RMT modélisation

www.modelia.org

ACTA

Le réseau des instituts des filières animales et végétales

INRA

Institut National de la Recherche Agronomique

EXPLORATION ET ANALYSE DE MODÈLES POUR L'AGRONOMIE ET L'ENVIRONNEMENT : CRÉATION D'UN PACKAGE DE RESSOURCES PÉDAGOGIQUES SOUS LE LOGICIEL R.

Mémoire de Master II

Sylvain TOULET

Sous la direction de : David MAKOWSKI, François BRUN & Daniel WALLACH.

Université de Toulouse

Université Paul Sabatier TOULOUSE III

M2I Modélisation des Systèmes Ecologiques 2011-2012

INTRODUCTION

Les modèles dynamiques sont largement utilisés dans de nombreux domaines scientifiques, ils permettent de simuler un système qui évolue au cours du temps de façon causale, son avenir ne dépendant que de phénomènes du passé et du présent, c'est à dire que pour une « condition initiale » donnée à l'instant « présent » ne correspondra qu'un seul état « futur » possible pour les versions déterministes (quand il n'y a pas de stochasticité dans le calcul de la réponse). Ce type de modèle dynamique est largement utilisé pour représenter les systèmes agronomiques ou d'élevage. Les modèles les plus couramment utilisés prennent la forme d'un système d'équations différentielles ou aux différences qui représentent la dynamique des différentes composantes du système (sol, plantes, pathogènes,...). La résolution numérique de ce système d'équations permet de simuler la dynamique du système. Ces modèles peuvent être utilisés pour explorer les effets induits sur le système modélisé de changements dans les caractéristiques du sol, du climat, et des pratiques agricoles. En particulier ils sont utilisés pour analyser l'impact des pratiques agricoles sur la production et l'environnement, pour évaluer et concevoir des pratiques innovantes, en tant qu'outil d'aide à la décision, ou pour aider à la planification d'expérimentations.

Pour la représentation des systèmes de production végétale, les modèles de culture se basent pour la plupart sur les mêmes mécanismes (Boote *et al.*, 1995 ; Wallach *et al.*, 2006) qui sont le développement (le stade phénologique dépend de la somme de températures), la croissance foliaire par l'augmentation de la surface foliaire et l'accumulation de biomasse (Monteith, 1972) qui permettent d'intercepter plus de lumière nécessaire à la croissance de la plante (Sinclair & Seligman, 1995). Ces trois mécanismes sont suffisants pour reproduire l'essentiel de la dynamique de croissance potentielle d'une plante en condition cultivée (Fig. 1). La particularité des modèles de cultures est de dépendre de données météorologiques (généralement

journalières) permettant de décrire le rayonnement, les températures de l'air et les précipitations entre autres. À chaque jour de simulation, le modèle prendra donc en compte une information externe de laquelle l'évolution du système dépendra. De nombreux modèles de culture sont basés sur ces quelques processus, mais sont rendus plus complexes par l'ajout dans le système de mécanismes de sénescence, de stress hydrique, azoté...

