conversion μ et le coefficient s^{pr}, et en supposant dans un premier temps que les paramètres d'organogenèse sont constants d'une plante à l'autre (chaque plante possède alors le même nombre de feuilles, toutes les plantes étant initiées en même temps). Puis, un jeu de données virtuelles plus réaliste a été généré, en considérant cette fois les trois paramètres aléatoires suivants : μ, s^{pr} et a_r, et en introduisant également une variabilité dans les paramètres d'organogenèse. Le paramètre a_r correspond à l'un des paramètres d'allocation de biomasse aux racines.

Ces paramètres étant positifs par définition, l'hypothèse d'une distribution normale faite lors de l'étape de variabilité inter-individuelle n'est pas réaliste. De plus, supposer une distribution de support $\mathbb{R}$ peut conduire à la simulation d'une valeur négative pour ces paramètres lors de l'algorithme MCMC, ce qui peut conduire à des problèmes numériques lorsque de telles valeurs sont ensuite utilisées pour produire des simulations du modèle Greenlab. Une transformation de type logarithmique a donc été appliquée pour ces paramètres, c'est-à-dire que l'on a supposé $\log \phi_i = \beta + \xi_i$ pour $i = 1, \ldots, s$, avec $\xi_i \sim \mathcal{N}(0, \Gamma)$, ce qui revient à supposer que les paramètres suivent des loi log-normales : $\phi_i \sim \mathcal{LN}(\beta, \Gamma)$. Nous noterons dans la suite $\beta_\mu := \mathbb{E}(\log \mu)$, $\beta_{s^{pr}} := \mathbb{E}(\log s^{pr})$, $\beta_{a_r} := \mathbb{E}(\log a_r)$, $\sigma_\mu = \sqrt{\mathrm{Var}(\log \mu)}$, $\sigma_{s^{pr}} = \sqrt{\mathrm{Var}(\log s^{pr})}$ et $\sigma_{a_r} = \sqrt{\mathrm{Var}(\log a_r)}$. Les valeurs de ces paramètres sont présentées dans le tableau 3.5.

Les paramètres d'organogenèse ont quant à eux été simulés à partir de lois normales, en suivant le modèle décrit dans la section 1.

Cinquante jeux de paramètres individuels ont ainsi été générés, puis les données virtuelles des cinquante plantes correspondantes ont été obtenues à l'aide du modèle Greenlab en ajoutant des bruits d'observations, eux-mêmes simulés à partir de lois normales centrées et dont les variances sont définies dans le tableau 3.5.

Au cours de l'étape de simulation, on obtient pour chaque plante $i, i = 1, \ldots, s$, une réalisation du vecteur aléatoire ϕ_i. En considérant ces réalisations comme des observations, on peut alors calculer l'estimateur du maximum de vraisemblance associé aux données entièrement observées, c'est-à-dire constituées

Tab. 3.5 – Valeurs des paramètres utilisés pour simuler les jeux de données dans le cas où l'on considère trois effets aléatoires.

Paramètre	Vraie valeur
β_μ	1. 7
σ_μ	0.15
$\beta_{s^{pr}}$	-3
$\sigma_{s^{pr}}$	0.5
β_{a_r}	1.45
σ_{a_r}	0.15

Paramètre	Vraie valeur
σ_b^2	0.15
σ_p^2	0.15
ρ	0.67
σ_r	0.15

à la fois des effets aléatoires simulés ϕ_i et des observations simulées y_i. Dans ce cas, la vraisemblance du modèle est la densité jointe des (y, ϕ) et s'écrit donc comme un modèle Gaussien. La comparaison des estimations obtenues avec les vraies valeurs utilisées pour générer les données est donnée dans le tableau 3.6. Les paramètres liés aux effets aléatoires sont plutôt bien estimés, mais les bruits d'observation associés aux limbes et aux pétioles ont été moins bien simulés. Plusieurs jeux de données ont été générés mais ont conduit à des valeurs similaires. Ce problème n'apparaît pas dans le cas où le bruit d'observation simulé est unidimensionnel, ce qui suggère un problème dans la procédure de génération de vecteurs aléatoires qui a été utilisée. D'autres tests devront être réalisés pour tenter d'identifier la source de ce biais.

Tab. 3.6 – Estimation des paramètres sur les données simulées complètes (observations + paramètres individuels).

Paramètre	**Vraie valeur**	**Estimation sur données complètes**
β_μ	1.7	1.7036
σ_μ	0.15	0.1544
$\beta_{s^{pr}}$	-3	-2.9062
$\sigma_{s^{pr}}$	0.5	0.5729
β_{a_r}	1.45	1.4906
σ_{a_r}	0.15	0.1232
σ_b^2	0.15	0.2019
σ_p^2	0.15	0.0827
ρ	0.67	0.5342
σ_r^2	0.15	0.1389

2.3.1 Algorithme MCMC-EM

Nous présentons dans cette section les résultats obtenus avec l'algorithme MCMC-EM. Les performances de l'algorithme de Metropolis-Hastings (MH) et de l'échantillonneur hybride de Gibbs (hGs) ont été comparées, en utilisant comme loi instrumentale, soit la loi marginale des données manquantes, soit une marche aléatoire Gaussienne. Pour cette dernière, les deux schémas adaptatifs présentés en section 2.3.2 ont été comparés. Dix réalisations indépendantes de chaque configuration ont été lancées, afin d'estimer l'erreur de Monte Carlo, et nous avons utilisé le programme OpenMP afin de paralléliser la partie du code correspondant à la génération des chaînes de Markov pour chaque plante, qui peuvent en effet être simulées indépendamment les unes des autres. Le programme a été lancé sur 6 cœurs du mésocentre de calcul de l'École Centrale [2].

2. `http ://www.mesocentre.ecp.fr/`

Tab. 3.7 – Résultats obtenus avec l'algorithme de Metropolis-Hastings à marche aléatoire adaptative proposé par Andrieu et Thoms (2008), avec une augmentation quadratique de la taille de la chaîne et une moyennisation des estimations à partir de l'itération 60.

	Vraie valeur	**Obs. complètes**	**Estimation**	**ET (FIM)**	**ET (10 réal.)**
β_μ	1.7	1.7036	1.6981	0.0210	0.00035
σ_μ	0.15	0.1544	0.1484	0.0148	0.00038
$\beta_{s^{pr}}$	-3	-2.9062	-2.9159	0.0820	0.00284
$\sigma_{s^{pr}}$	0.5	0.5729	0.5787	0.0586	0.00299
β_{a_r}	1.5	1.4906	1.4992	0.0163	0.00134
σ_{a_r}	0.15	0.1232	0.1154	0.0115	0.00099
σ_b^2	0.15	0.2019	0.2035	0.0055	0.00010
σ_p^2	0.15	0.0827	0.0844	0.0023	0.00002
ρ	0.67	0.5342	0.5292	0.0139	0.00019
σ_r^2	0.15	0.1389	0.1276	0.0255	0.00304
Nb itérations	100	Taille finale	10451	Temps d'éxécution	7h21

Dans un premier temps nous avons utilisé la méthode proposée par Fort et Moulines (2003) qui consiste à augmenter de façon déterministe la taille de la chaîne à chaque itération de l'algorithme, puis à utiliser une moyennisation des estimations à partir d'un certain rang. En suivant les recommandations des auteurs, nous avons choisi une augmentation de la taille de la chaîne de type quadratique, en fixant la taille de départ à 250 et un nombre d'itérations égal à 100, pour atteindre une taille finale de chaîne de l'ordre de 10000, et la procédure de moyennisation a été démarrée à l'itération 60. Les résultats sont présentés dans le tableau 3.7 pour l'algorithme de Metropolis Hastings avec marche aléatoire adaptative proposé par Andrieu et Thoms (2008). Les colonnes 3 et 4 correspondent à la moyenne, sur les dix réalisations indépendantes, des estimateurs et de leurs écarts-types calculés par l'intermédiaire de la matrice d'information de Fisher. La dernière colonne donne l'écart-type des estimations sur les dix réalisations indépendantes, et permet de rendre compte de la variabilité des résultats entre différentes réalisations de l'algorithme.

Les résultats obtenus sont satisfaisants, et sont proches des valeurs estimées sur l'échantillon complet, c'est-à-dire en considérant les paramètres individuels comme des données observées (voir section 2.3). Le temps d'exécution est d'environ 7h30 et est le même pour chaque réalisation (à quelques minutes près). Les résultats obtenus sur ces dix réalisations indépendantes suggèrent qu'une seule réalisation de l'algorithme est suffisante, car l'erreur de Monte Carlo associée est faible. La figure 3.5 représente l'évolution des estimations en fonction de l'itération pour le paramètre θ_1.

Dans un second temps, nous nous sommes intéressés à la version automatique de l'algorithme MCMC-EM, permettant une augmentation stochastique de la taille de la chaîne, et fournissant également un critère d'arrêt de l'algorithme. Pour cela, les deux méthodes d'estimation de la variance σ_Q^2 (batch means – BM – ou overlapping batch means – OBM –) ont été comparées. Les paramètres d'ajustement de l'algorithme automatique ont été fixés à $\alpha = 0.10$, $\beta = 0.10$, et $\gamma = 0.25$, et le nombre maximal d'itérations à 500. La taille de départ pour la chaîne et le burn-in ont été fixés à 250. Des premiers tests avec des valeurs plus élevées pour α et β ne permettaient pas une augmentation suffisante de la taille de la chaîne.

Les résultats de chacune des trois configurations des algorithmes MH et hGs, obtenus avec la méthode OBM, sont donnés dans le tableau 3.8. Les résultats obtenus avec la méthode BM sont sensiblement identiques et seront donc omis. Pour des raisons de clarté, nous adoptons les notations suivantes :

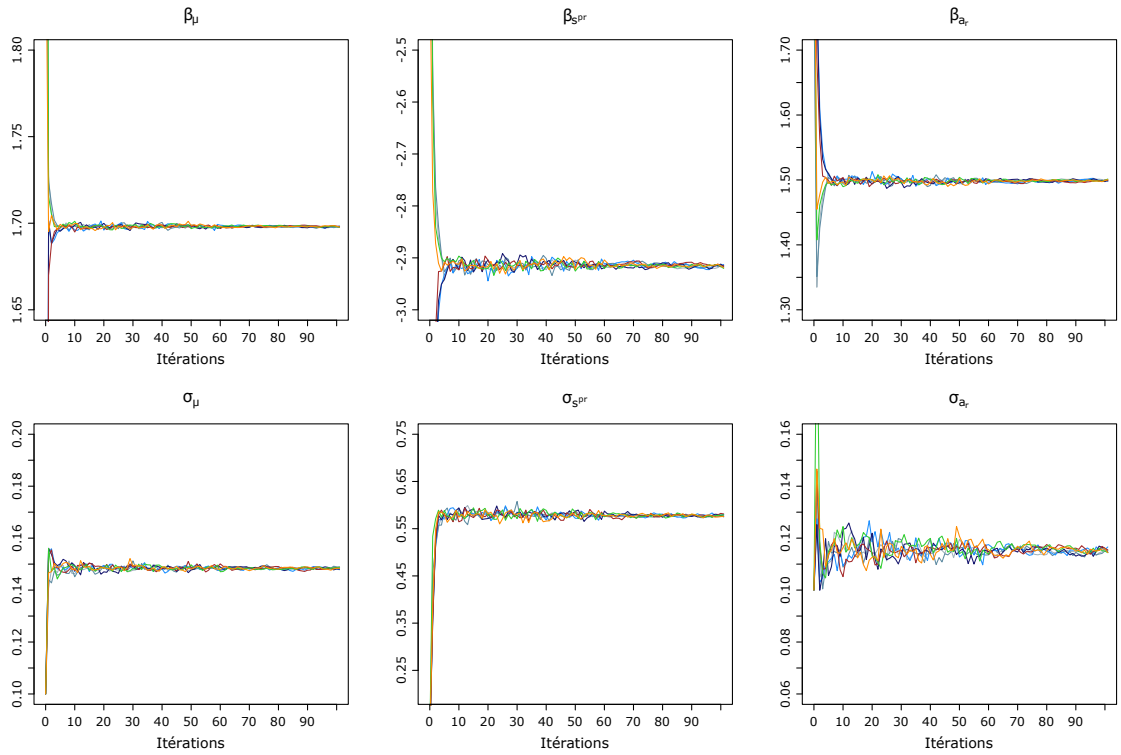


FIG. 3.5 – Évolution de $\theta_1^{(k)}$ en fonction du nombre d'itérations dans l'algorithme de Metropolis-Hastings à marche aléatoire adaptative, avec augmentation quadratique de la taille de la chaîne et moyennisation à partir de l'itération 60.

- l'algorithme MH adaptatif à marche aléatoire Gaussienne (Haario et al., 2001) présenté en 2.38 sera noté « AMH »
- l'algorithme MH adaptatif à marche aléatoire Gaussienne et avec un schéma adaptatif global (Andrieu et Thoms, 2008, Algorithme 4) présenté en 2.40 sera noté « AMH Global »
- l'algorithme hGs adaptatif basé sur une marche aléatoire et une adaptation composante par composante (Andrieu et Thoms, 2008, Algorithme 5) présenté en 2.43 sera noté « AhGs CW »
- l'algorithme hGs adaptatif basé sur une marche aléatoire globale et une adaptation composante par composante (Andrieu et Thoms, 2008, Algorithme 6) présenté en 2.44 sera noté « AhGs Global ».

Le tableau 3.9 présente une comparaison du nombre d'itérations, de la taille finale de la chaîne (générée à la dernière itération) et du temps d'exécution de chacun de ces algorithmes (le temps CPU est donc environ 6 fois plus élevé, puisque les calculs ont été parallélisés sur 6 cœurs). Les valeurs minimales, maximales et moyennes ont été obtenues sur les dix réalisations indépendantes. En particulier, l'utilisation d'un algorithme automatique permet de diminuer le temps d'exécution moyenne par rapport à l'algorithme à augmentation déterministe quadratique de la taille de la chaîne (5h12 contre 7h30).

De façon générale, les différents algorithmes fournissent des résultats satisfaisants ; cependant les meilleurs résultats sont obtenus en utilisant une loi instrumentale de type marche aléatoire adaptative. En effet, si les estimations obtenues dans chaque cas sont très similaires, on observe une plus grande variabilité entre les différentes réalisations indépendantes obtenues avec la loi marginale (voir tableau 3.8). On observe par exemple, dans le cas de l'algorithme MH, une variabilité 400 fois plus importante pour l'estimation de la variance du bruit d'observation associé aux limbes σ_b^2.

La loi marginale ne semble pas non plus être un bon choix de loi instrumentale dans le cas multivarié, ce qui se traduit notamment par une difficulté de l'algorithme automatique à augmenter la taille de la chaîne

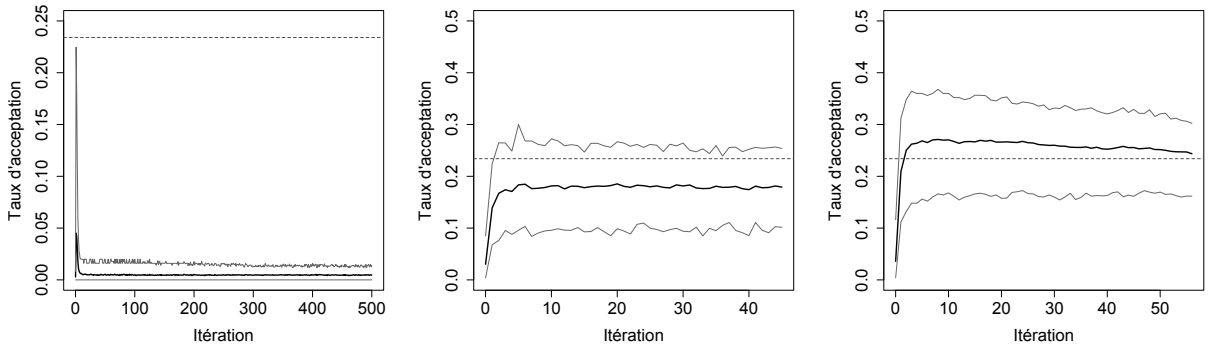
TAB. 3.8 – Comparaison des résultats des différents algorithmes en fonction du choix de la loi instrumentale (σ_Q^2 est estimée par la méthode OBM).

	Vraie valeur	Metropolis-Hastings			Échantillonneur de Gibbs		
		Estimation	ET (FIM)	ET (10 réal.)	Estimation	ET (FIM)	ET (10 réal.)
		Marginale			Marginale		
β_μ	1.7	1.6977	0.0212	0.00250	1.6974	0.0213	0.00027
σ_μ	0.15	0.1485	0.0150	0.00295	0.1489	0.0152	0.00112
$\beta_{s^{pr}}$	-3	-2.9168	0.0825	0.00611	-2.9211	0.0831	0.00232
$\sigma_{s^{pr}}$	0.5	0.5744	0.0581	0.01209	0.5748	0.0592	0.00353
β_{a_r}	1.5	1.5003	0.0180	0.00285	1.5017	0.0182	0.00098
σ_{a_r}	0.15	0.1168	0.0130	0.00489	0.1140	0.0136	0.00139
σ_b^2	0.15	0.2038	0.0058	0.00047	0.2036	0.0062	0.00004
σ_p^2	0.15	0.0843	0.0024	0.00047	0.0844	0.0026	0.00005
ρ	0.67	0.5297	0.0141	0.00197	0.5294	0.0145	0.00021
σ_r^2	0.15	0.1275	0.0303	0.01451	0.1293	0.0337	0.00268
		AMH			AhGs CW		
β_μ	1.7	1.6978	0.0213	0.00017	1.6975	0.0213	0.00022
σ_μ	0.15	0.1484	0.0152	0.00038	0.1486	0.0152	0.00031
$\beta_{s^{pr}}$	-3	-2.9185	0.0835	0.00168	-2.9211	0.0833	0.00270
$\sigma_{s^{pr}}$	0.5	0.5769	0.0598	0.00161	0.5764	0.0598	0.00192
β_{a_r}	1.5	1.5005	0.0183	0.00079	1.5017	0.0182	0.00122
σ_{a_r}	0.15	0.1148	0.0139	0.00054	0.1141	0.0136	0.00054
σ_b^2	0.15	0.2035	0.0062	0.00009	0.2035	0.0062	0.00011
σ_p^2	0.15	0.0844	0.0026	0.00002	0.0843	0.0026	0.00004
ρ	0.67	0.5291	0.0145	0.00018	0.5292	0.0145	0.00016
σ_r^2	0.15	0.1287	0.0341	0.00204	0.1304	0.0338	0.00284
		AMH Global			AhGs Global		
β_μ	1.7	1.6977	0.0213	0.00018	1.6978	0.0213	0.00032
σ_μ	0.15	0.1487	0.0152	0.00031	0.1485	0.0152	0.00058
$\beta_{s^{pr}}$	-3	-2.9190	0.0834	0.00123	-2.9186	0.0833	0.00205
$\sigma_{s^{pr}}$	0.5	0.5774	0.0598	0.00214	0.5766	0.0598	0.00081
β_{a_r}	1.5	1.5006	0.0181	0.00065	1.5004	0.0181	0.00100
σ_{a_r}	0.15	0.1145	0.0140	0.00057	0.1146	0.0139	0.00083
σ_b^2	0.15	0.2035	0.0062	0.00007	0.2035	0.0062	0.00004
σ_p^2	0.15	0.0843	0.0026	0.00002	0.0843	0.0026	0.00004
ρ	0.67	0.5293	0.0145	0.00013	0.5291	0.0145	0.00017
σ_r^2	0.15	0.1294	0.0348	0.00316	0.1304	0.0348	0.00238

Tab. 3.9 – Comparaison des différentes configurations de l'algorithme MCMC-EM.

	Nombre d'itérations			Taille de la chaîne			Temps d'exécution		
	Min	Moy	Max	Min	Moy	Max	Min	Moy	Max
	Metropolis-Hastings								
Marginale	500	500	500	337	1119	2966	8h21	12h08	21h24
AMH	21	48	72	3614	21811	46461	1h37	5h12	14h29
AMH Global	16	42	57	13354	23028	36157	2h46	5h04	6h52
	Échantilloneur de Gibbs hybride								
Marginale	55	91	128	7917	22980	40151	5h48	11h39	20h11
AhGs CW	30	72	145	5015	23465	41604	1h11	3h40	5h34
AhGs Global	26	45	72	15877	28656	45080	6h32	10h55	18h31

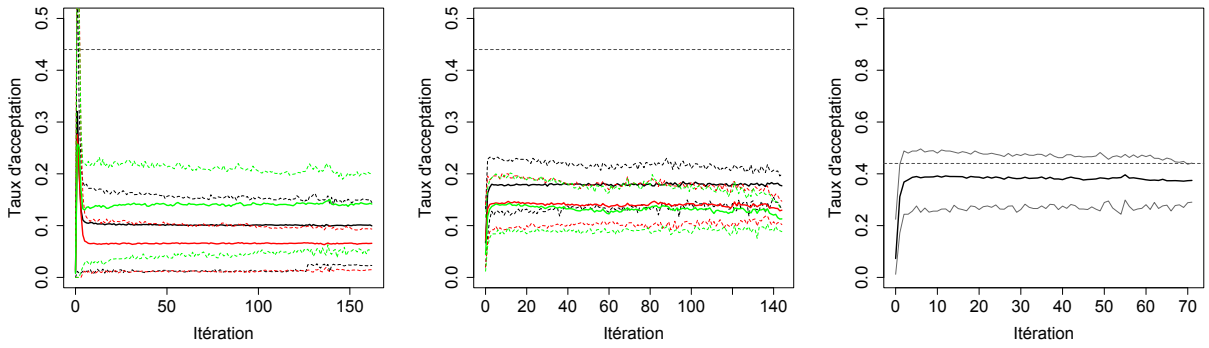
de façon suffisante (voir tableau 3.9). D'une part, l'algorithme atteint le nombre maximal d'itérations sans avoir satisfait le critère de convergence, et d'autre part la taille moyenne de la chaîne à la dernière itération est environ 20 fois plus faible qu'en utilisant une marche aléatoire adaptative. Le taux d'acceptation est également beaucoup trop faible (voir figure 3.6), bien inférieur au taux de 0.234, même si ce dernier correspond au taux optimal dans le cas d'une loi instrumentale de type marche aléatoire Gaussienne, et ne peut donc pas être extrapolé au cas de notre loi marginale. La loi marginale multivariée ne permet donc pas une bonne approximation de la loi cible. L'utilisation de la loi marginale avec l'échantillonneur de Gibbs semble moins souffrir de ces défauts. Le taux d'acceptation est plus élevé (voir figure 3.7), même s'il reste un peu trop faible, et l'algorithme automatique parvient à augmenter la taille de la chaîne et à satisfaire le critère de convergence avant d'avoir atteint le nombre maximal d'itérations. L'utilisation de lois univariées permet donc une meilleure approximation de la loi cible.



(a) Marginale (b) AMH (c) AMH Global

Fig. 3.6 – Moyenne (en noir) et quantiles d'ordre 5% et 95% (en gris) du taux d'acceptation en fonction du nombre d'itérations pour l'algorithme de Metropolis-Hastings. Chaque point correspond à la moyenne sur les 50 plantes et les 10 réalisations indépendantes de l'algorithme. Le trait en pointillés correspond au taux optimal de 0.234 (Roberts et al., 1997 ; Roberts et Rosenthal, 2001).

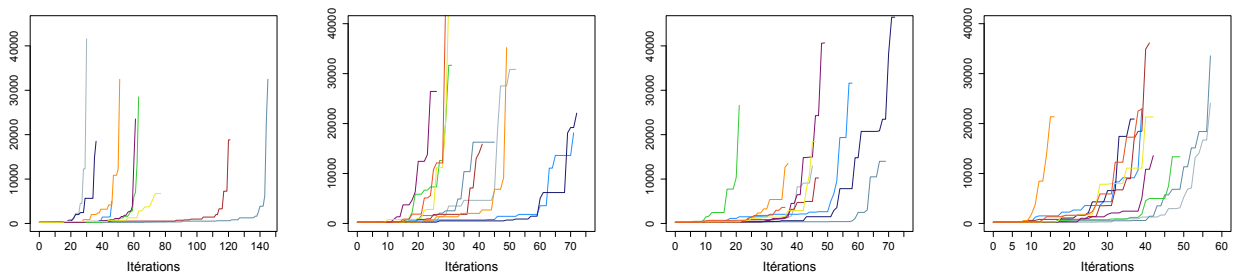
Les performances des algorithmes à marche aléatoire adaptative sont meilleures, et les résultats obtenus sont très similaires. En moyenne, sur la configuration que nous nous sommes donnée, l'algorithme MH converge avec un nombre plus faible d'itérations que l'algorithme hGs. Bien sûr, ce résultat est fortement



(a) Marginale (b) AhGs CW (c) AhGs Global

FIG. 3.7 – Moyenne (en noir ou en trait plein) et quantiles d'ordre 5% et 95% (en gris ou en pointillés) du taux d'acceptation en fonction du nombre d'itérations pour l'échantillonneur de Gibbs. Chaque point correspond à la moyenne sur les 50 plantes et les 10 réalisations indépendantes de l'algorithme. Le trait en pointillés correspond au taux optimal de 0.44 (Roberts et al., 1997 ; Roberts et Rosenthal, 2001).

dépendent de la valeur initiale des paramètres utilisée dans les algorithmes, et une initialisation différente pourrait donner des résultats différents. Une meilleure comparaison des performances des deux algorithmes serait obtenue en comparant plusieurs points de départ pour les algorithmes. Les deux algorithmes les plus rapides sont les algorithmes AMH Global et AhGs CW, et l'on observe également que les algorithmes hGs à la loi marginale et AhGs Global ont des temps de calculs relativement longs. En effet, chacun de ces algorithmes requiert à chaque itération m de la procédure MCMC, le calcul de probabilités d'acceptation « unidirectionnelles » : dans le cas de la loi marginale, il s'agit des probabilités d'acceptation pour chaque composante du vecteur d'états, et pour l'algorithme AhGs Global, même si l'acceptation du vecteur candidat se fait de façon globale en calculant une seule probabilité d'acceptation, la mise à jour du vecteur λ_m nécessite le calcul des probabilités d'acceptation correspondant à un déplacement de l'état actuel de la chaîne dans chacune des directions proposées (voir section 2.3.4). Dans la pratique, cela revient donc à générer à chaque itération de l'algorithme MCMC, autant de vecteurs candidats unidirectionnels qu'il n'y a de composantes, puis à générer des simulations du modèle Greenlab pour chacun d'eux, et à calculer la probabilité d'acceptation correspondante. Ces simulations supplémentaires entraînent une augmentation non négligeable du temps d'exécution.



(a) Algorithme AhGs CW (b) Algorithme AhGs Global (c) Algorithme AMH (d) Algorithme AMH Global

FIG. 3.8 – Taille de la chaîne en fonction du nombre d'itérations.

En comparant l'évolution de la taille de la chaîne en fonction du nombre d'itérations, on observe deux types de stratégies différentes (voir figure 3.8). L'augmentation de la taille de la chaîne se fait de façon plus

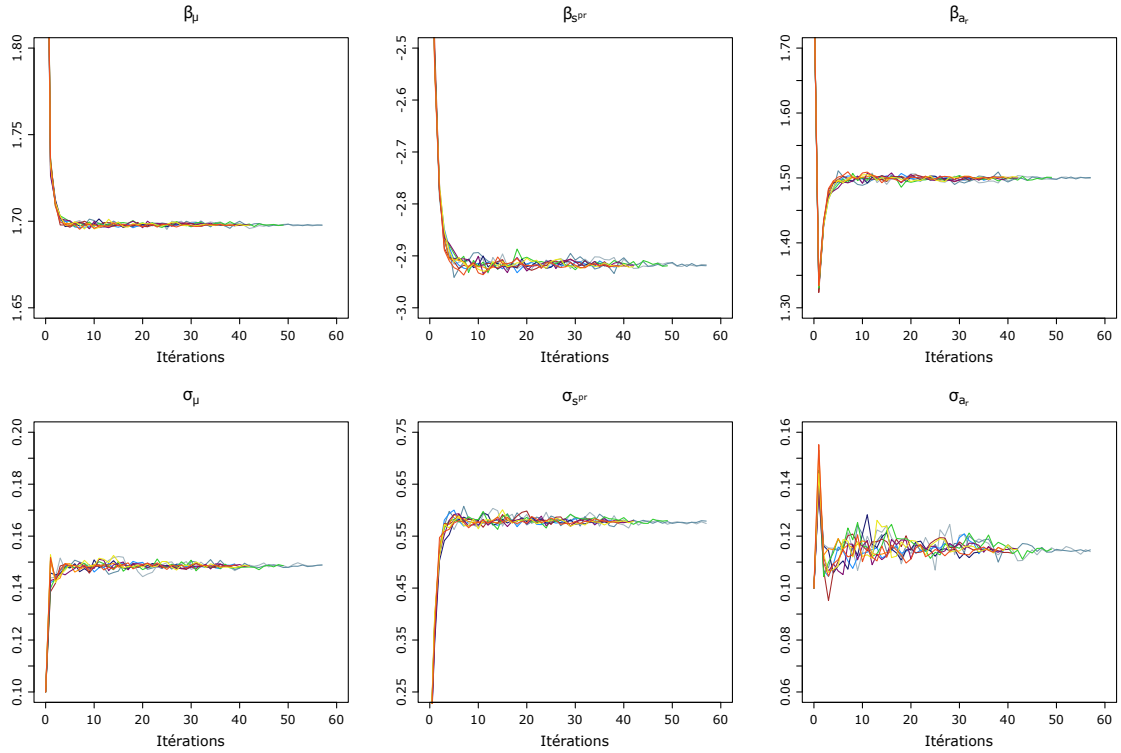


FIG. 3.9 – Évolution de $\theta_1^{(k)}$ en fonction du nombre d'itérations dans l'algorithme automatique de Metropolis-Hastings avec marche aléatoire adaptative (Andrieu et Thoms, 2008).

progressive avec les algorithmes AMH, AMH Global et AhGs Global, alors qu'avec l'algorithme AhGs CW, la taille de la chaîne reste plutôt modeste jusqu'aux deux ou trois dernières itérations à partir desquelles elle augmente brusquement, permettant à ce dernier d'avoir un temps d'exécution plus court. Cependant, le taux d'acceptation semble un peu trop faible dans le cas de l'algorithme AhGs CW (voir figure 3.7).

L'algorithme AMH semble quant à lui un peu moins performant que l'algorithme AMH Global, avec un temps d'exécution qui peut être deux fois plus important, même si les deux algorithmes ont une durée d'exécution en moyenne similaire sur nos données. L'algorithme AMH Global permet également d'obtenir un taux d'acceptation plus proche de la valeur optimale (voir figure 3.6) à mesure que le nombre d'itérations augmente, car l'estimateur de la variance Σ_π est alors calculé sur des chaînes dont la taille est de plus en plus grande, ce qui permet d'obtenir une meilleure exploration de l'espace d'états et une plus grande précision.

À titre d'illustration, nous présentons sur les figures 3.9 et 3.10 l'évolution des estimations en fonction du nombre d'itérations, pour les algorithmes AMH Global et MH à loi marginale, respectivement. La figure 3.10 illustre bien la persistance de l'erreur de Monte Carlo dans le cas de la loi marginale, qui reste assez élevée car l'algorithme ne parvient pas à augmenter la taille de la chaîne de façon suffisante.

Une comparaison des intervalles de confiance obtenus grâce à la méthode de Louis (1982) ou par Bootstrap paramétrique est présentée dans le Tableau 3.10. Cette comparaison a été faite lors des premiers tests des algorithmes, en utilisant une augmentation quadratique de la taille de la chaîne et une procédure de moyennisation des estimations. En raison des temps de calculs rédhibitoires, seulement cent échantillons Bootstrap ont été générés, et nous n'avons pas non plus relancé d'échantillons Bootstrap avec la version automatique de l'algorithme. Néanmoins les résultats obtenus avec l'algorithme automatique et avec l'algorithme déterministe étant assez proches, les intervalles de confiance obtenus par la méthode de Louis dans les deux cas sont similaires. La variabilité des estimateurs entre différentes réalisations indépendantes étant un peu plus élevée avec l'algorithme déterministe, les intervalles de confiance Bootstrap sont

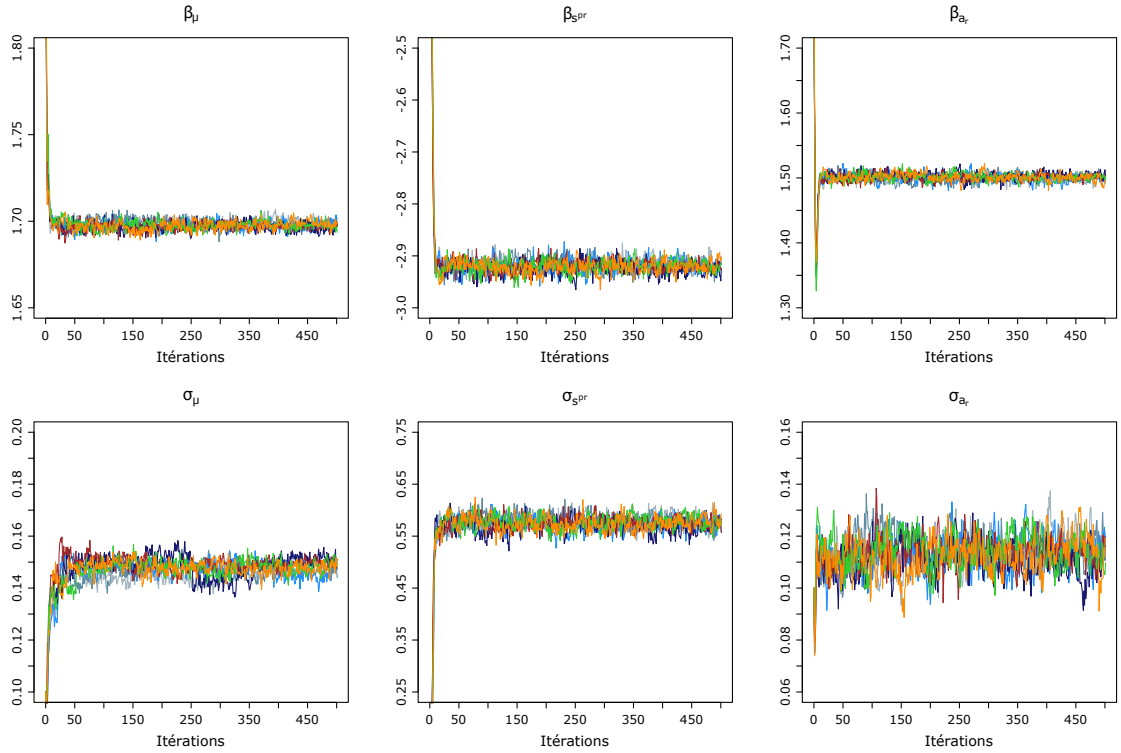


FIG. 3.10 – Évolution de $\theta_1^{(k)}$ en fonction du nombre d'itérations dans l'algorithme automatique de Metropolis-Hastings avec loi marginale comme loi instrumentale.

certainement plus larges que ceux que l'on aurait obtenus avec l'algorithme automatique.

Globalement, les résultats sont satisfaisants pour les paramètres associés aux effets aléatoires, mais sont moins bons pour les paramètres liés aux bruits d'observation. La variance σ_b^2 est sur-estimée dans l'échantillon Bootstrap, et la variance σ_p^2 est largement sous-estimée. Ceci peut être dû à la procédure de simulation, comme indiqué en section 2.3, qui a tendance à surestimer la variance des limbes et à sous-estimer celle des pétioles.

2.3.2 Algorithme SAEM

Nous présentons dans cette section les résultats obtenus avec l'algorithme SAEM, pour chacune des configurations possibles de l'algorithme MCMC. Après une période de burn-in de 250 itérations, la taille de la chaîne à chaque itération est fixée à 5, et l'on pose $K_1 = 100$ et $K_2 = 70$. Les résultats sont présentés dans le tableau 3.11.

Tout comme pour l'algorithme MCMC-EM, les résultats sont satisfaisants pour chaque configuration, même si les résultats obtenus en utilisant une loi instrumentale de type marche aléatoire Gaussienne sont plus stables. Le temps d'exécution est d'environ 1h40, sauf pour les algorithmes hGs à loi marginale et AhGs Global pour lesquels le temps d'exécution est de 2h40 et 3h15 respectivement. Il s'agit en effet des deux algorithmes reposant sur le calcul de probabilités d'acceptation composante par composante, ce qui introduit un coût supplémentaire en terme de temps de calcul.

Les figures 3.11 et 3.12 représentent l'évolution des estimations en fonction du nombre d'itérations pour les algorithmes AMH Global et MH à loi marginale respectivement. De la même façon qu'avec l'algorithme MCMC-EM, la loi marginale ne semble pas un bon choix pour la loi instrumentale, et la variabilité entre réalisations indépendantes de l'algorithme est beaucoup plus élevée qu'avec une loi instrumentale de type marche aléatoire Gaussienne (voir tableau 3.11 ou figure 3.12).

TAB. 3.10 – Comparaison des intervalles de confiance obtenus par Bootstrap paramétrique ou par la matrice d'information de Fisher.

	Vraie valeur	Obs complètes	MCMC-EM (FIM)			Bootstrap		
			Moyenne	ET	IC	Moyenne	ET	IC
β_0	1.7	1.7036	1.6981	0.0210	[1.657 ;1.739]	1.6965	0.0202	[1.656 ;1.734]
σ_0	0.15	0.1544	0.1484	0.0148	[0.119 ;0.178]	0.1458	0.0157	[0.114 ;0.173]
β_1	-3	-2.9062	-2.9159	0.0820	[-3.077 ;-2.755]	-2.9061	0.0855	[-3.088 ;-2.751]
σ_1	0.5	0.5729	0.5787	0.0586	[0.464 ;0.693]	0.5679	0.0600	[0.448 ;0.667]
β_2	1.5	1.4906	1.4992	0.0163	[1.467 ;1.531]	1.4948	0.0174	[1.46 ;1.532]
σ_2	0.15	0.1232	0.1154	0.0115	[0.093 ;0.138]	0.1145	0.0128	[0.076 ;0.136]
σ_b^2	0.15	0.2019	0.2035	0.0055	[0.193 ;0.214]	0.2277	0.0057	[0.216 ;0.239]
σ_p^2	0.15	0.0827	0.0844	0.0023	[0.08 ;0.089]	0.0380	0.0034	[0.032 ;0.043]
ρ	0.67	0.5342	0.5292	0.0139	[0.502 ;0.556]	0.6559	0.0407	[0.611 ;0.71]
σ_r	0.15	0.1389	0.1276	0.0255	[0.078 ;0.178]	0.1282	0.0389	[0.038 ;0.188]

En revanche, l'utilisation d'une marche aléatoire adaptative telle que celle proposée par Andrieu et Thoms (2008) permet d'obtenir des résultats beaucoup plus stables. La variabilité entre réalisations indépendantes de l'algorithme SAEM est similaire à celle que l'on observe entre réalisations indépendantes de l'agorithme MCMC-EM automatique, mais avec une taille de chaîne fixe et beaucoup plus faible, et un temps d'exécution plus faible. Nous avons ici fixé les paramètres K_1 et K_2, et il est évident qu'un mauvais choix de ces paramètres peut influencer significativement les résultats, en particulier si les valeurs proposées sont trop faibles et que l'algorithme est alors loin de la convergence. Il peut être nécessaire de relancer plusieurs fois l'algorithme avec différentes valeurs de ces paramètres, ou de développer des versions automatiques de l'algorithme.

Comme pour l'algorithme MCMC-EM, nous avons comparé les intervalles de confiance obtenus par la méthode de Louis (1982) et le calcul de la matrice d'information de Fisher, et ceux obtenus par Bootstrap paramétrique. Les résultats sont présentés dans le tableau 3.12. On observe le même phénomène que celui observé avec l'algorithme MCMC-EM sur les composantes de la matrice de covariance des bruits d'observation, à savoir une sur-estimation de la variance des limbes et une sous-estimation de la variance des pétioles. Globalement les résultats sont moins satisfaisant avec l'algorithme SAEM. Ceci est dû en partie à la plus grande variabilité des résultats entre réalisations indépendantes de l'algorithme, ce qui conduit à des intervalles de confiance Boostrap plus larges que ceux obtenus par la méthode de Louis. L'un des avantages de l'algorithme MCMC-EM automatique, est justement de fournir à la dernière itération une chaîne de taille beaucoup plus importante, ce qui permet d'obtenir de meilleurs estimations des intervalles de confiance.

2.4 Application sur données réelles

2.4.1 Données expérimentales

Betterave

Deux jeux de données réelles sont disponibles pour la betterave. Le premier correspond aux données de 2010 qui ont été utilisées dans le premier chapitre (voir section 2.1), et pour lequel nous disposons des masses sèches de racine, limbes et pétioles de neuf plantes choisies aléatoirement dans le champ. Cet échantillon sera noté « ITB » dans la suite.

Tab. 3.11 – Comparaison des résultats de l'algorithme SAEM.

	Vraie valeur	Metropolis-Hastings			Échantillonneur de Gibbs		
		Estimation	ET (FIM)	ET (10 réal.)	Estimation	ET (FIM)	ET (10 réal.)
		Marginale			**Marginale**		
β_μ	1.7	1.6980	0.0206	0.00108	1.6980	0.0208	0.00057
σ_μ	0.15	0.1455	0.0146	0.00329	0.1470	0.0147	0.00187
$\beta_{s^{pr}}$	-3	-2.9224	0.0812	0.01018	-2.9193	0.0820	0.00352
$\sigma_{s^{pr}}$	0.5	0.5743	0.0574	0.00895	0.5799	0.0580	0.00335
β_{a_r}	1.5	1.5015	0.0160	0.00405	1.5008	0.0163	0.00167
σ_{a_r}	0.15	0.1129	0.0113	0.00345	0.1148	0.0115	0.00269
σ_b^2	0.15	0.2037	0.0056	0.00032	0.2035	0.0057	0.00009
σ_p^2	0.15	0.0848	0.0023	0.00057	0.0845	0.0023	0.00006
ρ	0.67	0.5289	0.0139	0.00118	0.5293	0.0140	0.00033
σ_r^2	0.15	0.1314	0.0263	0.00649	0.1282	0.0257	0.00644
		AMH			**AhGs CW**		
β_μ	1.7	1.6977	0.0211	0.00039	1.6974	0.0210	0.00035
σ_μ	0.15	0.1488	0.0149	0.00042	0.1484	0.0149	0.00036
$\beta_{s^{pr}}$	-3	-2.9179	0.0822	0.00288	-2.9186	0.0816	0.00363
$\sigma_{s^{pr}}$	0.5	0.5809	0.0582	0.00237	0.5291	0.0140	0.00040
β_{a_r}	1.5	1.5000	0.0164	0.00144	1.5004	0.0160	0.00156
σ_{a_r}	0.15	0.1156	0.0117	0.00133	0.1131	0.0114	0.00206
σ_b^2	0.15	0.2035	0.0056	0.00010	0.2035	0.0056	0.00013
σ_p^2	0.15	0.0843	0.0023	0.00005	0.0843	0.0023	0.00008
ρ	0.67	0.5292	0.0140	0.00018	0.5766	0.0577	0.00231
σ_r^2	0.15	0.1277	0.0258	0.00697	0.1289	0.0259	0.00558
		AMH Global			**AhGs Global**		
β_μ	1.7	1.6979	0.0210	0.00031	1.6978	0.0211	0.00046
σ_μ	0.15	0.1485	0.0149	0.00044	0.1485	0.0149	0.00049
$\beta_{s^{pr}}$	-3	-2.9178	0.0825	0.00190	-2.9178	0.0824	0.00370
$\sigma_{s^{pr}}$	0.5	0.5830	0.0584	0.00262	0.5291	0.0140	0.00037
β_{a_r}	1.5	1.4998	0.0164	0.00096	1.5002	0.0164	0.00172
σ_{a_r}	0.15	0.1152	0.0116	0.00069	0.1156	0.0116	0.00230
σ_b^2	0.15	0.2034	0.0057	0.00011	0.2033	0.0057	0.00008
σ_p^2	0.15	0.0843	0.0023	0.00005	0.0843	0.0023	0.00004
ρ	0.67	0.5290	0.0140	0.00032	0.5823	0.0583	0.00247
σ_r^2	0.15	0.1217	0.0246	0.00366	0.1211	0.0244	0.00622