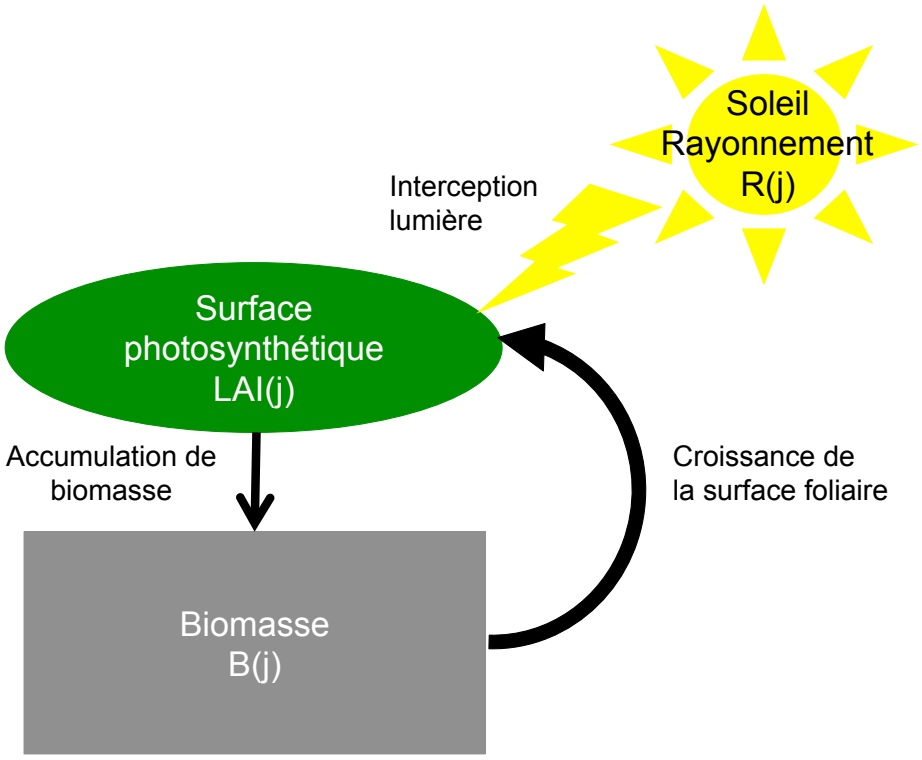

Figure 1 : *Exemple d'un modèle conceptuel basique de culture*

Le développement de tels modèles requiert une connaissance approfondie du fonctionnement du système considéré, mais la construction et l'analyse des modèles dynamiques posent aussi des questions méthodologiques qui relèvent des domaines des mathématiques et des statistiques (Fig. 2). Ces questions sont toutefois encore assez peu abordées dans le domaine de la modélisation en agronomie ce qui génère des difficultés dans la manipulation et l'utilisation des modèles : les modélisateurs sont généralement des spécialistes du système à modéliser mais manquent de méthodes mathématiques et statistiques pour travailler avec les modèles (paramétrage, analyses

d'incertitude et de sensibilité, évaluation) malgré l'existence de méthodes éprouvées et relativement accessibles associées à la puissance de calcul des ordinateurs qui permet de les mettre en application par tous les modélisateurs.

Le Réseau Mixte Technique (RMT) Modélisation (Modélisation et logiciels d'intérêt commun appliqués à l'Agriculture : www.modelia.org) a pour vocation d'organiser les échanges autour de la modélisation pour l'agriculture entre l'INRA et des instituts techniques agricoles (réseau ACTA). En particulier, le réseau organise des formations centrées autour du thème de la modélisation. Ces formations permettent de développer l'information et la connaissance sur la modélisation d'une manière générale mais toujours concernant des domaines d'applications liés à l'agronomie. Cependant, le réseau organise aussi régulièrement des formations focalisées sur des problèmes précis liés au processus de modélisation, notamment sur toutes les méthodes statistiques et mathématiques à utiliser.

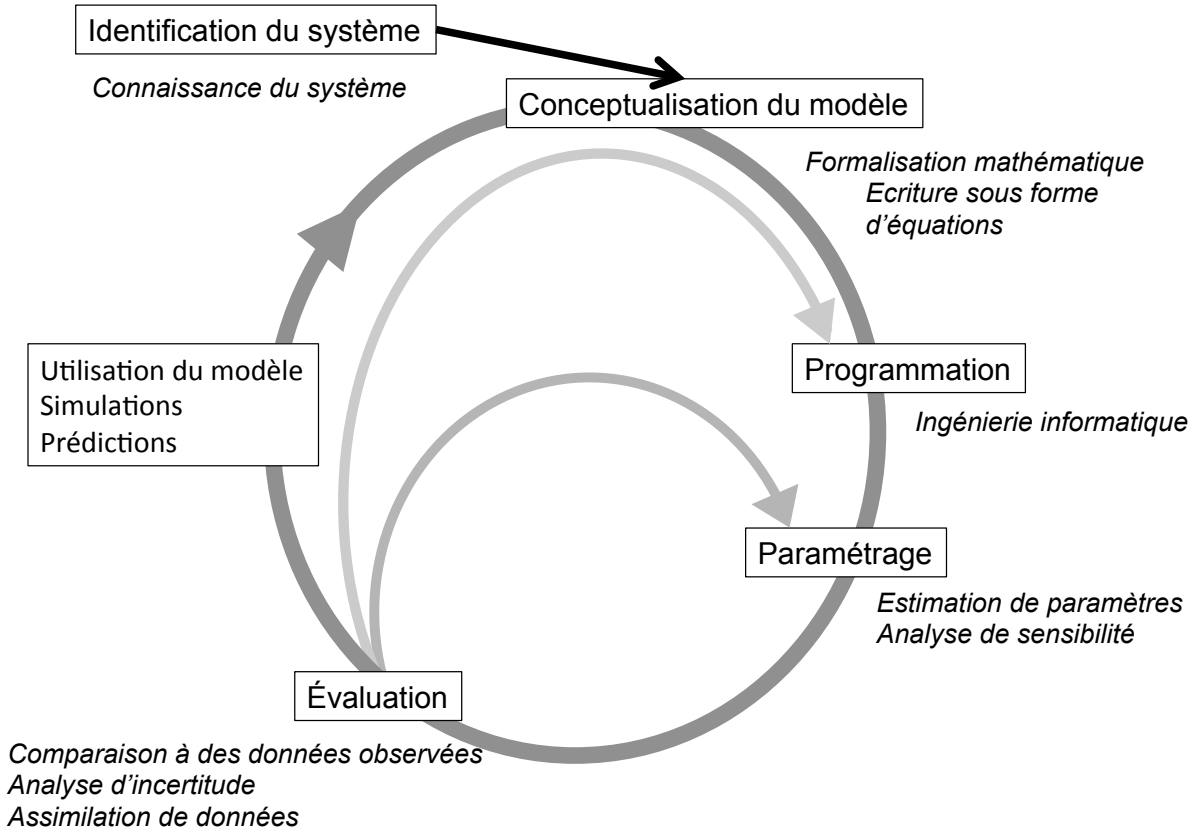

Figure 2 : *Étapes du processus de modélisation d'un système dynamique*

Parallèlement aux formations, David Makowski, Daniel Wallach, Jim Jones et François Brun travaillent sur la nouvelle édition du livre « Working with dynamic crop models » (Wallach *et al.*, 2006). Cet ouvrage est un livre pédagogique sur la modélisation appliquée aux systèmes agricoles. Il contient de nombreux exemples de modèles pour l'explication pratique de méthodes statistiques sur les modèles par le biais d'exemples et d'exercices fournis sous la forme de scripts R. Le public visé par ces formations et ce livre se compose des chercheurs, ingénieurs et étudiants en agronomie et qui sont amenés à aborder leurs sujets d'études par le biais de la modélisation.

Le sujet de mon stage s'inscrit dans ces deux actions que sont l'organisation de formations et l'écriture de ce livre. Dans le cadre de ce travail d'équipe, m'ont été confiées différentes tâches précises. L'objectif principal de ce stage est de contribuer à la création d'un package R permettant la mise en œuvre facile de différentes méthodes sur les modèles. Ainsi, dans le cadre de ce stage, mon travail a été de plusieurs types. Il a fallu dans un premier temps bien définir les modèles utilisés comme exemples dans le livre et lors des formations. Le but étant de définir des modèles de complexités différentes permettant de bien illustrer et appréhender les différents aspects des travaux de modélisation des systèmes dynamiques. Cette étape inclut la bonne compréhension du fonctionnement des modèles, et leur programmation en langage R d'une manière claire et transparente permettant à toute personne de comprendre le fonctionnement du modèle à partir du script R. Ensuite, à partir de ces modèles, l'étape suivante était de construire des fonctions R permettant de réaliser de nombreuses simulations de modèles et d'appliquer les différentes méthodes mathématiques et statistiques.