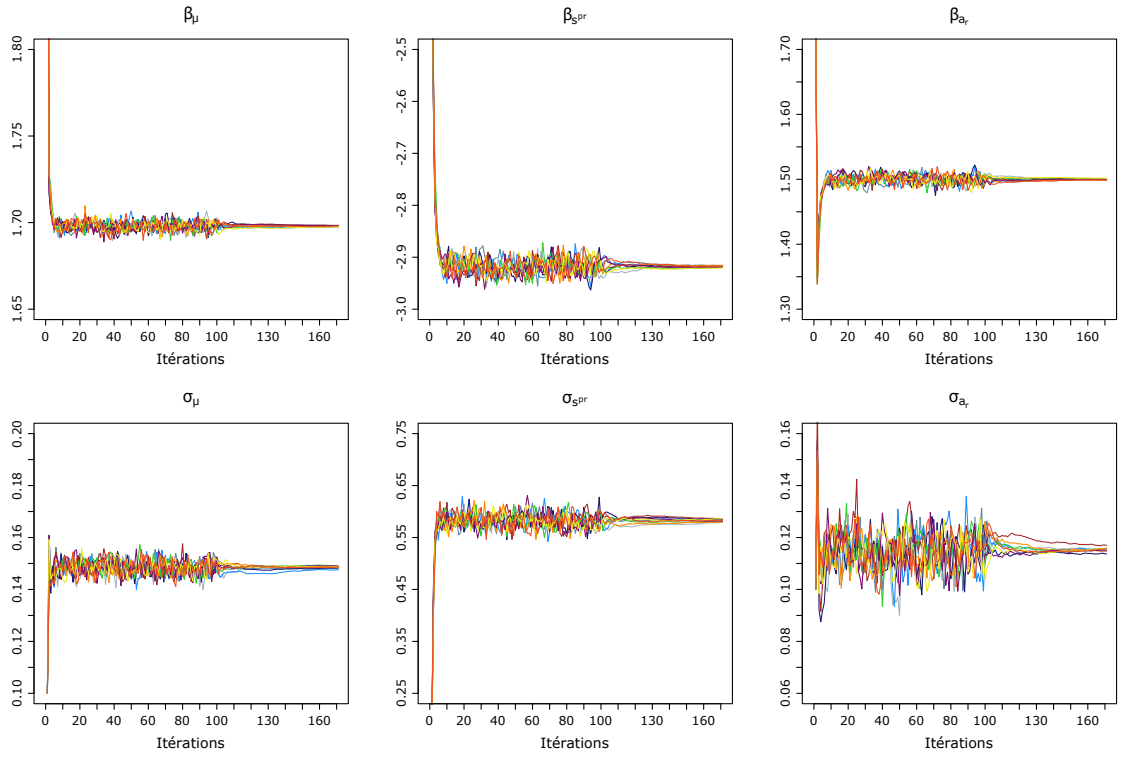


FIG. 3.11 – Évolution de $\theta_1^{(k)}$ en fonction du nombre d'itérations dans l'algorithme SAEM AMH Global

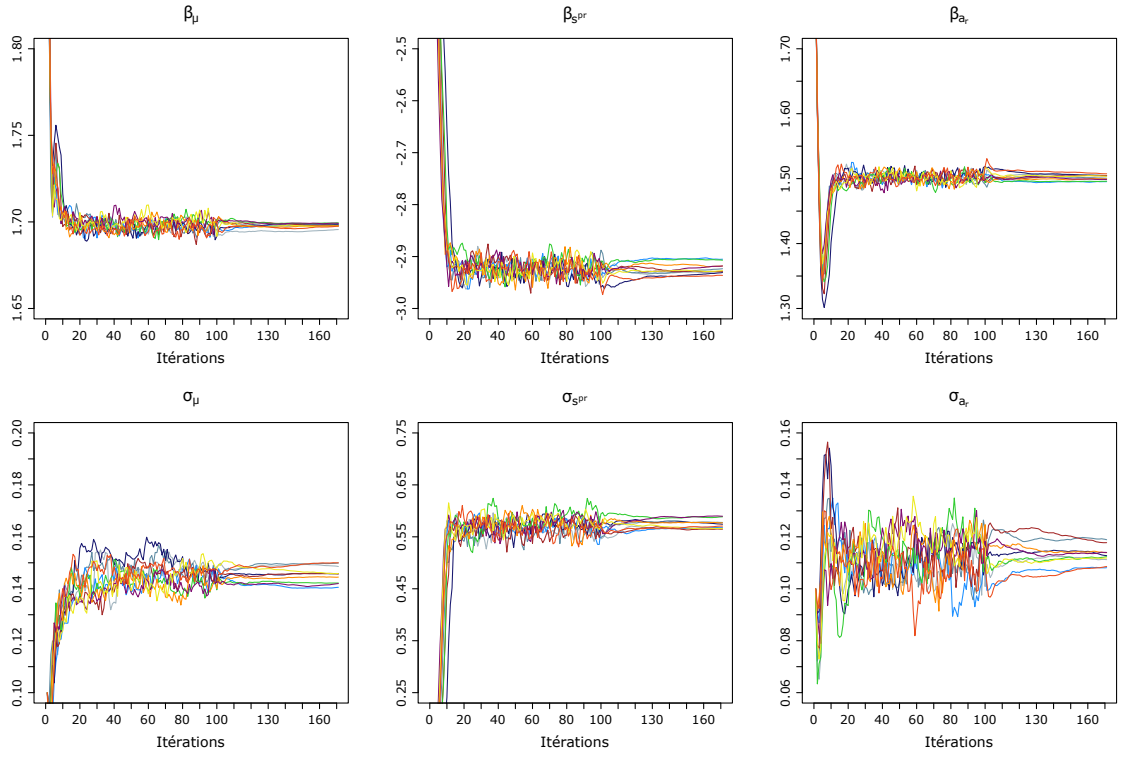


FIG. 3.12 – Évolution de $\theta_1^{(k)}$ en fonction du nombre d'itérations dans l'algorithme SAEM à loi instrumentale marginale

Tab. 3.12 – Comparaison des intervalles de confiance obtenus par bootstrap paramétrique (à droite, sur 100 échantillons Bootstrap) et par la méthode de Louis (à gauche), en utilisant l'algorithme de Metropolis-Hastings à marche aléatoire.

	Vraie valeur	Obs complètes	SAEM (FIM)			Bootstrap		
			Moyenne	ET	IC	Moyenne	ET	IC
β_μ	1.7	1.7036	1.6979	0.0210	[1.657 ;1.739]	1.7001	0.0191	[1.663 ;1.738]
σ_μ	0.15	0.1544	0.1485	0.0149	[0.119 ;0.178]	0.1488	0.0239	[0.102 ;0.196]
$\beta_{s^{pr}}$	-3	-2.9062	-2.9178	0.0825	[-3.08 ;-2.756]	-2.9182	0.0891	[-3.093 ;-2.744]
$\sigma_{s^{pr}}$	0.5	0.5729	0.5830	0.0584	[0.468 ;0.697]	0.593	0.067	[0.462 ;0.724]
β_{a_r}	1.5	1.4906	1.4998	0.0164	[1.468 ;1.532]	1.5039	0.0291	[1.447 ;1.561]
σ_{a_r}	0.15	0.1232	0.1152	0.0116	[0.092 ;0.138]	0.2004	0.0197	[0.162 ;0.239]
σ_b^2	0.15	0.2019	0.2034	0.0057	[0.192 ;0.214]	0.2285	0.0057	[0.217 ;0.24]
σ_p^2	0.15	0.0827	0.0843	0.0023	[0.08 ;0.089]	0.0381	0.0022	[0.034 ;0.042]
ρ	0.67	0.5342	0.5290	0.0140	[0.502 ;0.556]	0.6542	0.0219	[0.611 ;0.697]
σ_r^2	0.15	0.1389	0.1217	0.0246	[0.074 ;0.17]	0.1604	0.1089	[-0.053 ;0.374]

La faible taille de l'échantillon nous a incités à utiliser un deuxième jeu de données, correspondant également à une expérimentation réalisée en 2010 sur une parcelle voisine, et dont l'objectif initial était de comparer différents types de graines (deux génétiques différentes, puis deux calibres de graines différentes et enfin, graines traitées ou non par un procédé d'activation censé accélérer la levée). La taille de l'échantillon est plus importante sur ce deuxième jeu de données puisque 18 plantes ont été observées, cependant les plantes n'ont pas été choisies entièrement au hasard dans le champ : un tiers de l'échantillon correspond à des plantes dont la levée a été identifiée comme précoce, un deuxième tiers correspond à des plantes dont la date de levée était dans la moyenne, et le dernier tiers correspond à des plantes qui ont levé tardivement. À cause de ce biais de sélection, la variabilité inter-plantes dans l'échantillon est plus importante que dans la population totale. Ce deuxième jeu de données sera noté « SES ».

Colza

Les données sur le colza proviennent 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 (Jullien et al., 2011). Les graines ont d'abord été semées le 30 août 2012 dans des godets individuels, puis repiquées le 14 septembre 2012 après l'émergence des premières feuilles, et réparties dans dix bacs de tailles identiques (120×120×60 cm), remplis de terreau non azoté. La répartition des plantes dans ces dix bacs se fait selon 2 critères, l'un portant sur la densité de plantation et l'autre sur l'hétérogénéité de la taille des plantes au moment du repiquage (voir figure 3.13).

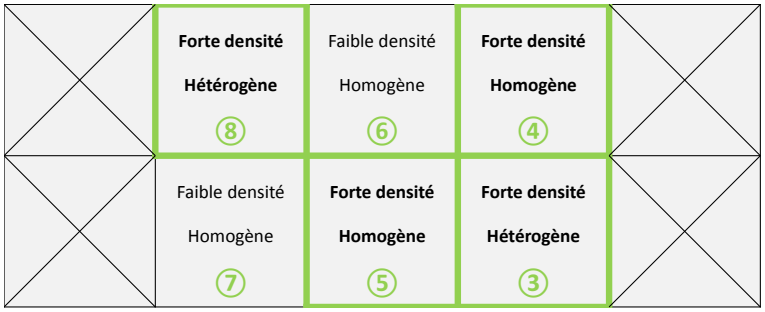


Fig. 3.13 – Protocole expérimental pour les données colza.

Dans les bacs de faible densité, 5 rangées de 4 plantes ont été plantées, contre 6 rangées de 7 plantes dans les bacs de forte densité, ce qui correspond à une densité de 42 plantes par m^2. Dans la pratique, le caractère hétérogène ou homogène étant basé sur la taille des jeunes plantules, il n'est pas réellement discriminant, et nous avons donc associé les données provenant des bacs de forte densité à peuplement « homogène » et « hétérogène ». Un test de Wilcoxon-Mann-Whitney a été réalisé sur les données de masses de feuilles et racines provenant de ces deux catégories et ne nous a pas permis de rejeter l'hypothèse nulle selon laquelle les deux échantillons proviennent de la même population ($p = 0.11$ pour les racines et $p = 0.90$ pour les feuilles).

Finalement, comme nous nous intéressons seulement à la première phase de croissance du colza, nous utiliserons les données provenant des bacs ③ et ④, pour lesquels les mesures ont été faites avant que la plante n'atteigne le stade de montaison, soit le 21 mars 2013. Les masses sèches individuelles des feuilles et de la racine de vingt plantes du bac ③ et de 14 plantes du bac ④ ont été relevées. Le nombre de feuilles a également été relevé hebdomadairement sur ces mêmes plantes.

2.4.2 Résultats

Le nombre d'observations par plante dépend directement des paramètres d'organogenèse, qui doivent donc faire partie des données d'entrées du modèle. Dans le cas des données simulées, ces paramètres sont générés lors de l'étape de simulation et sont donc considérés comme connus, mais ils doivent être estimés dans le cas des données réelles.

Pour cela, nous avons utilisé le modèle d'organogenèse développé dans la première section de ce chapitre, à partir des données de suivi du nombre de feuilles. Les paramètres individuels d'organogenèse sont estimés pour chaque plante i à l'aide du logiciel Monolix, et sont ensuite considérés comme des données d'entrées du modèle.

Les résultats des deux algorithmes MCMC-EM et SAEM ont été comparés, en utilisant les deux lois instrumentales AMHG et CWhGs, qui se sont révélées les plus performantes sur les données simulées. Les deux modèles d'erreur, additif ou log-additif, ont également été comparés sur les deux jeux de données betterave et colza, à l'aide des critères AIC et BIC.

Nous utiliserons dans la suite, comme sur le jeu de données simulées, une transformation logarithmique des paramètres du modèle, qui sont tous strictement positifs par définition.

Betterave

Comme avec le jeu de données simulées, les trois paramètres μ, s^{pr} et a_r ont été considérés comme aléatoires. Une première calibration du modèle Greenlab sur la plante « moyenne » a été réalisée, pour les deux modèles additif et log-additif, afin d'obtenir des valeurs initiales pour les algorithmes (voir tableau 3.13). Pour l'échantillon ITB, l'estimation des trois paramètres μ, s^{pr} et a_r s'est avérée suffisante pour obtenir une bonne calibration, mais pour l'échantillon SES, la calibration d'un plus grand nombre de paramètres a été nécessaire.

Une période de burn-in de 1000 itérations a été utilisée pour les deux algorithmes, ce qui est supérieur à la valeur utilisée dans le cas des données simulées, mais nécessaire pour s'assurer que la chaîne de Markov générée à chaque itération de l'algorithme est bien dans son régime stationnaire. Pour l'algorithme MCMC-EM automatique, le nombre maximal d'itérations a été fixé à 500, et plusieurs choix pour les paramètres α et β ont été testés (voir plus bas). Pour l'algorithme SAEM, un premier test a été lancé avec $K_1 = 500$ et $K_2 = 200$, et les résultats suggéraient que des valeurs plus faibles pouvaient être utilisées,

TAB. 3.13 – Calibration du modèle Greenlab sur la plante moyenne pour chaque échantillon et pour chaque modèle (additif ou log-additif). Ces valeurs sont utilisées comme valeurs initiales pour les algorithmes MCMC-EM et SAEM.

	Estimation sur plante moyenne		
Paramètre	ITB (log)	ITB (add)	SES
μ	7.179	7.951	4.940
$\log \mu$	1.9712	2.073	1.5974
s^{pr}	0.0334	0.0311	0.0244
$\log s^{pr}$	-3.999	-3.470	-3.713
p_p	1.0884	1.0884	0.28423
q_p	0.2346	0.2346	1.25
a_b	3	3	1.48
b_b	3	3	2.71
a_p	3	3	2.63
b_p	3	3	4.88
a_r	4.48	3.96	3.79
$\log a_r$	1.500	1.376	1.332
b_r	1.35	1.35	1.74

puisque les estimations semblaient se stabiliser après environ un centaine d'itérations. Nous avons donc fixé pour la suite $K_1 = 200$ et $K_2 = 100$. Compte tenu de la faible taille de l'échantillon, plusieurs chaînes ont également été générées en parallèle à chaque itération de l'algorithme, puis les résultats de ces différentes chaînes ont été combinés. Le nombre de chaînes a été fixé à 5, et la taille de chacune de ces chaînes à 10. Les modèles additif et log-additif ont été comparés à l'aide des critères AIC et BIC.

Quelques problèmes numériques ont été rencontrés lors de l'application des algorithmes sur les données betterave, en particulier sur l'échantillon ITB :

1. choix des paramètres α et β pour la version automatique de l'algorithme MCMC-EM : des valeurs peu élevées de ces paramètres entraînent une forte augmentation de la taille de la chaîne pour obtenir la précision requise, ce qui pose des problèmes de stockage en mémoire ou de temps d'exécution. En effet, le temps de calcul étant limité sur le mésocentre de calcul de Centrale, le programme est interrompu passé un délai de 24h, quel que soit son statut. Plusieurs valeurs de α et β ont été testées sur l'échantillon ITB, sans qu'aucune des configurations ne permette d'obtenir la convergence de l'algorithme dans le temps imparti. Certaines réalisations de l'algorithme en étaient encore aux toutes premières itérations au bout de 24h de calcul, ce qui signifie que l'algorithme automatique a nécessité dans ces cas là une augmentation très rapide et très forte de la taille de la chaîne.

2. choix de la loi instrumentale : dans le cas du modèle additif, l'utilisation de la loi instrumentale AMHG a posé quelques problèmes lors du calcul de la décomposition de Cholesky des matrices de covariance. Nous n'avons pas rencontré ces problèmes lors de l'utilisation de la loi instrumentale CWhGs.

Si la version automatique de l'algorithme MCMC-EM avec $\alpha = \beta = 0.15$ a permis d'obtenir des résultats sur l'échantillon SES sous les contraintes de calcul évoquées ci-dessus, nous avons eu recours à une augmentation quadratique de la taille de la chaîne sur l'échantillon ITB, en fixant le nombre d'itérations à 100, et la taille initiale de la chaîne à 250 ; pour obtenir une taille finale de 10000 environ. Le meilleur modèle au sens des critères AIC et BIC est le modèle à erreur log-additive, pour les deux échantillons ITB

TAB. 3.14 – Résultats obtenus sur l'échantillon ITB, avec loi instrumentale AMHG.

	MCMC-EM			SAEM		
	Estimation	Ecart-type	IC	Estimation	Ecart-type	IC
β_μ	2.7166	0.0765	[2.567 ;2.867]	2.7753	0.0165	[2.743 ;2.808]
σ_μ	0.0499	0.0491	[0 ;0.146]	0.0482	0.0118	[0.025 ;0.071]
$\beta_{s^{pr}}$	-2.0428	0.3646	[-2.757 ;-1.328]	-2.0865	0.1167	[-2.315 ;-1.858]
$\sigma_{s^{pr}}$	0.3182	0.2966	[0 ;0.899]	0.3492	0.0826	[0.187 ;0.511]
β_{a_r}	0.8208	0.0054	[0.810 ;0.831]	0.8000	0.0002	[0.7996 ;0.8004]
σ_{a_r}	0.0050	0.0090	[0 ;0.023]	0.0006	0.0002	[0.00033 ;0.00095]
σ_b^2	1.1067	0.0503	[1.008 ;1.205]	1.1031	0.1165	[0.875 ;1.332]
σ_p^2	0.9714	0.0441	[0.885 ;1.058]	0.9689	0.1018	[0.769 ;1.168]
ρ	0.9412	0.0059	[0.964 ;0.987]	0.9408	0.0059	[0.961 ;0.984]
σ_r^2	1.8747	0.8937	[0.123 ;3.626]	1.9575	0.9233	[0.148 ;3.767]

et SES. Nous présentons dans les tableaux 3.14 et 3.15 les résultats obtenus avec les deux algorithmes et en utilisant une loi instrumentale de type AMHG, sur les échantillons ITB et SES, respectivement.

Si les résultats des deux algorithmes sont similaires concernant l'estimation des paramètres (voir aussi figure 3.14), on observe une grande différence entre les écart-types des estimateurs obtenus par la méthode de Louis (1982), ceux de l'algorithme MCMC-EM étant plus élevés que ceux de l'algorithme SAEM. En particulier, tous les intervalles de confiance des écart-types des effets aléatoires obtenus par l'algorithme MCMC-EM contiennent la valeur 0. L'effet aléatoire ayant la plus grande variance est celui relié au paramètre s^{pr}, qui peut s'interpréter comme un indice de compétition entre plantes, puisqu'il correspond à la projection orthogonale de la surface occupée par la plante au sol : plus la densité de plantation, et donc la compétition pour les ressources augmente, plus ce paramètre diminue. La variance de l'efficience de conversion est plus faible, et celle du paramètre a_r est également très faible. On pourrait tester, sur un échantillon de taille plus importante, la significativité de la variabilité de ce paramètre dans la population. Les valeurs moyennes dans la population sont également assez différentes des valeurs obtenues sur la plante moyenne (voir tableau 3.13), et en particulier l'efficience de conversion. Rappelons que nous avons utilisé une transformation logarithmique des paramètres, ce qui correspond ici à une efficience moyenne de 15.13 g.MJ^{-1}.pl^{-1}. Associée à une valeur moyenne du paramètre s^{pr} de 0.1297, cette valeur reste clairement trop élevée, car elle correspond à une efficience au niveau du mètre carré de 19.34 g.MJ^{-1}.

Sur l'échantillon SES, même si le nombre de plantes est plus élevé que sur l'échantillon ITB, la forte variabilité induite par la méthode d'échantillonnage a également compliqué l'estimation des paramètres du modèle. En particulier, les résultats de différentes réalisations de l'algorithme, même en utilisant des valeurs initiales identiques, n'étaient pas toujours consistants. En revanche, l'algorithme MCMC-EM automatique avec $\alpha = 0.15$ et $\beta = 0.15$ n'a pas soulevé les mêmes difficultés que sur l'échantillon ITB. La loi instrumentale CWhGs a permis d'obtenir une taille de chaîne deux fois plus importante qu'avec la loi AMHG, ce qui a permis d'obtenir des intervalles de confiance de meilleure qualité. Avec la loi AMHG, le calcul de la matrice d'information de Fisher manquante, estimée à partir des réalisations de la chaîne de Markov, renvoyaient pour certains estimateurs des écart-types aberrants (négatifs). Nous présentons donc dans le tableau 3.15 les résultats obtenus à partir de la loi instrumentale CWhGs.