Dans ce rapport, je présenterai d'abord succinctement le contenu du package R créé, puis les différents modèles sélectionnés ainsi que les méthodes appliquées dans la suite de ce rapport. Cependant par souci de clarté et de pédagogie, le rapport ne

sera pas structuré selon le plan classique (Introduction - Matériel et méthodes – Résultats – Discussion). Il nous a semblé plus judicieux de présenter d'abord le résultat concret (la librairie R) puis d'organiser le rapport en fonction des méthodes présentées servant d'exemples d'application de la librairie. Ainsi après la description du contenu du package, dans un premier temps, je présenterai deux modèles sur lesquels ont été appliquées diverses méthodes d'analyse de sensibilité, un modèle de rendement du blé en fonction de l'infestation par une mauvaise herbe (modèle WEED) et un modèle de culture du colza (modèle AZODYN-Colza). Les méthodes d'analyse de sensibilité ainsi que des exemples de résultats seront décrites après la présentation des modèles. Deux méthodes seront particulièrement mises en avant : l'analyse de sensibilité par Analyse de variance (ANOVA) et la méthode de Morris.

Dans un second temps, je présenterai un modèle de culture du maïs (maize-model) qui servira d'exemple à la présentation des méthodes d'estimation de paramètres par les méthodes des moindres carrés ordinaires et des moindres carrés pondérés (OLS et WLS).

Dans une dernière partie de présentation de méthodes, j'introduirai la méthode d'évaluation de modèle ROC appliquée à des modèles d'estimation de la teneur en protéine de grains de blés.

Enfin, dans une dernière partie de conclusion, je présenterai le travail restant à accomplir durant le dernier mois de stage, notamment concernant la mise en forme de tous les modèles, de toutes les ressources pédagogiques créées ainsi que des jeux de données sous la forme d'un package R documenté au format R-Oxygen. Je conclurai en présentant les enjeux que représente la création de telles ressources pédagogiques et l'utilisation qui en sera faite dans un avenir proche.

DESCRIPTION DU PACKAGE « ZeBook »

Le package ZeBook comprend différentes fonctions correspondant aux différents modèles sélectionnés comme exemples. Ces fonctions sont nommées selon le modèle qu'elles utilisent (exemple `weed.model` voir extrait de code en annexe 1). Elles sont codées sous la forme scripts R documentés au format R-Oxygen. Ces fonctions permettent de réaliser des simulations en utilisant différents jeux de données (notamment météorologiques) qui sont inclus également dans le package ou encore en modifiant facilement les valeurs de certains paramètres (en arguments de la fonction).

Le package contient également de nombreuses fonctions permettant de mettre en œuvre différentes techniques statistiques appliquées aux modèles. Dans la suite, seront présentées plusieurs techniques ainsi que des exemple de leur fonctionnement sous R et des résultats pouvant être obtenus. Ci dessous est présentée une description non exhaustive sur le contenu du package créé.

- Modèles et fonctions associées : par exemple pour le modèle WEED :
    - `weed.model` pour réaliser des simulations du modèle,
    - `weed.define.param` pour définir les valeurs des paramètres,
    - `weed.simule` pour réaliser des simulations à partir d'un plan expérimental.
    - etc.
- Jeux de données :
    - données météorologiques NASA (de 1984 à 2011),
    - données de LAI observé sur cultures de maïs,
    - données de contenu en protéine de grains de blé,
    - etc.
- Fonctions pour l'application de méthodes, avec par exemple :
    - `OLS` (fonction permettant de réaliser l'estimation de paramètres par la méthode des moindre carrés pondérés ordinaires),
    - `weed.anova` (fonction permettant de réaliser l'analyse de sensibilité par ANOVA sur le modèle WEED).
    - etc.

- Documentation : pour chaque fonction : une documentation rédigée en format R-Oxygen pour expliquer les différents termes et le fonctionnement.
- Exemples : des exemples d'utilisation des fonctions, inclus dans la documentation, mais également pour certains servant d'exercices lors des formations.

Voir la fonction `weed.model` en exemple de code R inclus dans le package et documenté au format R-Oxygen en Annexe 1.

Par la suite, seront présentés plusieurs méthodes inclues dans le package ainsi que de nombreux exemples de résultats pouvant être obtenus par la mise en œuvre des différentes fonctions créées et regroupées au sein du package.