On observe tout d'abord quelques différences entre les résultats obtenus avec les deux algorithmes, en particulier sur les paramètres de variance des effets aléatoires qui sont en général plus faibles avec l'algorithme SAEM qu'avec l'algorithme MCMC-EM. Les résultats de la figure 3.15 semblent suggérer que

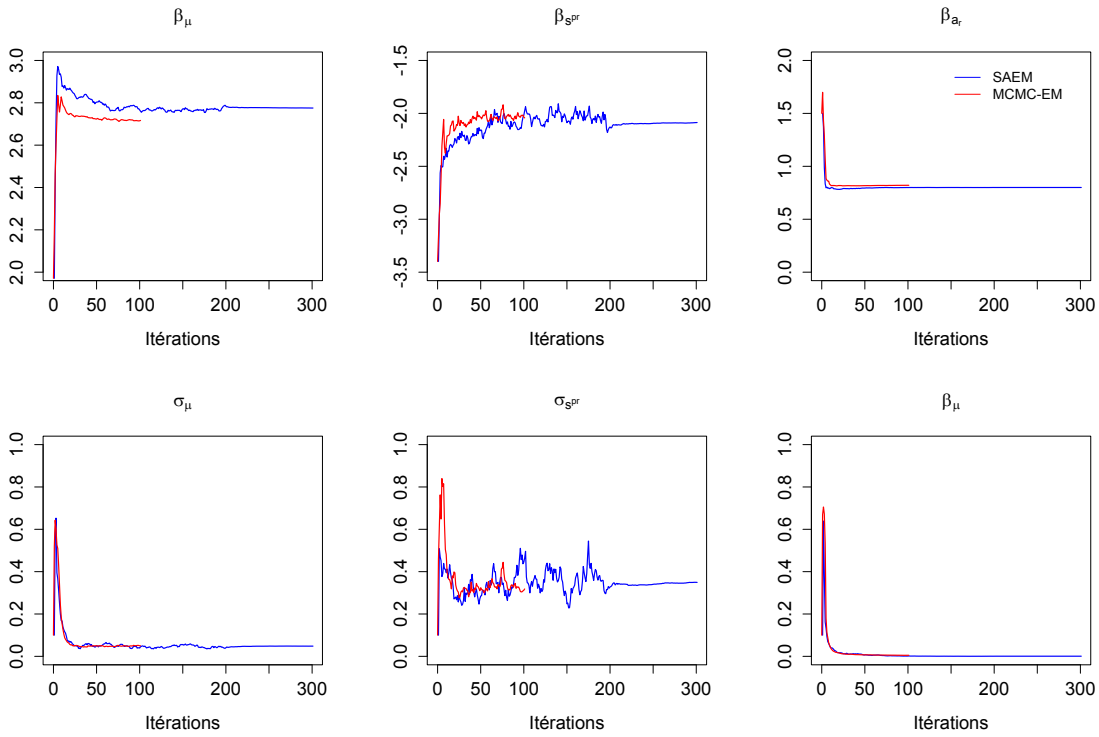


FIG. 3.14 – Comparaison des algorithmes SAEM et MCMC-EM (avec augmentation quadratique de la taille de la chaîne) sur les données ITB.

TAB. 3.15 – Résultats obtenus sur l'échantillon SES, avec loi instrumentale CWhGs

	MCMC-EM			SAEM		
	Estimation	Ecart-type	IC	Estimation	Ecart-type	IC
β_μ	3.4775	0.0950	[3.291 ;3.664]	3.1607	0.0007	[3.159 ;3.162]
σ_μ	0.0696	0.0863	[0 ;0.239]	0.0026	0.0005	[0.0016 ;0.0036]
$\beta_{s^{pr}}$	-4.9411	0.0950	[-5.127 ;-4.755]	-4.6517	0.0784	[-4.805 ;-4.498]
$\sigma_{s^{pr}}$	0.3305	0.0803	[0.173 ;0.488]	0.3322	0.0556	[0.223 ;0.441]
β_{a_r}	1.3142	0.0167	[1.281 ;1.347]	1.3173	0.0002	[1.317 ;1.318]
σ_{a_r}	0.0308	0.0221	[0 ;0.074]	0.0008	0.00014	[0.00052 ;0.00107]
σ_b^2	1.2569	0.0459	[1.167 ;1.347]	1.2630	0.1015	[1.064 ;1.462]
σ_p^2	2.4083	0.0892	[2.233 ;2.583]	2.4319	0.1951	[2.050 ;2.814]
ρ	0.9464	0.0043	[0.938 ;0.955]	0.9465	0.0043	[0.938 ;0.955]
σ_r^2	1.0882	0.3752	[0.353 ;1.824]	1.0667	0.3557	[0.369 ;1.764]

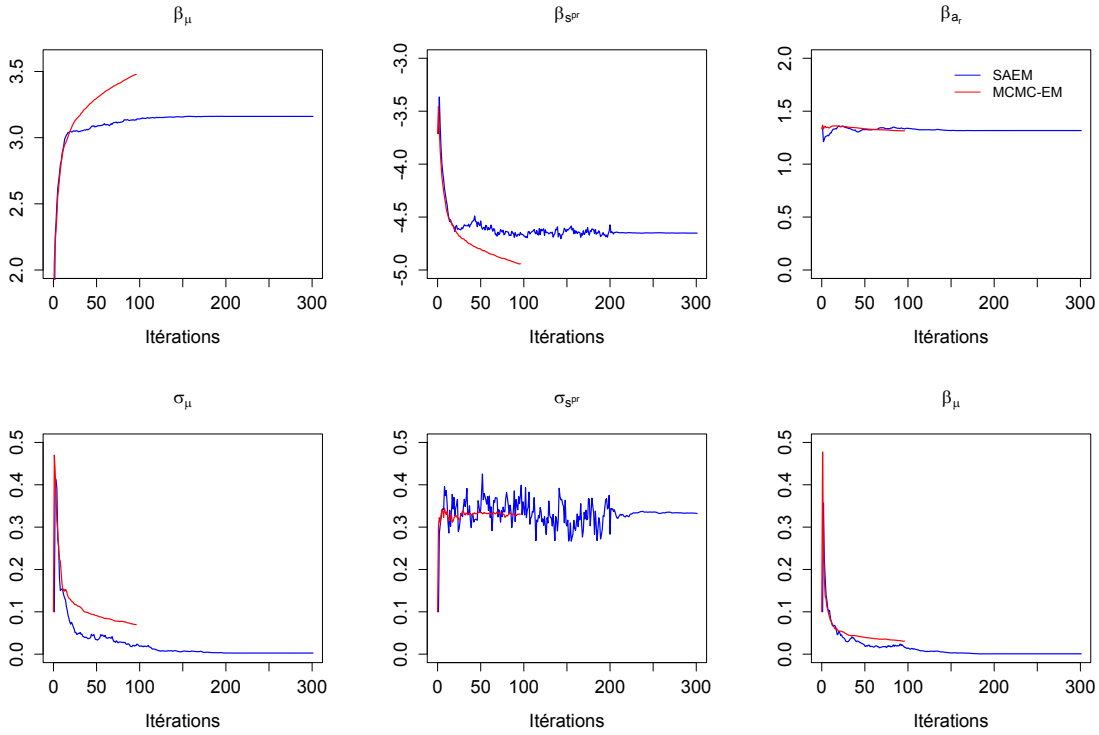


FIG. 3.15 – Comparaison des algorithmes SAEM et MCMC-EM (avec augmentation quadratique de la taille de la chaîne) sur les données ITB.

l'algorithme MCMC-EM automatique s'est arrêté trop tôt, ce qui peut être dû à la définition du critère d'arrêt à partir des quantités δ et γ (voir chapitre 2, section 2.3). Ce critère d'arrêt dépend également de la quantité $\hat{\sigma}^2_Q/\sqrt{m_k}$ à l'itération k, or ce terme décroît avec la taille de la chaîne, et donc à mesure que le nombre d'itérations de l'algorithme MCMC-EM augmente. Il est alors possible de définir un critère d'arrêt plus restrictif en choisissant des valeurs plus élevées pour le paramètre γ. Les résultats obtenus montrent, comme dans le cas de l'échantillon ITB, une sur-estimation de la valeur moyenne de μ, encore plus flagrante ici, puisque l'on obtient une efficience moyenne de 32.38 g.MJ^{-1}.pl^{-1} ($= \exp(3.4775)$), ce qui est beaucoup trop élevé d'un point de vue biologique. Cependant, le coefficient s^{pr} est beaucoup plus faible sur cet échantillon, avec une valeur moyenne estimée à 0.0071 m^2.pl^{-1}, ce qui correspond finalement à une efficience au niveau du mètre carré de l'ordre de 2.31 g.MJ^{-1}. Il pourrait donc s'agir plutôt d'un problème d'identifiabilité entre ces paramètres, des phénomènes de compensation pouvant exister (voir chapitre 1). La forte variabilité de l'échantillon se traduit par une forte variabilité résiduelle sur les bruits d'observation, plus élevée que sur l'échantillon ITB. Comme sur les données ITB, la plus forte variabilité est celle du paramètre s^{pr}, les deux autres étant plus faibles.

Les deux échantillons de données betterave nous ont permis une première application des modèles sur données réelles, et une première appréciation des problèmes numériques potentiels. Cependant, ils se sont avérés insuffisants, soit par leur petite taille, soit par la méthode d'échantillonnage. C'est pourquoi nous proposons dans le paragraphe suivant une application au cas du colza au stade rosette, pour lequel nous disposons d'un plus grand nombre d'observations (34 plantes), et avec un modèle Greenlab plus simple que dans le cas de la betterave.

Tab. 3.16 – Calibration du modèle Greenlab sur la plante moyenne pour chaque échantillon et pour chaque modèle (additif ou log-additif). Ces valeurs sont utilisées comme valeurs initiales pour les algorithmes MCMC-EM et SAEM.

	Modèle	
Paramètre	Additif	Log-additif
$\log \mu$	1.1554	1.1731
$\log s^{pr}$	-3.7297	-3.7297
$\log a_l$	1.2182	0.9518
$\log b_l$	1.0986	1.0986
μ	3.2321	3.1754
s^{pr}	0.024	0.024
a_l	3.3812	2.5904
b_l	3	3

Colza

Le modèle Greenlab que nous utilisons pour le colza au stade rosette est plus simple que pour la betterave, avec notamment moins de paramètres, ce qui implique a priori moins de possibilités de compensation entre les paramètres et moins de problèmes d'identifiabilité. Une première calibration du modèle sur la plante moyenne a permis d'obtenir des valeurs initiales pour les algorithmes (voir tableau 3.16).

Les paramètres du modèle étant tous positifs par définition, nous avons appliqué une transformation logarithmique dans le modèle Greenlab de population, et la calibration sur plante moyenne a donc également été faite en considérant le logarithme de chaque paramètre. Nous présentons dans le tableau 3.16 les résultats obtenus sur les paramètres transformés, et nous indiquons également à titre d'information et pour l'interprétation biologique, la valeur correspondant au paramètre initial non transformé. Seuls quatre paramètres interviennent dans le modèle puisque l'on considère uniquement le stade rosette : l'efficience de conversion, le paramètre s^{pr} relié à la projection au sol de la surface occupée par la plante, et les deux paramètres de la fonction d'allocation de biomasse aux feuilles, a_l et b_l. Dans la pratique, le paramètre s^{pr} est souvent considéré comme constant, et fixé égal à l'inverse de la densité (Jullien et al., 2011), ce qui correspond ici à une valeur de 0.024, soit -3.7297 lorsque l'on utilise une transformation logarithmique. La calibration des deux paramètres μ et b_l a permis d'obtenir la meilleure calibration sur la plante moyenne.

Concernant le modèle d'organogenèse, deux phases de développement sont normalement observées au cours de la phase rosette, comme dans le cas de la betterave, et la deuxième phase débute en général autour de la mi-janvier, lorsque le semis a lieu début septembre. La température de base pour le colza étant de 4.5°C, cela correspond à un temps thermique compris entre 715°et 717°C. Or, peu de points de mesures sont disponibles après cette date (voir figure 3.16), ce qui rend difficile l'estimation de 4 paramètres d'organogenèse. Pour éviter les problèmes d'identifiabilité, nous avons donc considéré une seule phase de développement, décrite par seulement deux paramètres : le temps thermique d'initiation et le phyllochrone. Ce modèle s'est avéré beaucoup plus stable que le précédent, mais a l'inconvénient de sous estimer le nombre final de feuilles par plante, car la deuxième phase de développement, que nous ne prenons pas en compte dans ce modèle simplifié, correspond en fait à une augmentation du rythme d'apparition des feuilles.

Les paramètres individuels estimés par le modèle d'organogenèse ont ensuite été utilisés comme paramètres d'entrée pour le modèle de population. Pour ce dernier, plusieurs versions ont été testées en fonction du nombre de paramètres aléatoires et du type d'erreur (additive ou log-additive) considérés, et ont été comparées à l'aide des critères AIC et BIC. La configuration de chacun des algorithmes est résumée dans

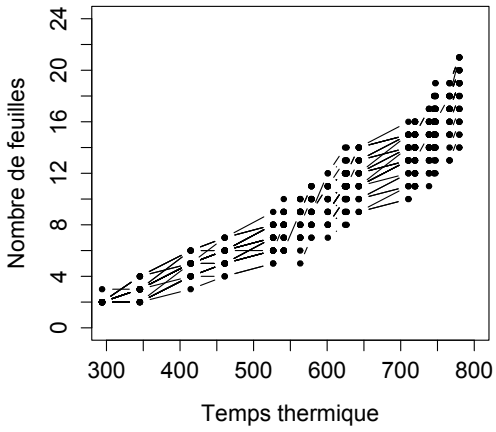


FIG. 3.16 – Nombre de feuilles en fonction du temps thermique pour les 34 plants de colza.

le tableau 3.17.

TAB. 3.17 – Configuration des algorithmes d'estimation MCMC-EM et SAEM sur les données colza.

MCMC-EM		SAEM	
Paramètre	Valeur	Paramètre	Valeur
α	0.15	K_1	100
β	0.15	K_2	70
γ	0.25	L (nombre de chaînes)	1
Burn-in	1000	Burn-in	1000
Taille initiale de la chaîne	250	Taille de la chaîne	10
Nombre maximal d'itérations	500		

L'algorithme MCMC-EM automatique s'est avéré beaucoup plus rapide que l'algorithme SAEM dans le cas du modèle à erreur additive : 15 minutes seulement contre 1h40 pour l'algorithme SAEM, mais la différence est beaucoup plus faible dans le cas du modèle à erreur log-additive puisque le temps d'exécution de l'algorithme MCMC-EM passe alors à 1h30, alors que le temps d'exécution du SAEM reste inchangé. Ceci s'explique principalement par la taille de la chaîne générée à chaque itération de l'algorithme : lors des dernières itérations de l'algorithme automatique, la taille de la chaîne générée dans le cas du modèle à erreur log-additive est en moyenne deux fois deux élevée que dans le cas du modèle à erreur additive. La calibration préalable du modèle sur la plante moyenne nous a également permis d'obtenir de bonnes valeurs initiales pour les algorithmes. A titre d'illustration, nous avons également relancé l'algorithme SAEM en posant $K_1 = 25$ et $K_2 = 15$ (ces valeurs étant suggérées par un examen visuel des courbes obtenues avec les premières valeurs définies dans le tableau 3.17), et nous avons obtenu des résultats proches de ceux présentés ici, mais avec un temps d'exécution de 8 minutes seulement.

Dans les deux cas, les résultats obtenus avec les deux algorithmes sont similaires. Le modèle à erreur additive est globalement meilleur que le modèle à erreur log-additive sur ce jeu de données (voir tableau 3.19). Dans le cas d'une erreur log-additive, le meilleur modèle au sens des critères AIC et BIC est celui où les deux paramètres μ et a_l sont considérés comme aléatoires, les autres étant fixés. Dans le cas du modèle à erreur additive, c'est le même modèle qui est sélectionné par le critère BIC, mais c'est le modèle complet qui est le meilleur au sens du critère AIC. Cependant, les valeurs de AIC obtenues pour le modèle complet

TAB. 3.18 – Résultats obtenus avec les algorithmes MCMC-EM automatiques et SAEM et la loi instrumentale CWhGs. Une transformation logarithmique des paramètres a été utilisée, et nous avons donc $\beta_\mu := \mathbb{E}(\log \mu)$, $\beta_{b_l} := \mathbb{E}(\log b_l)$, $\sigma_\mu = \sqrt{\mathrm{Var}(\log \mu)}$, et $\sigma_{b_l} = \sqrt{\mathrm{Var}(\log b_l)}$.

Paramètre	MCMC-EM			SAEM		
	Estimation	Écart-type	IC	Estimation	Écart-type	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_l}	0.5799	0.0204	[0.5400 ; 0.6198]	0.5804	0.0119	[10.5571 ; 0.6037]
σ_{a_l}	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]

et celui où μ et a_l sont considérés comme aléatoires sont très proches. Nous choisissons donc de retenir le modèle à deux paramètres aléatoires, μ et a_l. Les résultats obtenus pour ce modèle avec les algorithmes MCMC-EM et SAEM et loi instrumentale CWhGs sont présentés dans le tableau 3.18. Différentes valeurs initiales ont également été testées et ont conduit aux mêmes résultats (voir figure 3.21).

Comme lors de la calibration sur plante moyenne, l'estimation des deux paramètres μ et a_l suffit à obtenir une bonne calibration du modèle de population. Les résultats obtenus suggèrent également que le paramètre d'allocation a_l permet de mieux prendre en compte la variabilité inter-individuelle que le paramètre b_l, puisque les deux modèles contenant ce paramètre sont meilleurs que ceux contenant b_l, au sens des deux critères AIC et BIC. À b_l fixé, ce paramètre permet de modéliser différentes stratégies d'allocation de la biomasse aux feuilles : plus la valeur de a_l est faible, plus l'allocation se fait tôt (voir figure 3.17). Ce paramètre porte donc sur la première partie de la courbe d'allocation, qui correspond au début de l'expansion des feuilles et qui est donc particulièrement importante ici puisque tous les organes n'ont pas atteint leur taille maximale, certains étant encore en expansion au moment où les mesures ont été faites.

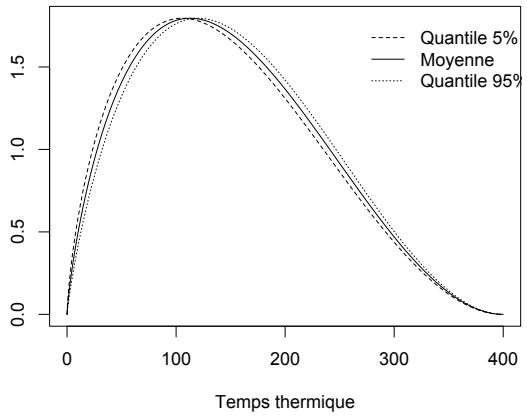


FIG. 3.17 – Comparaison des fonctions d'allocation de biomasse aux feuilles pour le modèle Greenlab colza, en fonction du paramètre a_l, lorsque le paramètre b_l est fixé égal à 3. La courbe en trait continu correspond à la valeur moyenne de $\log a_l$ estimée dans la population, et les deux courbes en pointillés correspondent aux quantiles d'ordre 5% et 95%.

En revanche, l'ajout du paramètre s^{pr} comme effet aléatoire n'améliore pas suffisamment la vraisemblance du modèle pour compenser l'augmentation du nombre de paramètres. Avec toutes les précautions

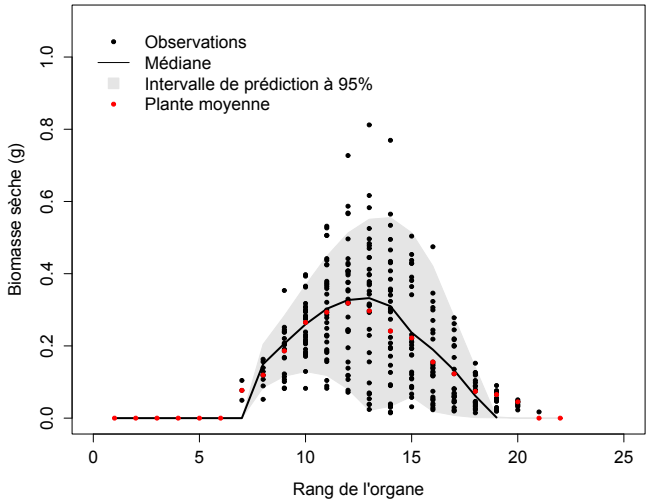


FIG. 3.18 – Prédictions du modèle Greenlab de population pour le colza. Les points rouges correspondent à la plante « moyenne » qui est utilisée classiquement pour calibrer le modèle Greenlab individuel, la ligne continue correspond à la médiane des prédictions et la zone grisée aux quantiles d'ordre 5% et 95%.

qui s'imposent en l'absence de tests statistiques, cela pourrait toutefois 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_l. En effet, dans les modèles contenant l'effet aléatoire s^{pr}, l'écart-type de l'estimateur de la variance $\sigma^2_{s^{pr}}$ est assez élevé et conduit à des intervalles de confiance contenant la valeur 0. Ceci contraste avec les résultats obtenus sur la betterave, mais 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 (Jullien et al., 2011).

La figure 3.18 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. On observe une bonne prise en compte de la variabilité inter-individuelle par le modèle, sauf pour les organes de rang élevé (supérieur à 18). Ceci est dû à la sous-estimation du nombre total de feuilles par le modèle d'organogenèse. En effet, supposons que le nombre total de feuilles pour la plante i, calculé à partir des estimations individuelles des paramètres d'organogenèse, soit estimé à $\tilde{n}_i < n_i$, où n_i est le nombre total de feuilles *observé* pour cette même plante. Alors, les biomasses des feuilles de rang $\tilde{n}_i + 1, \ldots, n_i$, calculées à chaque itération des algorithmes MCMC-EM et SAEM (voir 3.25) sont nulles, et il en est de même pour chaque prédiction de la plante i. Sur notre jeu de données, cette sous-estimation se traduit par une décroissance un peu trop rapide de la biomasse des feuilles à partir du quinzième rang. De la même façon, la médiane des prédictions décroît plus vite que les données de la plante moyenne. Ceci a également une influence sur les résidus du modèle ou erreurs de prédiction décrites au chapitre 2, en section 3, car les termes correspondant sont alors tous égaux à 1 (les biomasses prédites étant nulles, elles sont toujours inférieures aux biomasse observées), ce qui se traduit par un pic dans la distribution des termes d'erreurs autour de la valeur 1. Pour cette raison, le test réalisé sur les erreurs de prédictions nous a conduit à rejeter l'hypothèse nulle de normalité ($p < 0.001$).

L'un des avantages de la méthode proposée, basée sur les modèles mixtes, est de pouvoir obtenir des estimations individuelles pour les paramètres. Nous présentons en figures 3.19-3.20 les calibrations pour

chaque plante, obtenues à partir de ces paramètres individuels. Les résultats sont plutôt satisfaisants, étant donnée la variabilité des profils individuels. Si certaines plantes un peu atypiques sont moins bien modélisées, comme par exemple la plante 15 ou la plante 28, dans l'ensemble les profils foliaires sont bien restitués.

Tab. 3.19 – Comparaison des différents modèles testés pour le colza. En gras sont indiqués les valeurs minimales des critères AIC et BIC.

Erreur	Modèle	$-2\,\ell$	AIC	BIC
Additive	μ, b_l	-633.15	-623.15	-604.37
	μ, s^{pr}, b_l	-634.50	-620.50	-594.22
	μ, a_l	-708.47	-698.47	**-679.70**
	μ, s^{pr}, a_l	-707.05	-693.05	-666.77
	μ, s^{pr}, a_l, b_l	-721.02	**-703.02**	-669.27
Log-additive	μ, b_l	132.18	142.18	160.97
	μ, s^{pr}, b_l	131.93	145.93	172.22
	μ, a_l	130.39	**140.39**	**159.17**
	μ, s^{pr}, a_l	130.04	144.04	170.33
	μ, s^{pr}, a_l, b_l	132.15	150.15	183.95

2.5 Discussion

Dans cette section, nous avons proposé une extension du modèle Greenlab à l'échelle de la population, où certains des paramètres du modèle sont considérés comme aléatoires. Nous avons proposé une estimation des paramètres par la méthode du maximum de vraisemblance, en utilisant deux versions stochastiques de l'algorithme Espérance-Maximisation (EM). Comme la vraisemblance complète du modèle appartient à la famille exponentielle, les deux étapes de l'algorithme EM s'écrivent simplement en fonction des statistiques exhaustives du modèle, et l'étape de maximisation peut se résoudre de façon explicite.

Les deux algorithmes MCMC-EM et SAEM, basés tous les deux sur des méthodes de type Monte Carlo par Chaîne de Markov, et avec différents choix d'algorithmes et de lois instrumentales ont été comparés. Les différentes configurations testées ont fourni des résultats similaires et satisfaisants sur les données simulées, même si l'utilisation des algorithmes de Metropolis-Hastings à marche aléatoire adaptative globale (AMHG) et de Gibbs hybride à marche aléatoire adaptative composante par composante (hGsCW) ont été les plus performants du point de vue du temps de calcul. L'utilisation de la loi marginale comme loi instrumentale ne semble pas être un bon choix, que ce soit en terme de temps de calcul ou de taux d'acceptation. Dans le cas multivarié de l'algorithme de Metropolis-Hastings, on observe même une variabilité supplémentaire entre réalisations indépendantes de l'algorithme où chaque fois, le nombre maximal d'itérations sans que l'algorithme automatique ne soit parvenu à augmenter la taille de la chaîne de façon satisfaisante.

L'implémentation d'une version automatique de l'algorithme MCMC-EM nous a permis de gagner en moyenne 2h de temps de calcul par rapport à une augmentation quadratique de la taille de la chaîne, sur nos données simulées. L'intérêt de cet algorithme est de proposer une augmentation suffisante de la taille de la chaîne permettant de compenser l'erreur de Monte-Carlo et basée sur la propriété de monotonie de l'algorithme EM, ainsi qu'un critère d'arrêt. D'un autre coté, l'algorithme SAEM que nous avons étudié ici est également particulièrement efficace en terme de temps de calcul, du fait de la faible taille de la chaîne générée à chaque itération. La variabilité observée entre différentes réalisations indépendantes est plus forte

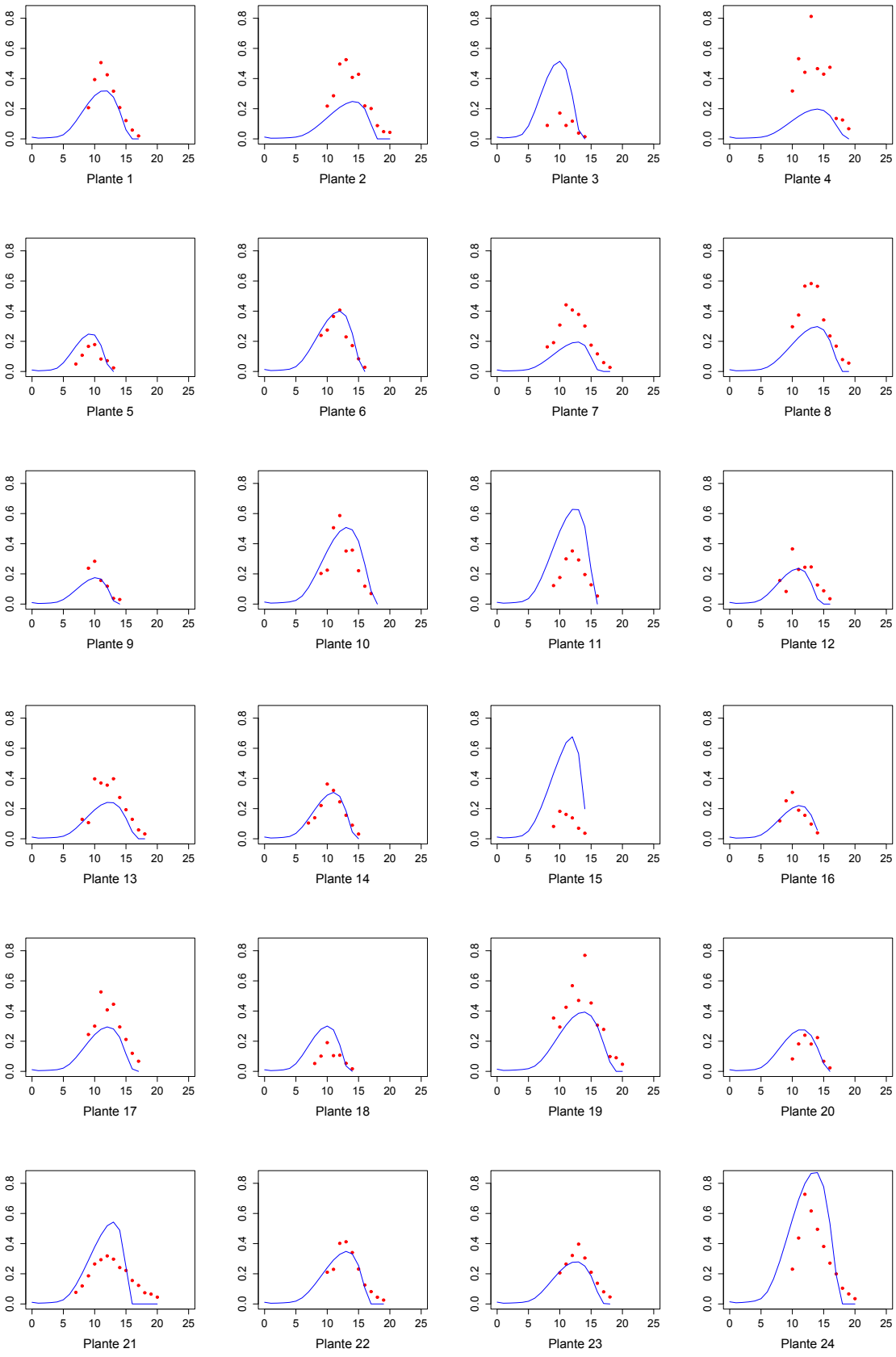


FIG. 3.19 – Prédictions à partir des paramètres individuels estimés par le modèle de population.

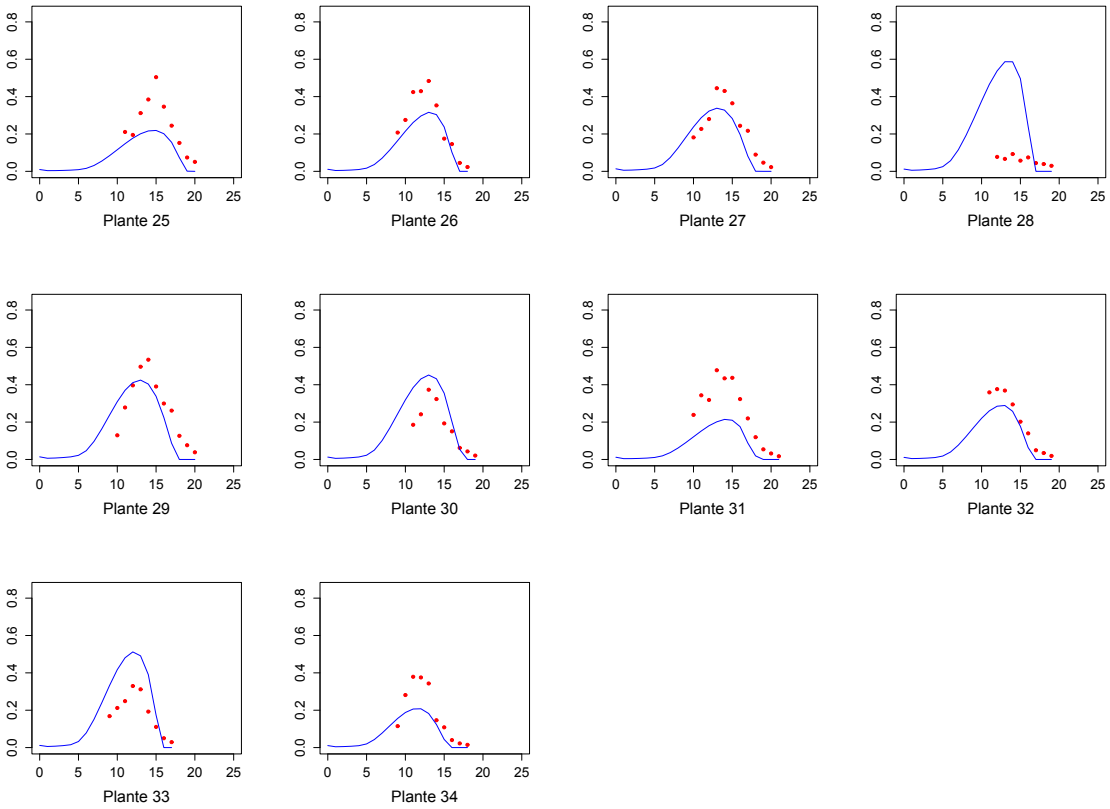


FIG. 3.20 – Prédictions à partir des paramètres individuels estimés par le modèle de population (suite).

qu'avec l'algorithme MCMC-EM, mais cette variabilité pourrait être réduite en développant une version automatique de l'algorithme.

Les résultats obtenus sur les données réelles sont également encourageants, même si dans le cas de la betterave nous avons eu plus de difficultés qu'avec les données colza. Ceci peut s'expliquer par la qualité des deux jeux de données betterave, le premier étant de taille relativement faible, et le second étant construit sur une méthode d'échantillonnage ayant introduit un biais de sélection des plantes et une forte variabilité dans l'échantillon. Les résultats obtenus sur les données colza sont toutefois très encourageants. Parmi les différents modèles testés, le meilleur modèle retenu est celui correspondant à un modèle d'erreur additif avec deux paramètres aléatoires, l'efficience de conversion μ et le paramètre a_l de la loi d'allocation, qui permet de différencier plusieurs stratégies d'allocation de la biomasse aux feuilles. En particulier, ce paramètre a une influence sur la première phase de ce processus, ce qui est particulièrement important lorsque les organes n'ont pas tous terminé leur expansion. Les prédictions obtenues avec cette formulation sont satisfaisantes, mais le couplage avec le modèle d'organogenèse n'est pas parfait, car il conduit à une sous-estimation du nombre total de feuilles et donc une moins bonne prise en compte de la variabilité des feuilles de rang élevé. Ceci pourrait être amélioré en augmentant la fréquence de suivi du nombre de feuilles par exemple, ou la durée d'observation, afin de pouvoir correctement estimer les 4 paramètres d'organogenèse. L'intérêt de notre approche est également de proposer des estimations individuelles des paramètres pour chaque plante dans la population.

D'un point de vue pratique, l'algorithme MCMC-EM automatique s'exécute en moyenne en une quinzaine de minutes sur les données colza avec modèle d'erreur additif, même si le temps d'exécution sans parallélisation est environ 6 fois plus élevé. L'algorithme SAEM est plus long, mais cela tient au fait que nous n'avons pas encore implémenté de version automatique de l'algorithme. En théorie, ceci

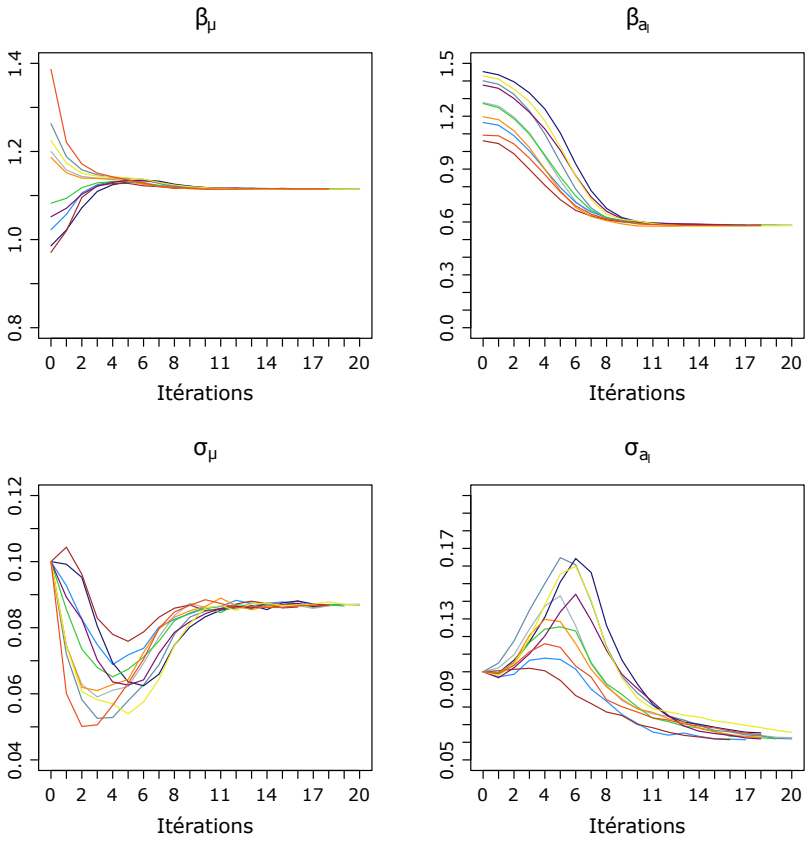


FIG. 3.21 – Évolution de $\theta_1^{(k)}$ en fonction du nombre d'itérations pour différentes valeurs initiales de β (algorithme MCMC-EM et loi instrumentale CWhGs).

permettrait d'améliorer sensiblement les performances de l'algorithme en terme de temps de calcul, et si la convergence peut être obtenue avec un même nombre d'itérations que pour l'algorithme MCMC-EM, le temps d'exécution et les besoins en mémoire seront moindre avec l'algorithme SAEM, comme le montre l'exemple où l'on a réduit les deux paramètres K_1 et K_2. C'est un point qu'il est nécessaire d'approfondir.

Une des limites de notre étude concerne les hypothèses restrictives selon lesquelles : (i) tous les effets introduits dans le modèle sont aléatoires, et (ii) la matrice de covariance des effets aléatoires est diagonale. Ces hypothèses peuvent bien sûr être relâchées, en autorisant par exemple la variance de certains effets aléatoires à être nulle, ce qui revient à considérer l'effet correspondant comme fixe, ou en supposant une matrice symétrique définie positive de forme générale pour la matrice de covariance. La possibilité de considérer certains paramètres comme effets fixes paraît essentielle, et permettrait sans aucun doute d'améliorer les résultats obtenus sur les jeux de données réels, en particulier pour la betterave, où nous avons observé de faibles valeurs pour les variances de certains effets aléatoires. Il est également possible avec ce type de formulation, de tester si certains effets doivent être introduits dans le modèle comme aléatoires, ou s'ils doivent être considérés comme fixes, en utilisant des tests du rapport de vraisemblance. Le modèle obtenu n'appartient plus à la famille exponentielle, et la maximisation des paramètres provenant des effets fixes ne se fait plus nécessairement de façon explicite, mais des méthodes d'optimisation de type quasi-Newton peuvent alors être utilisées.

Discussion et perspectives

"En vérité, le chemin importe peu, la volonté d'arriver suffit à tout"
Albert Camus, *Le mythe de Sisyphe*

Nous nous sommes efforcés de répondre dans cette thèse aux deux problématiques majeures soulevées dans l'introduction, concernant d'une part l'évaluation et la comparaison des modèles de croissance de plantes utilisés comme outils prédictifs, et d'autre part la prise en compte de la variabilité inter-individuelle dans les populations de plantes. Si les travaux présentés proposent des éléments de réponses à certaines des questions posées, certaines restent encore en suspens, et d'autres émergent également de ces travaux. Nous présentons dans ce dernier chapitre un rappel des résultats obtenus et des contributions de cette thèse, puis nous proposons plusieurs axes de développement futurs.

1 Principaux résultats et contributions

1.1 Sélection de modèles pour la prévision

L'un des objectifs de cette thèse était de proposer une approche de type « benchmarking » dans le domaine des modèles de croissance de plantes. En effet, un grand nombre de modèles ont été développés depuis les années 1970, avec des objectifs précis et parfois distincts. Malgré cela, ces modèles sont parfois utilisés hors du cadre dans lesquels ils ont été élaborés, et il n'est pas toujours évident d'anticiper leurs performances en dehors de ce contexte. De plus, très peu d'études se sont attachées à comparer les performances des modèles disponibles.

Dans la première partie de cette thèse, nous nous sommes intéressés à la prévision de la production de biomasse et du rendement chez la betterave sucrière, pour laquelle différents modèles sont couramment utilisés, certains nécessitant la calibration d'un grand nombre de paramètres, comme le modèle STICS, ou des données de calibration qui peuvent être difficiles à obtenir expérimentalement, comme le modèle Greenlab. Nous avons mis en place une approche en deux étapes, dans le but de construire et d'évaluer cinq modèles de croissance de plantes utilisés comme outils prédictifs : Greenlab, LNAS, STICS, Pilote, CERES.

Une première étape d'analyse de sensibilité a permis d'identifier les paramètres les plus influents, et d'élaborer une version plus robuste de chaque modèle, en diminuant le nombre de paramètres à estimer. Ainsi, dans le contexte de notre étude, nous avons montré qu'il était possible de réduire le nombre de paramètres à estimer dans le modèle STICS, en passant de 18 à seulement 1 paramètre, les autres étant fixés à des valeurs de référence données dans la littérature. Les différentes analyses de sensibilité ont également confirmé quantitativement l'importance capitale du paramètre d'efficience, qui a été classé comme le plus

influent dans chaque cas. La deuxième étape de notre étude a consisté en la comparaison à proprement dite des qualités prédictives des cinq modèles, évaluées sur un jeu de données indépendant de celui sur lequel a eu lieu la calibration.

Plusieurs résultats intéressants ont été obtenus dans cette étude. D'un point de vue pratique, nous avons notamment montré qu'une formulation linéaire de la production de biomasse dans le modèle STICS permettait d'obtenir de meilleurs résultats que la version initiale du modèle, où cette relation est quadratique. Le modèle STICS s'est par ailleurs révélé très robuste sur le jeu de données test de 2008. D'autre part, les modèles reposant sur un indice de récolte empirique pour la répartition de la biomasse entre racine et partie aérienne se sont avérés globalement corrects, même si nous avons observé que cet indice empirique est lui-même très peu robuste. Les modèles de type structure-fonction ont montré de bonnes performances, et ont également l'avantage de rendre accessible certains processus impliqués dans le fonctionnement de la plante et donc de permettre la prévision d'un plus grand nombre de variables. Par exemple, il est possible avec Greenlab de prédire l'évolution des profils individuels de feuilles, ce qui n'est pas possible avec des modèles comme Pilote ou CERES. De façon plus générale, les modèles de culture ou agronomiques peuvent être vus comme des « projections » des modèles de type structure-fonction.

Nous avons également mis en évidence l'importance capitale du temps thermique d'initiation. Une mauvaise estimation de ce paramètre peut en effet conduire à de très mauvaises performances prédictives, ce qui se caractérise notamment par une sous-estimation systématique lorsque ce dernier est trop précoce, ou au contraire à une sur-estimation lorsqu'il est trop tardif. Il existe également une forte variabilité de ce temps thermique d'initiation, qui peut également expliquer en partie la difficulté de sa prise en compte dans les modèles de type individu-centré. Une étude récente (Feng et al., 2014) a notamment montré comment la prise en compte de la variabilité de ce paramètre pouvait améliorer les performances des modèles de type individu-centré.

1.2 Variabilité inter-individuelle

Le deuxième objectif de cette thèse portait sur la prise en compte de la variabilité inter-individuelle dans les populations de plantes.

Dans une première étude sur la variabilité du processus d'organogenèse chez la betterave, nous avons proposé l'utilisation d'un modèle mixte, permettant de prendre en compte cette variabilité, et de comparer différentes populations en tenant compte de cette variabilité. Nous avons notamment montré que la variabilité des quatre paramètres d'organogenèse était significative, et que le temps de rupture correspondant au changement de rythme d'émission des feuilles, n'était pas corrélé aux autres paramètres. Ce changement de phase semble donc plutôt lié à un nombre relativement stable de feuilles, estimé à environ 35. Dans un second temps nous avons comparé l'effet de différentes doses d'azote sur ces quatre paramètres, et nous avons montré que l'administration d'azote retardait l'initiation, mais augmentait le rythme d'émission des feuilles. Ces résultats se rapprochent de ceux observés sur d'autres plantes comme le navet, le colza, …, (Fletcher et al., 2012). Par ailleurs, aucune différence significative n'a été observée entre les deux doses d'azote, suggérant que la dose standard d'azote est suffisante pour obtenir un nombre maximal de feuilles.

Même si plusieurs auteurs ont déjà étudié les caractéristiques de ces paramètres d'organogenèse, et l'effet de différents traitements agricoles sur ces paramètres, ces études sont la plupart du temps basées sur des valeurs moyennes, ou supposent l'indépendance entre les mesures d'une même plante. On peut citer notamment Clerget et al. (2008) pour l'étude du lien avec la température du sol chez le sorgho, Cao et Moss (1989) pour une étude similaire sur le blé et l'orge, ou plus généralement Wilhelm et McMaster (1995)

pour une revue des facteurs ayant une influence sur le phyllochrone. Lemaire et al. (2008) ont également proposé une étude du lien entre ces paramètres et la densité de plantation. Le modèle que nous avons proposé, au contraire, utilise l'ensemble des données, sans avoir recours à une plante « moyenne », et ne fait pas l'hypothèse restrictive d'indépendance. Chaque paramètre du modèle a également une interprétation biologique. De plus, il est alors possible de comparer différentes populations de plantes *en prenant en compte* la variabilité du processus d'organogenèse, ce qui n'était pas le cas des études précédentes basées sur des données moyennes.

2 Perspectives

Nous proposons dans cette section plusieurs axes de développement possibles pour chacune des deux problématiques abordées dans cette thèse.

2.1 Sélection de modèles pour la prévision

Concernant la sélection de modèles pour la prévision, les résultats que nous avons présentés sont basés sur deux jeux de données test, provenant des expérimentations 2008 et 2011. Seulement, un épisode de grêle a eu lieu en 2011, ce qui a conduit à la destruction totale ou partielle d'un grand nombre de feuilles, et ce qui nous a donc empêché d'évaluer correctement les performances des modèles en 2011. De façon plus générale, il faudrait étoffer les résultats de l'étude en évaluant les modèles sur un plus grand nombre de jeux de données, correspondant à des conditions environnementales ou à des variétés différentes. On s'attend cependant à de moins bonnes performances si les caractéristiques de l'échantillon test sont trop éloignées de celles de l'échantillon d'apprentissage. Une façon d'améliorer les performances des modèles serait alors de prendre notamment en compte les stress environnementaux, ce qui est directement possible pour le stress hydrique, dans les modèles Pilote et STICS. Pilote a en effet été conçu initialement pour permettre de piloter l'irrigation en proposant une estimation du bilan hydrique de la culture, et STICS possède plusieurs modules dédiés aux stress environnementaux, et notamment hydrique, thermique, azoté, ... Une étape supplémentaire pourrait être de développer pour chaque modèle une version avec stress, afin de déterminer l'importance de leur prise en compte pour la prédiction. Cependant, un des dangers de ce type d'approche réside dans une possible sur-paramétrisation du modèle, comme illustrée par exemple dans le cas du modèle STICS, pour lequel une formulation quadratique s'est avérée moins pertinente qu'une formulation linéaire pour la production de biomasse. Ce point est également discuté par Chen et al. (2013a) dans le cadre de l'assimilation de données. Yin et Struik (2010) mettent également en évidence la forte interaction qui existe entre les différents processus éco physiologiques mis en jeu lors de la croissance d'une plante, et la nécessité de prendre en compte la variabilité génétique, pour modéliser les interactions génotype-environnement et améliorer la robustesse des modèles.

Concernant le temps thermique d'initiation, l'influence que peut avoir ce paramètre sur la suite des prédictions du modèle plaide en faveur d'une meilleure prise en compte du processus de germination et/ou d'émergence dans les modèles de croissance de plantes. Plusieurs modèles empiriques existent déjà, et pourraient être utilisés en amont de chaque modèle prédictif. Il pourrait également être intéressant de comparer avec le modèle STICS, qui contient déjà un module lié à l'initiation, les performances avec ou sans calcul de ce temps d'initiation, ce qui pourrait révéler l'intérêt d'en tenir compte dans un objectif de prédiction.

Par ailleurs, nous n'avons étudié ici que le cas de la betterave sucrière, et pour cette plante, nous nous sommes seulement intéressés à la prédiction de la production de biomasse et du rendement. Ces deux variables ont été choisies d'une part pour leur importance dans le cas de la betterave sucrière, mais également parce qu'elles étaient facilement calculables par les cinq modèles étudiés. Cependant, il faut noter qu'avec les modèles agronomiques comme Pilote ou CERES, il n'est souvent pas possible de prédire d'autres quantités que la production de biomasse et le rendement. En ce sens, ils peuvent être vus comme des simplifications des modèles de type structure-fonction. Une comparaison complète des différents types de modèle doit donc prendre en compte le fait qu'en dépit d'une plus grande complexité en général, des modèles de type structure-fonction, ces derniers permettent l'estimation ou la prédiction d'un plus grand nombre de variables.

Une autre façon d'améliorer les prédictions des modèles serait d'utiliser des méthodes de type assimilation de données, qui reposent sur une formulation du type modèle de Markov caché. Les méthodes utilisées appartiennent à la classe des méthodes de filtrage particulaires, et permettent notamment d'actualiser à chaque pas de temps l'estimation des paramètres et des états du système en fonction de points déjà observés. Dans ses travaux de thèse, Yuting Chen a notamment montré que ce type d'approches permettait d'améliorer les performances en divisant l'erreur relative de prédiction par six, lorsque les $n-2$ points de mesures sont utilisés pour actualiser les estimations, les deux derniers étant prédits Chen et al. (2013a).

2.2 Variabilité inter-individuelle

2.2.1 Modèle d'organogenèse

Le modèle mixte d'organogenèse mériterait d'être testé sur un plus grand jeu de données, puisque seulement 20 plantes ont été utilisées dans notre cas. Ces mesures sont faciles à réaliser dans la pratique et peuvent se faire de façon non destructive, il paraît donc possible, si l'on décide d'utiliser ce type d'approche, d'obtenir de plus grands échantillons. Il pourrait également être intéressant de comparer d'autres caractéristiques que la dose d'azote, ou d'appliquer la même approche à d'autres espèces. En effet, notre modèle peut facilement s'étendre au cas linéaire ou linéaire par morceaux avec plus de deux phases.

2.2.2 Modèle Greenlab de population

Concernant le modèle Greenlab de population, plusieurs choses restent à faire pour une étude complète de cette variabilité inter-individuelle.

Bruits de modélisation

Un premier axe de développement serait d'ajouter des bruits de modélisation dans le modèle. En effet, pour l'instant nous avons uniquement considéré des bruits d'observation, or le modèle Greenlab est composé de plusieurs modules (production et allocation) qui peuvent également être entachés d'incertitude. Supposons par exemple que l'une des composantes du modèle, que l'on note $z_{i,j}$, n'est plus déterministe, mais aléatoire, de moyenne $h(t_{ij}, \phi_i)$ et de variance σ_z^2. L'étape d'observation reste inchangée, mais l'étape de variabilité intra-individuelle peut se ré-écrire de la façon suivante :

$$\begin{aligned} y_{ij} &= g(t_{ij}, \phi_i, z_{i,j}) + \varepsilon_{ij}, \quad & \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \\ z_{i,j} &= h(t_{ij}, \phi_i) + \zeta_{i,j}, & \zeta_{i,j} &\sim \mathcal{N}(0, \sigma_z^2) \end{aligned} \tag{4.45}$$

Les mêmes méthodes d'estimation peuvent être utilisées, avec $x = (y, \phi, z)$ les données complètes, y les données observées, et $\psi = (\phi, z)$ les données manquantes ou non observées. La prise en compte de bruits de modélisation a donc ajouté une dimension à l'espace des données manquantes, ce qui va avoir une incidence à l'étape E de l'algorithme, lors de l'application des algorithmes de Metropolis-Hastings ou de Gibbs, et également à l'étape M, puisqu'on a ajouté un paramètre à estimer. En particulier, dans ce cas, l'échantillonneur hybride de Gibbs s'avère plus adapté, puisqu'il peut être plus difficile d'utiliser une loi multidimensionnelle pour le vecteur $(z, \phi)^t$ (voir plus bas dans le cas où le bruit de modélisation porte sur la production de biomasse).

La vraisemblance complète s'écrit alors (voir équation 2.5 dans le cas où l'on ne considère pas de bruit de modélisation) :

$$\begin{aligned} f(y;\theta) = \int_{\mathbb{R}^{sP+N^*}} f(y,\phi,z;\theta)d\psi &= \int_{\mathbb{R}^{sP+N^*}} f(y \mid \phi,z;\theta)f(\phi,z;\theta)d\psi \\ &= \int_{\mathbb{R}^{sP+N^*}} f(y \mid \phi,z;\theta)f(z \mid \phi;\theta)f(\phi;\theta)d\psi, \end{aligned} \tag{4.46}$$

où $f(y \mid \phi, z; \theta)$ et $f(z \mid \phi; \theta)$ sont définies par (4.45), et où N^* est la dimension de z. Le vecteur de paramètres peut alors se décomposer en trois : $\theta_1 = (\beta, \Gamma)$ et $\theta_2 = \sigma^2$ restant inchangés, et $\theta_3 = \sigma_z^2$, et la fonction Q est donnée par :

$$\begin{aligned} Q(\theta;\theta^{(k)}) &= \mathbb{E}\left(\log f(\phi;\theta_1) \mid y;\theta^{(k)}\right) + \mathbb{E}\left(\log f(y \mid \phi,z;\theta_2) \mid y;\theta^{(k)}\right) + \mathbb{E}\left(\log f(z \mid \phi;\theta_3) \mid y;\theta^{(k)}\right) \\ &= Q_1(\theta_1;\theta^{(k)}) + Q_2(\theta_2;\theta^{(k)}) + Q_3(\theta_3;\theta^{(k)}). \end{aligned}$$

À titre d'illustration, supposons que la production de biomasse au jour t est stochastique, ce qui introduit une dépendance entre la biomasse produite au jour t et la séquence de biomasses produites depuis t^s_{max} jusqu'à $t-1$ (voir Figure 4.22), où t^s_{max} représente la durée de vie maximale des feuilles de la plante (cette quantité est supposée identique pour chaque plante). Une telle formulation a été proposée dans le cas du modèle Greenlab discret sans effets aléatoires par Trevezas et Cournède (2013) (voir aussi Trevezas et al. (2012)).

En utilisant la notation $q_{i,t:t'}$ pour désigner le vecteur $(q_i(t), \ldots, q_i(t'))$, où $q_i(u)$ est la biomasse produite par la plante i au jour u, on a dans le cas du modèle log-additif :

$$\begin{aligned} y_i(t_n) &= G_n(q_{i,(t_{obs}-t^s_{max})^+:t_{obs}-1};\phi_i) \circ e^{\varepsilon_{i,n}}, \\ q_i(t) &= F\left(q_{i,(t-t^s_{max})^+:t-1};t;\phi_i\right) e^{\omega_t} \end{aligned}$$

où $(x)^+ := \max(x, 0)$, F correspond à la fonction de production de biomasse, $\omega_t \sim \mathcal{N}(0, \sigma_q^2)$ et $\varepsilon_{i,n} \sim \mathcal{N}_{d_n}(0, \Sigma)$, et où l'on a spécifié la dépendance de F et G_n en la séquence de biomasses $(q_i((t_{obs} - t^s_{max})^+), \ldots, q_i(t_{obs}-1))$. Le modèle ainsi décrit appartient à la classe des *modèles espace-état* ou *modèles de Markov cachés*, où la séquence d'états cachés est $\{q_i(t)\}_{t=0,\ldots,t_{obs}-1}$, et où la séquence d'observations est $\{y_i(t_1), \ldots, y_i(t_{n_i})\}$. Les deux étapes de l'algorithme s'écrivent alors :

Étape E : La loi cible est maintenant $f(\phi_i, q_i \mid y_i; \theta^k)$, et la configuration des données manquantes nous suggère d'utiliser l'échantillonneur de Gibbs :

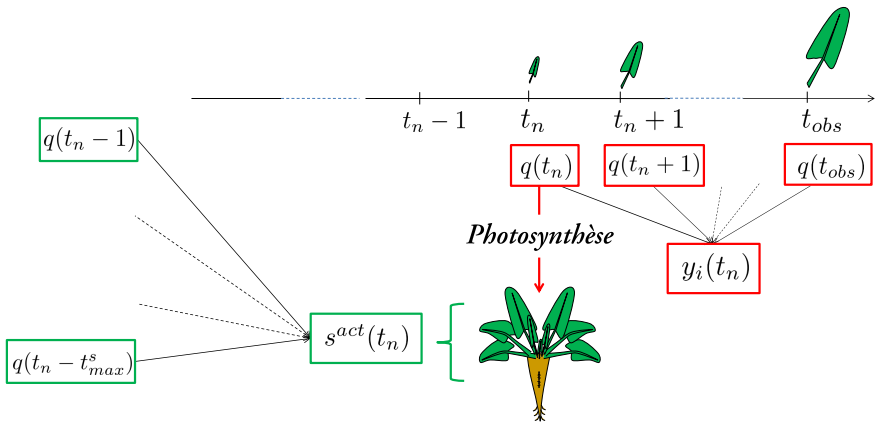


FIG. 4.22 – La biomasse $y_i(t_n)$ des organes de rang n (le limbe et le pétiole de la feuille de rang n, pour $n > 1$) de la plante i est créée au temps t_n. Elle dépend à la fois des biomasses qui seront produites chaque jour depuis l'initiation de la feuille n jusqu'à la fin de son expansion (censurée éventuellement par la date t_{obs} si la feuille est encore en expansion au moment de l'observation du système), et des biomasses produites depuis $t_n - 1$ jusqu'à t^s_{max} à travers la surface photosynthétiquement active.

On initialise pour chaque plante i, $\psi_i^{(0)} = (\phi_i^{(0)}, q_i^{(0)})$, puis pour $m = 1, \ldots, M$:

1. on génère $\phi_i^{k,(m+1)} \sim f(\cdot \mid q_i^{k,(m)}, y; \theta^k)$.

   Pour cela, on peut utiliser un algorithme de Metropolis-Hastings pour simuler le vecteur $\phi_i^{k,(m+1)}$, ou l'échantillonneur de Gibbs hybride pour générer chaque composante $\phi_{i,j}^{k,(m+1)}$.

2. on génère $q_i^{k,(m+1)} \sim f(\cdot \mid \phi_i^{k,(m+1)}, y_i; \theta^k)$

Étape M : La maximisation par rapport à θ_1 et θ_2 reste inchangée, et la maximisation par rapport à θ_3 donne :

$$\hat{\sigma}^2_Q = \frac{1}{s\,(t_{obs} - 1)} \sum_{i=1}^{s} \sum_{t=1}^{t_{obs}-1} \mathbb{E}_{\theta^k} \left[\left(q_i(t) - F(q_{i,(t-t^s_{max})^+:t-1}; t; \phi_i) \right)^2 \,\middle|\, y_i \right]$$

L'ajout de bruits de modélisation peut permettre de mieux prendre en compte et identifier les différentes sources d'incertitude et de variabilité intervenant dans le modèle. Nous avons pris ici l'exemple de la production, mais il est bien sûr possible d'ajouter des bruits de modélisation sur le processus d'allocation, même si la complexité du modèle Greenlab doit nous pousser à être prudent dans ce dernier cas : la fonction d'allocation étant définie pour chaque organe, le nombre de paramètres à estimer et la complexité de l'étape E peuvent s'avérer trop élevés. En effet, à chaque étape E de l'algorithme il faut simuler chacun des N^* états cachés, ce qui peut induire un temps de calcul rédhibitoire.

Une autre difficulté que l'on pourrait rencontrer est liée à l'identifiabilité des différentes sources de variabilité. En effet, l'introduction d'effets aléatoires dans le modèle permet en général de capturer une grande part de la variabilité de l'échantillon, et la distinction entre les deux autres sources de variabilité pourrait s'avérer délicate.

Effets fixes

Jusqu'à présent, nous avons fait l'hypothèse restrictive que chaque effet était aléatoire, c'est-à-dire avec une variance non nulle. Ceci nous a permis notamment d'obtenir une vraisemblance complète appartenant

à la famille exponentielle, et d'écrire l'étape de maximisation explicitement. Cependant, il pourrait être intéressant de considérer que certains effets comme fixes, c'est-à-dire avec une variance nulle. Dans ce cas le modèle n'appartient plus à la famille exponentielle, et l'étape de maximisation peut ne pas être explicite pour les paramètres considérés, auquel cas des méthodes d'optimisation de type quasi-Newton peuvent être envisagées.

Une autre méthode, suggérée dans le guide d'utilisateur du logiciel Monolix (The Monolix Team, 2011), pourrait être de considérer l'effet en question comme aléatoire, mais avec une variance décroissant vers 0. Pendant un nombre d'itérations fixé K_0, l'algorithme se déroule classiquement, en ajoutant une contrainte supplémentaire de non corrélation entre l'effet considéré comme fixe et le reste des paramètres, puis à partir de l'itération $K_0 + 1$, la variance correspondante n'est plus estimée, et est forcée à diminuer à chaque itération suivante, pour atteindre un seuil préalablement fixé. Cette méthode pourrait être facilement implémentée.

Une fois ces méthodes implémentées, il sera possible de tester, comme dans le modèle d'organogenèse, la significativité des effets aléatoires, en testant si la variance est nulle à l'aide d'un test du rapport de vraisemblance. Ceci permettrait d'identifier les paramètres dont la variabilité inter-individuelle est significative.

Une approche qu'il pourrait également être intéressant d'implémenter, dans le but de tester la significativité des variances des effets aléatoires, est la méthode RE-ML, pour Restricted Maximum Likelihood (voir Foulley et al. (2000)). En effet, alors que la méthode du maximum de vraisemblance fournit des estimations biaisées pour les paramètres de variance, Meza et al. (2007) ont montré que l'approche RE-ML permettait de réduire ce biais, en maximisant non pas la vraisemblance du modèle, mais la vraisemblance restreinte, obtenue en intégrant la vraisemblance des observations par rapport aux effets fixes.

Annexe A

Paramètres des modèles du Chapitre 1

Le tableau ci-dessous résume les différents paramètres utilisés dans les équations des cinq modèles du chapitre 1, ainsi que les unités correspondantes et leur interprétation biologique ou mathématique.

Modèle	Paramètre	Unité	Signification
Tous	RUE	$g \,.\, MJ^{-1}$	Efficience de conversion (Radiation Use Efficiency)
	PAR	$MJ \,.\, m^{-2}$	Radiation photosynthétiquement active
	k_B	-	Coefficient d'extinction de la loi de Beer-Lambert
Greenlab	μ	$g \,.\, MJ^{-1}.pl^{-1}$	Efficience individuelle de conversion
	s^{pr}	$m^2.pl^{-1}$	Paramètre relié à la projection orthogonale de la surface occupée par la plante, sur le sol
	e_b	$g \,.\, m^{-2}$	Masse surfacique des limbes
	p_o	-	Force de puits de l'organe o
	q_p	-	Correction de la force de puits des pétioles en fonction de la compétition pour la lumière entre limbes et pétioles
	a_o, b_o	-	Paramètres de la loi beta pour l'organe o
LNAS	e_g	$g \,.\, m^{-2}$	Masse surfacique des feuilles
	μ_a	-	Médiane de la loi log-normale pour le processus d'allocation
	σ_a	-	Écart-type de la loi log-normale pour le processus d'allocation
	μ_s	-	Médiane de la loi log-normale pour le processus de sénescence
	σ_s	-	Écart-type de la loi log-normale pour le processus de sénescence
	γ_0	-	Proportion initiale de biomasse allouée aux feuilles
	γ_f	-	Proportion finale de biomasse allouée aux feuilles
PILOTE	α, β	-	Paramètres empiriques pour la courbe du LAI
	τ_{max}	°C jour	Temps thermique correspondant à un LAI maximal
	LAI_{max}	-	Valeur maximale du LAI
	τ_e	°C jour	Temps thermique d'émergence
STICS	α, β	-	Paramètres empiriques pour la courbe du LAI
	u_{mat}	-	Unité de développement foliaire au point d'inflexion de la courbe logistique du LAI

Les valeurs des paramètres considérés comme fixes lors de la calibration des modèles du Chapitre 1 sont résumées dans le tableau ci-dessous.

Tab. A.1 – Valeurs des paramètres considérés comme fixes.

Paramètre	Valeur	Paramètre	Valeur
LNAS		**STICS**	
τ_{init}	103 °Cj	ADENS	-0.47
τ_{sen}	644 °Cj	BDENS	7 pl.m^{-2}
q_0	0.003 g	AFPFP	50
k_B	0.7	BFPFP	1
γ_0	80 %	COEFBG	0 (0.00815 pour la version initiale)
γ_f	10 %	LAICOMP	75
Greenlab		DURVIEFV	400 Q^{10}
e	83 g.m^{-2}	DURVIEIP	100 Q^{10}
k_B	0.7	DUREEFRUIT	2000 °Cj
q_0	0.003 g	DURVIESUPMAX	1
τ_{init}	103 °Cj	EXTIN	0.7 (0.58 pour la version initiale)
γ_1	36.1 °Cj	TCMAX	30°
τ_{rupt}	1393 °Cj	TCMIN	0°
γ_2	74.1 °Cj	TDMAX	25°
PILOTE		TDMIN	20°
k_B	0.7	TCXSTOP	35°
τ_e	139 °Cj	PARSURRGCC	0.48
HI	70 %	RATIOSEN	0.87
CERES		SLAMAX	120 cm^2.g^{-1}
k_B	0.7	SLAMIN	30 cm^2.g^{-1}
γ_1	36.1 °Cj	SPLAIMAX	1
γ_2	74.1	SPLAIMIN	0.0057
τ_{init}	103	STAMFLAXV	5000 °Cj
τ_{rupt}	1393 °Cj	STLEVAMFV	500°Cj
HI	70 %	STLEVDRPV	500 °Cj
		STRESSDEV	1
		LAICOMP	75
		TIGEFEUILLE	1.41
		UDLAIMAX	3
		VLAIMAX	2.2
		ILEV	13

Annexe B

Calcul de la matrice d'information de Fisher

Nous présentons ici les détails du calcul permettant de relier la matrice d'information de Fisher des données observées, ou *incomplètes* et celle des données complètes, dans le cadre du chapitre 3. On suppose pour des raisons de simplicité que $\theta \in \mathbb{R}$, mais le résultats s'étendent facilement au cas $\theta \in \mathbb{R}^P$. On note y les données incomplètes, et $x = (y, \phi)$ les données complètes.

Tout d'abord, on a $S(y;\theta) = \mathbb{E}_\theta\left[S(x;\theta) \mid y\right]$. En effet,

$$\begin{aligned}
\mathbb{E}_\theta\left[S(x;\theta) \mid y\right] &= \mathbb{E}_\theta\left[S(\phi \mid y;\theta) \mid y\right] + \mathbb{E}_\theta\left[S(y;\theta) \mid y\right] \\
&= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(\phi \mid y;\theta) \mid y\right] + S(y;\theta) \\
&= \int \left(\frac{\partial}{\partial\theta}\log f(\phi \mid y;\theta)\right) f(\phi \mid y;\theta)\, d\phi + S(y;\theta) \\
&= \int \frac{\partial}{\partial\theta}\, f(\phi \mid y;\theta)\, d\phi + S(y;\theta) \\
&= \frac{\partial}{\partial\theta}\left(\int f(\phi \mid y;\theta)\, d\phi\right) + S(y;\theta) \\
&= S(y;\theta),
\end{aligned}$$

car $f(\phi \mid y;\theta)$ définie une densité de probabilité dont l'intégrale vaut 1. Puis on a :

$$\begin{aligned}
\mathbb{E}_\theta\left[S(x;\theta)\right] &= \int \frac{\partial}{\partial\theta}\log f(x;\theta)\; f(x;\theta)\, dx \\
&= \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}\, f(x;\theta)\, dx \\
&= \frac{\partial}{\partial\theta}\left(\int f(x;\theta)\, dx\right) \\
&= 0,
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right] &= \mathbb{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)\; \frac{\partial}{\partial\theta} f(x;\theta) - (\frac{\partial}{\partial\theta} f(x;\theta))^2}{f(x;\theta)^2}\right] \\
&= \mathbb{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2} f(x;\theta)}{f(x;\theta)}\right] - \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] \\
&= \int \frac{\partial^2}{\partial\theta^2}\, f(x;\theta)\, dx - \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] \\
&= -\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right]
\end{aligned}$$

Puis, on a :

$$\begin{aligned}
\frac{\partial}{\partial\theta}\, S(y;\theta) &= \frac{\partial}{\partial\theta}\, \mathbb{E}_\theta\left[S(x;\theta) \mid y\right] \\
&= \frac{\partial}{\partial\theta}\left(\int \frac{\partial}{\partial\theta} \log f(x;\theta) f(\phi \mid y;\theta)\, d\phi\right) \\
&= \int \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta} \log f(x;\theta) f(\phi \mid y;\theta)\right] d\phi \\
&= \int \frac{\partial^2}{\partial\theta^2} \log f(x;\theta) f(\phi \mid y;\theta)\, d\phi + \underbrace{\int \frac{\partial}{\partial\theta} \log f(x;\theta) \frac{\partial}{\partial\theta} f(\phi \mid y;\theta)\, d\phi}_{:=A} \\
&= \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f(x;\theta) \mid y\right] + A
\end{aligned}$$

L'intégrale A peut ensuite être calculée de deux manières différentes, en utilisant la décomposition $f(x;\theta) = f(\phi \mid y;\theta) f(y;\theta)$ sur le premier ou le deuxième terme de l'intégrande :

$$\begin{aligned}
A &= \int \frac{\partial}{\partial\theta} \log f(x;\theta)\, \frac{\partial}{\partial\theta} \frac{f(x;\theta)}{f(y;\theta)}\, d\phi \\
&= \int \frac{\partial}{\partial\theta} \log f(x;\theta)\, \frac{\frac{\partial}{\partial\theta} f(x;\theta) f(y;\theta) - f(x;\theta) \frac{\partial}{\partial\theta} f(y;\theta)}{f^2(y;\theta)}\, d\phi \\
&= \int \frac{\partial}{\partial\theta} \log f(x;\theta) \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(y;\theta)}\, d\phi - \int \frac{\partial}{\partial\theta} \log f(x;\theta) \frac{f(x;\theta) \frac{\partial}{\partial\theta} f(y;\theta)}{f^2(y;\theta)}\, d\phi \\
&= \int \left(\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}\right)^2 \frac{f(x;\theta)}{f(y;\theta)}\, d\phi - \int \frac{\partial}{\partial\theta} \log f(x;\theta) \frac{\partial}{\partial\theta} \log f(y;\theta) \frac{f(x;\theta)}{f(y;\theta)}\, d\phi \\
&= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2 \mid y\right] - \mathbb{E}_\theta\left[S(x;\theta) S(y;\theta) \mid y\right] \\
&= \mathbb{E}_\theta\left[\left(S(x;\theta)\right)^2 \mid y\right] - S(y;\theta) \mathbb{E}_\theta\left[S(x;\theta) \mid y\right] \\
&= \mathbb{E}_\theta\left[\left(S(x;\theta)\right)^2 \mid y\right] - \mathbb{E}_\theta\left[S(x;\theta) \mid y\right]^2 \\
&= \mathrm{Var}_\theta\left(S(x;\theta) \mid y\right).
\end{aligned}$$

Finalement :

$$\begin{aligned}
A &= \int \frac{\partial}{\partial\theta} (\log f(\phi \mid y;\theta) + \log f(y;\theta)) \frac{\partial}{\partial\theta} f(\phi \mid y;\theta)\, d\phi \\
&= \int \frac{\partial}{\partial\theta} \log f(\phi \mid y;\theta) \frac{\partial}{\partial\theta} f(\phi \mid y;\theta)\, d\phi + \int \frac{\partial}{\partial\theta} \log f(y;\theta) \frac{\partial}{\partial\theta} f(\phi \mid y;\theta)\, d\phi \\
&= \int \left(\frac{\frac{\partial}{\partial\theta} f(\phi \mid y;\theta)}{f(\phi \mid y;\theta)}\right)^2 f(\phi \mid y;\theta)\, d\phi + \frac{\partial}{\partial\theta} \log f(y;\theta) \int \frac{\partial}{\partial\theta} f(\phi \mid y;\theta)\, d\phi \\
&= \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \log f(\phi \mid y;\theta)\right)^2 \mid y\right] \\
&= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f(\phi \mid y;\theta) \mid y\right]
\end{aligned}$$

Glossaire

capacité au champ : quantité maximale d'eau que peut contenir le sol. Elle correspond à la quantité restante dans un sol, 24 ou 48 heures après qu'il ait été saturé d'eau, par les précipitations ou l'irrigation.

chloroplaste : organite situé dans les cellules végétales et contenant les pigments de chlorophylle. Les chloroplastes sont le siège de la photosynthèse.

cotylédons : première feuille ou première paire de feuilles qui sont contenues dans l'embryon de la plante (dans la graine). On distingue les plantes monocotylédones et dicotylédones. Le(s) cotylédon(s) est(sont) en général très différents des autres feuilles de la plante (en terme de forme notamment). Ils peuvent rester enfouis sous terre, ou au contraire sortir et participer ainsi à la photosynthèse.

croissance déterminée : les plantes à croissance déterminée sont des plantes dont le méristème apical finit par cesser de croître, la plante restant alors à l'état végétatif ou périclitant, par opposition aux plantes à **croissance indéterminée**.

croissance indéterminée : les plantes à croissance indéterminée sont des plantes dont le méristème apical continue de croître et de se différencier tant que vit la plante.

indice de surface foliaire (Leaf Area Index, LAI) : ratio entre la surface totale supérieure des feuilles vertes et la surface de sol sur laquelle se développe la culture (Watson, 1947). C'est une grandeur sans dimension, qui varie de 0 pour un sol nu à environ 8 pour une forêt tempérée. Dans certaines forêts tropicales denses ou de conifères, cet indice peut même dépasser 15 (Schulze, 1982).

méristème : tissu formé de cellules indifférenciées (embryonnaires) à multiplication rapide, qui permettent la croissance.

métamère : (ou phytomère) unité botanique élémentaire de la plante, constituée d'un nœud, d'une ou plusieurs feuilles insérées à ce nœud, et de l'entrenœud sous-jacent. Le mot métamère s'emploie dans un cadre général (règnes animal et végétal), alors que le mot phytomère est spécifique au monde végétal.

organogenèse : création de nouveaux organes par la plante.

phyllochrone : temps thermique séparant l'apparition de deux feuilles successives.

point de flétrissement : teneur en eau du sol en-deçà de laquelle la plante ne peut plus puiser d'eau, flétrit et meurt, car la tension capillaire des racines n'est pas suffisante pour absorber l'eau du sol.

stomate : orifice présent habituellement sur la face inférieure des feuilles et permettant les échanges gazeux entre la plante et l'atmosphère.

Publications

Les travaux présentés dans ce manuscrit ont fait l'objet de plusieurs publications :

Revues à comités de lecture

- C. Baey, S. Trevezas, et P.-H. Cournède. A nonlinear mixed model to explain variability in plant populations : maximum likelihood estimation via stochastic variants of the EM algorithm. Soumis à *Communications in Statistics - Theory and Methods* (2013).
- C. Baey, A. Didier, S. Lemaire, F. Maupas, et P.-H. Cournède. Parametrization of five classical plant growth models applied to sugar beet and comparison of their predictive capacities on root yield and total biomass. Accepté pour publication dans *Ecological Modelling* (2014) (disponible en ligne).
- C. Baey, A. Didier, S. Lemaire, F. Maupas, et P.-H. Cournède. Modelling the interindividual variability of organogenesis in sugar beet populations using a hierarchical segmented model. *Ecological Modelling*, 263 :56–63, 2013.

Conférences avec actes

- C. Baey, S. Trevezas, et P.-H. Cournède. A nonlinear mixed effects model to explain inter-individual variability in plant populations, *15th Applied Stochastic Models and Data Analysis International Conference* - Barcelone, Espagne, Juin 2013.
- C. Baey, A. Didier, S. Lemaire, F. Maupas, et P.-H. Cournède. Evaluation of the predictive capacity of five plant growth models for sugar beet, *4th International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications (PMA)* - Shanghai, Chine, Novembre 2012, IEEE Press, pp 30-37.
- C. Baey, A. Didier, S. Lemaire, F. Maupas, et P.-H. Cournède. Using a hierarchical segmented model to assess the dynamics of leaf appearance in plant populations, *14th Applied Stochastic Models and Data Analysis International Conference* - Rome, Italie, Juin 2011.

Bibliographie

S. Allassonnière, E. Kuhn et A. Trouvé : Construction of Bayesian Deformable Models via Stochastic Approximation Algorithm : A Convergence Study. *Bernouilli*, 16(3) :641–678, 2010.

B. Andrieu, J.-M. Allirand et K. W. Jaggard : Ground cover and leaf area index of maize and sugar beet crops. *Agronomie*, 17 :315–321, 1997.

C. Andrieu, E. Moulines et P. Priouret : Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 05 :1–29, 2005.

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

D. Barthélémy et Y. Caraglio : Plant architecture : a dynamic, multilevel and comprehensive approach to plant form, structure and ontogeny. *Annals of Botany*, 99(3) :375–407, 2007.

A. Bauer, A. B. Frank et A. L. Black : Estimation of spring wheat grain dry matter assimilation from air temperature. *Agronomy journal*, 77(5) :743–752, 1985.

S. L. Beal et L. B. Sheiner : Estimating population kinetics. *Critical Reviews in Biomedical Engineering*, 8 (3) :195–222, 1982.

M. Bédard : Weak convergence of metropolis algorithms for non-iid target distributions. *The Annals of Applied Probability*, 17(4) :1222–1244, 2007.

M. Bédard : Optimal acceptance rates for metropolis algorithms : Moving beyond 0.234. *Stochastic Processes and their Applications*, 118(12) :2198–2222, 2008.

J. Bertheloot, B. Andrieu, C. Fournier et P. Martre : A process-based model to simulate nitrogen distribution in wheat (Triticum aestivum) during grain-filling. *Functional Plant Biology*, 35(10) :781, 2008.

J. Bertheloot, Q. Wu, P.-H. Cournède et B. Andrieu : NEMA, a functional–structural model of nitrogen economy within wheat culms after flowering. II. Evaluation and sensitivity analysis. *Annals of Botany*, 108(6) :1097–1109, 2011.

P. Billingsley : *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 4th édn, 2012.

P. V. Biscoe et J. N. Gallagher : Weather, dry matter production and yield. *In* J. Landsberg et C. V. Cutting, éds : *Environmental effects on crop physiology*. Academic Press, 1977.

B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens et J.-S. S. White : Generalized linear mixed models : a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 24(3) :127–135, 2009.

J. G. J. Booth et J. P. J. Hobert : Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society : Series B*, 61(1) :265–285, 1999.

J. G. Booth, J. P. Hobert et W. Jank : A survey of monte carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling*, 1(4) :333–349, 2001.

R. A. Boyles : On the convergence of the EM algorithm. *Journal of the Royal Statistical Society : Series B*, 45 :47–50, 1983.

P. Bratley, B. Fox et L. Schrage : *A Guide*. Springer-Verlag, 1987.

N. E. Breslow et D. G. Clayton : Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421) :9–25, 1993.

N. Brisson, C. Gary, E. Justes, R. Roche, B. Mary, D. Ripoche, D. Zimmer, J. Sierra, P. Bertuzzi et P. Burger : An overview of the crop model STICS. *European Journal of agronomy*, 18 :309–332, 2003.

N. Brisson, M. Launay, B. Mary et N. Beaudoin : *Conceptual Basis, Formalisations and Parameterization of the Stics Crop Model*. Quae, 2008.

N. Brisson, B. Mary, D. Ripoche, M.-H. Jeuffroy, F. Ruget, B. Nicoullaud, P. Gate, F. Devienne-Barret, R. Antonioletti, C. Durr, G. Richard, N. Beaudoin, S. Recous, X. Tayot, D. Plenet, P. Cellier, J.-M. Machet, J.-M. Meynard et R. Delécolle : STICS : a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parametrization applied to wheat and corn. *Agronomie*, 18 :311–346, 1998.

J. Brouwer, L. K. Fussell et L. Herrmann : Soil and crop growth micro-variability in the west african semi-arid tropics : a possible risk-reducing factor for subsistence farmers. *Agriculture, Ecosystems and Environment*, 45(3-4) :229 – 238, 1993.

L. Brown : *Fundamentals of Statistical Exponential Families : With Applications in Statistical Decision Theory*. IMS Lecture Notes. Institute of Mathematical Statistics, 1986.

K. Burnham et D. Anderson : *Model selection and multimodel inference : a practical information-theoretic approach*. Springer Verlag, 2nd édn, 2002.

B. S. Caffo, W. Jank et G. L. Jones : Ascent-based Monte Carlo expectation- maximization. *Journal of the Royal Statistical Society : Series B*, 67(2) :235–251, 2005.

W. Cao et D. N. Moss : Temperature and Daylength Interaction on Phyllochron in Wheat and Barley. *Crop Science*, 29(4) :1046, 1989.

O. Cappé, E. Moulines et T. Rydén : *Inference in hidden Markov models*. Springer, 2005.

J. Cariboni, D. Gatelli, R. Liska et A. Saltelli : The role of sensitivity analysis in ecological modelling. *Ecological Modelling*, 203(1-2) :167–182, 2007.

M. Casagrande, C. David, M. Valantin-Morison, D. Makowski et M.-H. Jeuffroy : Factors limiting the grain protein content of organic winter wheat in south-eastern France : a mixed-model approach. *Agronomy for Sustainable Development*, 29(4) :565–574, 2009.

Y. Chen, S. Trevezas et P.-H. Cournède : A regularized particle filter EM algorithm based on Gaussian randomization with an application to plant growth modeling. *Submitted*, 2013a.

Y. Chen, S. Trevezas et P.-H. Cournède : Iterative convolution particle filtering for nonlinear parameter estimation and data assimilation with application to crop yield prediction. *In Proceeding of Society for Industrial and Applied Mathematics (SIAM) Control & its Applications*, San Diego, États-Unis, Juillet 2013b.

B. Clerget, M. Dingkuhn, E. Gozé, H. F. W. Rattunde et B. Ney : Variability of phyllochron, plastochron and rate of increase in height in photoperiod-sensitive sorghum varieties. *Annals of Botany*, 101(4) : 579–94, 2008.

N. Colbach, N. Molinari et C. Clermont-Dauphin : Sensitivity analyses for a model simulating demography and genotype evolutions with time : Application to genesys modelling gene flow between rape seed varieties and volunteers. *Ecological Modelling*, 179(1) :91–113, 2004.

E. Comets, K. Brendel et F. Mentré : Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models : the npde add-on package for R. *Computer Methods and Programs in Biomedicine*, 90(2) :154–166, 2008.

E. Comets, C. Verstuyft, M. Lavielle, P. Jaillon, L. Becquemont et F. Mentré : Modelling the influence of MDR1 polymorphism on digoxin pharmacokinetic parameters. *European Journal of Clinical Pharmacology*, 63(5) :437–449, mai 2007.

B. Courbaud, G. Vieilledent et G. Kunstler : Intra-specific variability and the competition-colonisation trade-off : coexistence, abundance and stability patterns. *Theoretical Ecology*, 5(1) :61–71, 2012.

P.-H. Cournède, V. Letort, A. Mathieu, M.-Z. Kang, S. Lemaire, S. Trevezas, F. Houllier et P. de Reffye : Some Parameter Estimation Issues in Functional-Structural Plant Modelling. *Mathematical Modelling of Natural Phenomena*, 6(2) :133–159, 2011.

P.-H. Cournède, A. Mathieu, F. Houllier, D. Barthélémy et P. de Reffye : Computing competition for light in the Greenlab model of plant growth : a contribution to the study of the effects of density on resource acquisition and architectural development. *Annals of Botany*, 101(8) :1207–1219, 2008.

P.-H. Cournède, Y. Chen, Q. Wu, C. Baey et B. Bayol : Development and Evaluation of Plant Growth Models : Methodology and Implementation in the PyGMAlion platform. *Mathematical Modelling of Natural Phenomena*, 8(4) :112–130, 2013.

P.-H. Cournède, M.-Z. Kang, A. Mathieu, J.-F. Barczi, H.-P. Yan, B.-G. Hu et P. de Reffye : Structural factorization of plants to compute their functional and architectural growth. *Simulation*, 82(7) :427–438, 2006.

R. I. Cukier, C. M. Fortuin, K. E. Shuler, A. G. Petschek et J. H. Schaibly : Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. *Journal of Chemical Physics*, 1973.

N. Damay et J. Le Gouis : Radiation use efficiency of sugar beet in northern France. *European Journal of Agronomy*, 2 :179–184, 1993.

M. Davidian et D. M. Giltinan : *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall/CRC, 1ère édn, 1995.

M. Davidian et A. R. Gallant : The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80(3) :475 –488, 1993.

M. Davidian et D. M. Giltinan : Some simple methods for estimating intraindividual variability in nonlinear mixed effects models. *Biometrics*, p. 59–73, 1993.

P. de Reffye et B.-G. Hu : Relevant qualitative and quantitative choices for building an efficient dynamic plant growth model : GreenLab case. *In* B. G. Hu et M. Jaeger, éds : *Proceedings of the First International Symposium on Plant Growth Models and Applications (PMA03)*, p. 87–107. Tsinghua University Press and Springer, 2003.

P. de Reffye, S. Lemaire, N. Srivastava, F. Maupas et P.-H. Cournède : Modeling inter-individual variability in sugar beet populations. *In* B. Li, M. Jaeger et Y. Guo, éds : *Proceedings of the Third International Symposium on Plant Growth Models and Applications (PMA03)*, Beijing, China, 9-12 Novembre 2009a. IEEE Computer Society (Los Alamitos, California).

P. de Reffye, F. Blaise, S. Chemouny et S. Jaffuel : Calibration of hydraulic growth model on the architecture of cotton plants. *Agronomie*, 19 :265–280, 1999.

P. de Reffye et P. Dinouard : Basic concepts of computer plants growth simulation. *In Proceedings of NICOGRAPH*, p. 219–234, Tokyo, 1990.

P. de Reffye, C. Edelin, J. Françon, M. Jaeger et C. Puech : Plant models faithful to botanical structure and developmentr. *In Proceedings of SIGGRAPH '88*, p. 151–158, 1988.

P. de Reffye, E. Elguero et E. Costes : Growth units construction in trees : a stochastic approach. *Acta Biotheoretica*, 39(3) :325–342, 1991.

P. de Reffye, T. Fourcaud, F. Blaise, D. Barthélémy, F. Houllier et al. : A functional model of tree growth and tree architecture. *Silva Fennica*, 31(3) :297–311, 1997.

P. de Reffye, M. Jaeger et P.-H. Cournède : Une histoire de la modélisation des plantes. *Interstices*, 2009b.

C. de Wit, R. Brouwer, F. de Vries et I. Setlik : The simulation of photosynthetic systems. *In Prediction and measurement of photosynthetic productivity. Proceedings of the IBP/PP Technical Meeting*, p. 47–70, Tfebon, Tchécoslovaquie, Septembre 1969 1970. Wageningen, Netherlands, PUDO.

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

A. Dempster, N. M. Laird et D. Rubin : Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B*, 39(1) :1–38, 1977.

A. Didier : *Modélisation de la croissance, des relations sources-puits et du rendement en sucre de la betterave sucrière (Beta vulgaris L.) sous des régimes contrastés de nutrition azotée*. Thèse de doctorat, AgroParisTech, 2013.

M. C. Dietze, M. S. Wolosin et J. S. Clark : Capturing diversity and interspecific variability in allometries : A hierarchical approach. *Forest Ecology and Management*, 256(11) :1939 – 1948, 2008.

Q. Dong, G. Louarn, Y. Wang, J. Barczi et P. de Reffye : Does the structure-function model GreenLab deal with crop phenotypic plasticity induced by plant spacing ? A case study on tomato. *Annals of Botany*, 101(8), 2008.

A. P. Draycott et D. R. Christensen : *Nutrients for sugar beet production : Soil-plant relationships*. CABI Publishing, 2003.

A. Draycott : *Sugar Beet*. World Agriculture Series. Wiley, 2008.

C. Dürr, J. N. Aubertot, G. Richard, P. Dubrulle, D. Y et B. J : SIMPLE : a model for simulation of plant emergence predicting the effects of soil tillage and sowing operations. *Soil Science Society American Journal*, 65 :414–423, 2001.

C. Dürr et J. Boiffin : Sugarbeet seedling growth from germination to first leaf stage. *The Journal of Agricultural Science*, 124 :427–435, 6 1995. ISSN 1469-5146.

C. Dürr, K. Fares, N. Damay, A. Carrera, N. Beaudoin, J.-M. Machet, R. Duval, F. Maupas et P. Postel : A description of the development and growth of sugar beet for crop modelling. *Advances in Sugar Beet Research - Sugar Beet Growth and Growth Modelling*, 5 :71–85, 2003.

M. Duval : *Modélisation et estimation de variances hétérogènes dans les modèles non linéaires mixtes*. Thèse de doctorat, AgroParisTech, 2008.

B. Efron et D. V. Hinkley : Assessing the accuracy of the maximum likelihood estimator : Observed versus Expected Fisher information. *Biometrika*, 65(3) :457 – 487, 1978.

P. Federl et P. Prusinkiewicz : Virtual laboratory : an interactive software environment for computer graphics. *In Proceedings of Computer Graphics International CGI-99*, p. 93–100. IEEE, 1999.

L. Feng, J.-c. Mailhol, H. Rey, S. Griffon, D. Auclair et P. D. Reffye : Comparing an empirical crop model with a functional structural plant model to account for individual variability. *European Journal of Agronomy*, 53 :16–27, 2014.

J. M. Flegal et G. L. Jones : Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2) :1034–1070, avr. 2010.

A. L. Fletcher, E. Chakwizira, S. Maley et M. George : Canopy development and radiation use efficiency of four forage brassica crops. *In* I. Yunusa, éd. : *Proceedings of the 16th Australian Agronomy Conference*, Armidale, NSW, 2012.

F. Forcella, R. L. B. Arnold, R. Sanchez et C. M. Ghersa : Modeling seedling emergence. *Field Crops Research*, 67 :123–139, 2000.

G. Fort et E. Moulines : Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4) :1220–1259, 2003.

J.-L. Foulley, F. Jaffrézic et C. Robert-Granié : EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis. *Genetics, selection, evolution : GSE*, 32(2) :129–41, 2000.

T. Fourcaud, X. Zhang, A. Stokes, H. Lambers et C. Korner : Plant growth modelling and applications : The increasing importance of plant architecture in growth models. *Annals of Botany*, 101 (8) :1053–1063, 2008.

C. Fournier et B. Andrieu : ADEL-maize : An L-system based model for the integration of growth processes from the organ to the canopy. Application to regulation of morphogenesis by light availability. *Agronomy*, 19(3-4) :313–327, 1999.

A. Frank et A. Bauer : Phyllochron differences in wheat, barley and forage grasses. *Crop Science*, 35 (1) :19–23, 1995.

P. Garnier : Sensitivity analysis of pastis, amodel of nitrogen transport and transformation in the soil. *In* D. Wallach, D. Makowski et J. W. Jones, éds : *Working with Dynamic Crop Models*, p. 367–375. Elsevier, 2006.

C. Gaucherel, F. Campillo, L. Misson, J. Guiot et J.-J. Boreux : Parameterization of a process-based tree-growth model : comparison of optimization, mcmc and particle filtering algorithms. *Environmental Modelling & Software*, 23(10) :1280–1288, 2008.

A. E. Gelfand, S. E. Hills, A. Racine-Poon et A. F. M. Smith : Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85(412) :972–985, 1990.

A. Gelman, G. Roberts et W. Gilks : Efficient metropolis jumping hules. *Bayesian statistics*, 5 :599–608, 1996.

S. Geman et D. Geman : Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 :721–741, 1984.

G. C. Goodwin et R. L. Payne : *Dynamic system identification*. Academic Press, 1977.

M. Guérif et C. Duke : Calibration of the SUCROS emergence and early growth module for sugar beet using optical remote sensing data assimilation. *European Journal of Agronomy*, 9 :127–136, 1998.

Y. Guo, Y. Ma, Z. Zhan, B. Li, M. Dingkuhn, D. Luquet et P. de Reffye : Parameter optimization and field validation of the functional-structural model GREENLAB for maize. *Annals of Botany*, 97 :217–230, 2006.

H. Haario, E. Saksman et J. Tamminen : Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 4 :1–32, 1999.

H. Haario, E. Saksman et J. Tamminen : An Adaptive Metropolis Algorithm. *Bernouilli*, 7(2) :223–242, 2001.

D. Hall et R. Bailey : Modeling and prediction of forest growth variables based on multilevel nonlinear mixed models. *Forest Science*, 47(3) :311–321, 2001. ISSN 0015-749X.

F. Hallé et R. Oldeman : *Essai sur l'architecture et la dynamique de croissance des arbres tropicaux*. Collection de monographies de botanique et de biologie végétale. Monographie 6. Masson, 1970.

F. Hallé, R. Oldeman et P. Tomlinson : *Tropical trees and forests : an architectural analysis*. New York : Springer-Verlag, 1978.

W. Hastings : Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, 1970.

J. Hillier, D. Makowski et B. Andrieu : Maximum likelihood inference and bootstrap methods for plant organ growth via multi-phase kinetic models and their application to maize. *Annals of Botany*, 96 : 137–148, 2005.

W. Jank : Implementing and diagnosing the stochastic approximation EM algorithm. *Journal of Computational and Graphical Statistics*, 1815 :1–30, 2006.

C. A. Jones et J. Kiniry, éds. *CERES-Maize : a simulation model of maize growth and development*. Texas A&M University Press, College Station, Temple, TX, 1986.

G. L. Jones, M. Haran, B. S. Caffo et R. Neath : Fixed-width output analysis for markov chain monte carlo. *Journal of the American Statistical Association*, 101(476) :1537–1547, 2006.

A. Jullien, A. Mathieu, J.-M. Allirand, A. Pinet, P. de Reffye, P.-H. Cournède et 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.

P. Juskiw, J. Helm et J. Nyachiro : Measuring phyllochrons in barley to use for seeding rate recommendations. *In 18th North American Barley Researchers Workshop*, Juillet 2005.

M. Kang, P.-H. Cournède, P. de Reffye, D. Auclair et B. Hu : Analytical study of a stochastic plant growth model : application to the Greenlab model. *Mathematics and Computers in Simulation*, 78(1) :57–75, 2008.

R. Karwowski et P. Prusinkiewicz : Design and implementation of the L+ C modeling language. *Electronic Notes in Theoretical Computer Science*, 86(2) :134–152, 2003.

R. Karwowski et P. Prusinkiewicz : The L-system-based plant-modeling environment L-studio 4.0. *In 4th International Workshop on Functional-Structural Plant Models*, p. 403–405, 2004.

J. King : *Le monde fabuleux des plantes : Pourquoi la Terre est verte.* Bibliothèque Pour la science. Belin, 2004. Traduction française de *Reaching for the Sun : How Plants Work* par J. M. Walter.

O. Kniemeyer, G. H. Buck-Sorlin et W. Kurth : GroIMP as a platform for functional-structural modelling of plants. *In* J. Vos, L. F. M. Marcelis, P. H. B. de Visser, P. C. Struik et J. B. Evers, éds : *Functional-Structural Plant Modelling in Crop Production*, vol. Chapter 04, p. 43–52. Citeseer, 2007.

E. Korpilaht, éd. *Helsinki Workshop on Functional-Structural Tree Models*, vol. 31(3). Silva Fennica, 1997.

E. Kuhn et M. Lavielle : Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8 :115–131, 2004.

E. Kuhn et M. Lavielle : Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4) :1020–1038, 2005.

W. Kurth : Morphological models of plant growth : possibilities and ecological relevance. *Ecological Modelling*, 75 :299–308, 1994.

N. M. Laird et J. H. Ware : Random-effects models for longitudinal data. *Biometrics*, p. 963–974, 1982.

M. Launay et M. Guérif : Ability for a model to predict crop production variability at the regional scale : an evaluation for sugar beet. *Agronomie*, 23(2) :135–146, 2003.

M. Lavielle, A. Samson, A. K. Fermin et F. Mentré : Maximum Likelihood Estimation of Long Term HIV Dynamic Models and Antiviral Response. *Biometrics*, 2010.

X. Le Roux, A. Lacointe, A. Escobar-Gutiérrez et S. Le Dizès : Carbon-based models of individual tree growth : a critical appraisal. *Annals of Forest Science*, 58(5) :469–506, 2001.

G. S. Lee et W. Schmehl : Effect of planting date and nitrogen fertility on appearance and senescence of sugarbeet leaves. *Journal of Sugar Beet Research*, 25(1) :28–41, 1988.

S. Lemaire : *Système dynamique de la croissance et du développement de la betterave sucrière (Beta vulgaris L.)*. Thèse de doctorat, AgroParisTech, 2010.

S. Lemaire, F. Maupas, P.-H. Cournède et P. de Reffye : A morphogenetic crop model for sugar-beet (*Beta vulgaris L.*). *In International Symposium on Crop Modeling and Decision Support : ISCMDS 2008*, Nanjing, China, 19-22 Avril 2008.

S. Lemaire, F. Maupas, P.-H. Cournède, J.-M. Allirand, P. de Reffye et B. Ney : Analysis of the density effects on the source-sink dynamics in Sugar-Beet growth. *In* B. Li, M. Jaeger et Y. Guo, éds : *Proceedings of the Third International Symposium on Plant Growth Models and Applications (PMA03)*, Beijing, China, November 9-12 2009. IEEE Computer Society (Los Alamitos, California).

V. Letort : *Adaptation du modèle de croissance GreenLab aux plantes à architecture complexe et analyse multi-échelle des relation source-puits pour l'identification paramètrique*. Thèse de doctorat, École Centrale Paris, 2008.

B. Leviel : *Evaluation des risques et maîtrise des flux d'azote au niveau d'une parcelle agricole dans la plaine roumaine et bulgare. Application aux cultures de mais, blé, colza et betterave.* Thèse de doctorat, Institut National Polytechnique de Toulouse, 2000.

R. Levine et J. Fan : An automated (Markov chain) Monte Carlo EM algorithm. *Journal of Statistical Computation and Simulation*, 74(5) :349–360, 2004.

R. A. Levine et G. Casella : Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3) :422–439, 2001.

Z. Li, V. Le Chevalier et P.-H. Cournède : Towards a continuous approach of functional-structural plant growth. *In* B.-G. Li, M. Jaeger et Y. Guo, éds : *Proceedings of the Third International Symposium on Plant Growth Models and Applications (PMA03)*, Beijing, China, November 9-12 2009. IEEE.

A. Lindenmayer : Mathematical models for cellular interactions in development i. filaments with one-sided inputs. *Journal of theoretical biology*, 18(3) :280–299, 1968.

M. J. Lindstrom et D. M. Bates : Nonlinear Mixed Effects Models. *Biometrics*, 46 :673–687, 1990.

W. Liu, M. Tollenaar, G. Stewart et W. Deen : Response of corn grain yield to spatial and temporal variability in emergence. *Crop Science*, 44 :847–854, 2004.

T. A. Louis : Finding the Observed Information Matrix when using the EM-algorithm. *Journal of the Royal Statistical Society : Series B*, 44(2) :226–233, 1982.

Y. Ma, M. Wen, Y. Guo, B. Li, P.-H. Cournède et P. de Reffye : Parameter optimization and field validation of the functional-structural model GREENLAB for maize at different population densities. *Annals of botany*, 101(8) :1185–94, mai 2008.

J.-C. Mailhol, A. Olufayo et P. Ruelle : Sorghum and sunflower evapotranspiration and yield from simulated leaf area index. *Agricultural Water Management*, 35 :167–182, 1997.

J.-C. Mailhol, P. Revol et P. Ruelle : Pilote : un modèle opérationnel pour déceler l'apparition de stress hydrique. *In ICID 16th international congress on irrigation and drainage : workhop on crop-water-environment models*, Cairo, Egypt, juil. 1996.

D. Makowski et M. Lavielle : Using SAEM to estimate parameters of models of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1) :45–60, 2006.

D. Makowski, D. Wallach et M. Tremblay : Using a bayesian approach to parameter estimation ; comparison of the glue and mcmc methods. *Agronomie*, 22(2) :191–203, 2002.

D. Makowski, J. Hillier, D. Wallach, B. Andrieu et M.-H. Jeuffroy : Parameter estimation for crop models. *In* D. Wallach, D. Makowski et J. W. Jones, éds : *Working with Dynamic Crop Models*, p. 101–140. Elsevier, 2006.

L. Marcelis, E. Heuvelink et J. Goudriaan : Modelling of biomass production and yield of horticultural crops : a review. *Scientia Horticulturae*, 74 :83–111, 1998.

A. Marshall : The use of multi-stage sampling shcemes in Monte Carlo computations. *In* M. Meyer, éd. : *Symposium on Monte Carlo Methods*, p. 123–140. Wiley, 1956.

A. Mathieu, P.-H. Cournède, V. Letort, D. Barthélémy et P. de Reffye : A dynamic model of plant growth with interactions between development and functional mechanisms to study plant structural plasticity related to trophic competition. *Annals of Botany*, 103(8) :1173–1186, 2009.

A. Mathieu : *Essai sur la modélisation des interactions entre la croissance et le développement d'une plante - Cas du modèle GreenLab*. Thèse de doctorat, Ecole Centrale Paris, 2006.

D. Mayer et D. Butler : Statistical validation. *Ecological Modelling*, 68 :21–32, juil. 1993.

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

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

G. McLachlan et T. Krishnan : *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 2nd édn, 2007.

X.-L. Meng et D. B. Rubin : Maximum likelihood estimation via the ECM algorithm : A general framework. *Biometrika*, 80(2) :267–278, 1993.

F. Mentré et S. Escolano : Prediction discrepancies for the evaluation of nonlinear mixed-effects models. *Journal of pharmacokinetics and pharmacodynamics*, 33(3) :345–67, 2006.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller et E. Teller : Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6) :1087, 1953.

C. Meza, F. Jaffrézic et J.-L. Foulley : REML estimation of variance parameters in nonlinear mixed effects models using the SAEM algorithm. *Biometrical journal. Biometrische Zeitschrift*, 49(6) :876–88, 2007.

G. F. J. Milford, T. O. Pocock et J. Riley : An analysis of leaf growth in sugar beet. I. Leaf appearance and expansion in relation to temperature under controlled conditions. *Annals of Applied Biology*, 106 (1) :163–172, 1985a.

G. F. J. Milford, T. O. Pocock et J. Riley : An analysis of leaf growth in sugar beet. II. Leaf appearance in field crops. *Annals of Applied Biology*, 106(1) :173–185, 1985b.

G. F. J. Milford et J. Riley : The effects of temperature on leaf growth of sugar beet varieties. *Annals of Applied Biology*, 94(3) :431–443, 1980.

H. Monod, C. Naud et D. Makowski : Uncertainty and sensitivity analysis for crop models. *In* D. Wallach, D. Makowski et J. W. Jones, éds : *Working with Dynamic Crop Models*, p. 55–96. Elsevier, 2006.

J. Monteith : Climate and the efficiency of crop production in Britain. *Proceedings of the Royal Society of London*, 281 :277–294, 1977.

C. Morrell, J. Pearson, H. Carter et L. Brant : Estimating Unknown Transition Times Using a Piecewise Nonlinear Mixed-Effects Model in Men with Prostate Cancer. *Journal of the American Statistical Association*, 90(429), 1995.

M. D. Morris : Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33 (2) :161–174, 1991.

P. Mykland, L. Tierney et B. Yu : Regeneration in markov chain samplers. *Journal of the American Statistical Association*, 90(429) :233–241, 1995.

S. Newcomb : A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8(4) :343–366, 1886.

A. Nothdurft, E. Kublin et J. Lappi : A non-linear hierarchical mixed model to describe tree height growth. *European Journal of Forest Research*, 125(3) :281–289, mars 2006. ISSN 1612-4669.

T. Orchard et M. Woodbury : A missing information principle : theory and applications. *In* L. Le Cam, J. Neyman et J. Scott, éds : *Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Theory of Statistics*, vol. 1 : Theory of Statistics. University of Californy Press, 1972.

J. Perttunen, R. Sievänen, E. Nikinmaa, H. Salminen, Saarenmaa, A. Väkev et Others : LIGNUM : a tree model based on simple structural units. *Annals of Botany*, 77(1) :87, 1996.

G. Piñeiro, S. Perelman et J. Guerschman : How to evaluate models : Observed vs ; predicted or predicted vs. observed ? *Ecological Modelling*, 216(3-4) :316–322, 2008.

J. C. Pinheiro et D. M. Bates : Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4(1) :12–35, 1995.

B. T. Polyak et A. B. Juditsky : Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4) :838–855, 1992.

P. Prusinkiewicz et A. Lindenmayer : *The algorithmic Beauty of Plants*. Springer New York, 1990.

P. Prusinkiewicz, A. Lindenmayer et J. Hanan : Development models of herbaceous plants for computer imagery purposes. *In Proceedings of SIGGRAPH '88*, p. 141–150, Atlanta, États-Unis, 1988.

A. Racine-Poon : A bayesian approach to nonlinear random effects models. *Biometrics*, p. 1015–1023, 1985.

J.-F. Renno et T. Winkel : Phenology and reproductive effort of cultivated and wild forms of Pennisetum glaucum under experimental conditions in the Sahel : implications for the maintenance of polymorphism in the species. *Canadian Journal of Botany*, 74 :959–964, 1996.

H. Robbins et S. Monro : A stochastic approximation method. *The Annals of Mathematical Statistics*, p. 400–407, 1951.

C. Robert, T. Rydén et D. Titterington : Convergence controls for MCMC algorithms, with applications to hidden Markov chains. *Journal of Statistical Computation and Simulation*, 64 :327–355, 1999.

C. Robert et G. Casella : *Monte Carlo Statistical Methods*. Springer Texts in Statistics Series. Springer-Verlag GmbH, 1999.

G. O. Roberts et J. S. Rosenthal : Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4) :351–367, nov. 2001.

G. Roberts, A. Gelman et W. Gilks : Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied …*, 7(1) :110–120, 1997.

G. Roberts et R. Tweedie : *Understanding MCMC*. Springer, 2008.

F. Ruget, N. Brisson, R. Delécolle et R. Faivre : Sensitivity analysis of a crop simulation model, STICS, in order to choose the main parameters to be estimated. *Agronomie*, 22 :133–158, 2002.

A. Saltelli, K. Chan et E. M. Scott : *Sensitivity Analysis*. Wiley, 2000.

A. Saltelli, M. Ratto, T. Andres et F. Campolongo : *Global sensitivity analysis : the primer*. Wiley, 2008.

A. Saltelli, S. Tarantola, F. Campolongo et M. Ratto : *Sensitivity Analysis in Practice : A Guide to Assessing Scientific Models*. John Wiley & Sons, 2004.

E. D. Schulze : *Plant life forms and their carbon, water, and nutrient relations*, p. 615–676. Springer-Verlag, 1982.

S. G. Self et K.-Y. Liang : Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398) :605–610, 1987.

L. Sibma : Maximization of arable crop yields in the netherlands. *Netherlands Journal of Agricultural Science*, 25 :278–287, 1977.

R. Sievänen, E. Nikinmaa, P. Nygren, H. Ozier-Lafontaine, J. Perttunen et H. Hakula : Components of functional-structural tree models. *Annals of Forest Science*, 57(5) :399–412, 2000.

I. Sobol : Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1993.

M. Stout : A new look at some nitrogen relationships affecting the quality of sugar beets. *J. Am. Soc. Sugar Beet Technol*, 11(5) :388–398, 1961.

N. A. Streck, R. A. Bellé, E. K. da Rocha et M. Schuh : Estimating leaf appearance rate and phyllochron in safflower (*Carthamus tinctorius L.*). *Ciència Rural*, 35 :1448–1450, 2005.

R. Sundberg : Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, 1 :49–58, 1974.

A. Taky : *Maîtrise des excès d'eau hivernaux et de l'irrigation et de leurs conséquences sur la productivité de la betterave sucrière dans le périmètre irrigué du Gharb (Maroc). Analyse expérimentale et modélisation*. Thèse de doctorat, AgroParisTech, 2008.

The Monolix Team : User Guide to Monolix, 2011.

W. M. Thorburn : The myth of occam's razor. *Mind*, 107(27) :345–353, 1918.

M. Tremblay et D. Wallach : Comparison of parameter estimation ethods for crop models. *Agronomie*, 24 :351–365, 2004.

S. Trevezas et P. Cournède : A sequential Monte Carlo approach for MLE in a plant growth model. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(2) :250–270, 2013.

S. Trevezas, S. Malefaki et P.-H. Cournède : Simulation techniques for parameter estimation via a stochastic ECM algorithm with applications to plant growth modeling. *Submitted*, 2012.

G. Vieilledent, B. Courbaud, G. Kunstler, J.-F. Dhôte et J. S. Clark : Individual variability in tree allometry determines light resource allocation in forest ecosystems : a hierarchical Bayesian approach. *Oecologia*, 163(3) :759–73, 2010.

E. F. Vonesh : Non-linear models for the analysis of longitudinal data. *Statistics in Medicine*, 11(14-15) :1929–1954, 1992.

J. Vos, L. Marcelis et J. Evers : *Functional-structural plant modelling in crop production*, chap. 1. Springer, 2007.

D. Wallach et B. Goffinet : Mean Squared Error of Prediction in Models for Studying Ecological and Agronomic Systems. *Biometrics*, 43(3) :561, sept. 1987.

D. Wallach, D. Makowski et J. Jones : *Working with Dynamic Crop Models : Evaluation, Analysis, Parameterization, and Applications*, chap. Evaluating crop models, p. 11–53. Elsevier Science Ltd, 2006.

D. Wallach, B. Goffinet, J.-e. Bergez, P. Debaeke et D. Leenhardt : Parameter Estimation for Crop Models : A New Approach and Application to a Corn Model. *Agronomy Journal*, 93(4) :757–766, 2001.

D. Watson : Comparative physiological studies in the growth of filed crops. I. Variation in net assimilation rate and leaf area between species and varieties, and within and between years. *Annals of Botany*, 11 : 41–76, 1947.

G. C. G. Wei et 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.

W. W. Wilhelm et G. S. McMaster : Importance of the phyllochron in studying in development and growth in grasses. *Crop Science*, 35(1) :1–3, 1995.

C. F. J. Wu : On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1) :95–103, 1983.

Q. Wu : *Sensitivity Analysis for Functional Structural Plant Modelling*. Thèse de doctorat, École Centrale Paris, 2012.

Q. Wu, B. Bayol, F. Kang, J. Lecoeur et P.-H. Cournède : Sensitivity analysis for plant models with correlated parameters : Application to the characterization of sun flower genotypes. *In 7th International Conference on Sensitivity Analysis of Model Output*, Nice, France, juillet 2013.

Q. Wu, P.-H. Cournède et A. Mathieu : An efficient computational method for global sensitivity analysis and its application to tree growth modelling. *Reliability Engineering and System Safety*, juil. 2011.

Q. Xue, A. Weiss et P. S. Baenziger : Predicting leaf appearance in field-grown winter wheat : evaluating linear and non-linear models. *Ecological Modelling*, 175(3) :261–270, 2004.

H.-P. Yan, M.-Z. Kang, P. de Reffye et M. Dingkuhn : A Dynamic, Architectural Plant Model Simulating Resource-dependent Growth. *Annals of Botany*, 93(5) :591, 2004.

X. Yin, P. Stam, M. Kropff et A. Schapendonk : Crop modeling, QTL mapping, and their complementary role in plant breeding. *Agronomy Journal*, 95(1) :90–98, 2003.

X. Yin et P. C. Struik : Modelling the crop : from system dynamics to systems biology. *Journal of Experimental Botany*, 61(8) :2171–2183, 2010.

Z.-G. Zhan, P. De Reffye, F. Houllier et B. Hu : Fitting a functional-structural growth model with plant architectural data. *In Proceedings of the First International Symposium on Plant Growth Models and Applications (PMA03)*, Beijing, China, 2003.