MÉTHODES D'ANALYSE DE SENSIBILITÉ

Ces méthodes visent à identifier dans quelles mesures les facteurs d'entrée d'un modèle influent sur les sorties issues de simulations.

Modèles utilisés :

Dans le cadre de la présentation des différentes analyses de sensibilité, nous avons sélectionnés deux modèles pour qu'ils servent d'exemples afin d'expliquer les différentes méthodes. Le premier est un modèle de rendement du blé en fonction de l'infestation par une mauvaise herbe. Le second est un modèle simulant le rendement et la teneur en protéines des grains à la récolte d'une culture de colza.

- Modèle de rendement du blé : `weed.model`

Ce modèle est un Modèle de système dynamique simulant la croissance d'une culture de blé et d'une population de vulpin sur la même parcelle représentés par le rendement en blé en $T.ha^{-1}$ (Y), le nombre de graines de vulpin par m^2 (S), la densité de plantes de vulpin à émergence (début de saison : d), le nombre de graines de vulpin par m^2 dans le sol en surface après travail du sol (SSBa) et le nombre de

graines de vulpin par m^2 dans le sol en profondeur après travail du sol (DSBa) (Munier-Jolain *et al.*, 2002 ; Munier-Jolain *et al.*, 2008) .

Ce modèle utilise en entrée les décisions de techniques culturales appliquées pour les années que l'on veut simuler. Ces décisions concernent le travail du sol, l'application d'herbicide et le type de culture de blé pratiquée et sont codées en binaires :

- Pour ce qui est du travail du sol : Soil = 1 si labour, Soil = 0 si travail superficiel,
- Pour l'herbicide : Herb = 1 si application d'un traitement, Herb = 0 sinon,
- Pour le type de culture : Crop = 1 si on considère la culture du blé d'hiver, Crop = 0 sinon.

Le modèle est défini par 8 équations avec un total de 16 paramètres. Quatre paramètres déterminent l'effet du travail du sol sur l'enfouissement et la mise à jour des graines de vulpin (β_0, β_1, χ_0 et χ_1), 5 définissent l'augmentation du nombre de graines (δ_{new}, δ_{old}, S_{max0}, S_{max1} et v), 4 déterminent les pertes en graines (mh, mc, ϕ et μ) et 3 qui définissent la production de blé (Y_{max}, r_{max} et γ).

Les deux premières équations permettent de définir les quantités de graines de vulpin présentes dans les deux zones considérées (Surface et Profondeur).

Le stock de graines de vulpin en surface au début de l'année i avant travail du sol ($SSBb_i$) correspond au stock précédent moins la proportion ayant dépéri ou étant devenue plantule plus la quantité nette de graines produites (équation (1)). Le stock de graines en profondeur au début de l'année i avant travail du sol ($DSBb_i$) correspond au stock précédent moins la proportion ayant dépéri (équation (2)).

$$SSBb_i = (1-\mu)[SSBa_{i-1} - d_{i-1}] + v(1-\phi)S_{i-1} \quad (1)$$

$$DSBb_i = (1-\mu)DSBa_{i-1} \quad (2)$$

Deux autres équations rendent comptent de l'effet du travail de la terre sur le transfert de graines entre les deux zones (flux). β décrit le flux de graines de la surface vers la profondeur et χ le flux inverse (de la profondeur vers la surface). En fonction du type de travail du sol (Soil) on utilisera les paramètres possédant les indices correspondant (par exemple β_1 quand Soil = 1, c'est à dire que le travail du sol cet année là sera un réel labour et non un travail superficiel). Ces deux équations nous permettent donc de simuler la quantité de graines présentes dans les deux zones à l'année i après travail du sol ($SSBa_i$ et $DSBa_i$ respectivement équations (3) et (4)).

$$SSBa_i = (1-\beta)SSBb_i + \chi SSBb_i \quad (3)$$

$$DSBa_i = (1-\chi)DSBb_i + \beta SSBb_i \quad (4)$$

L'équation suivante permet de simuler ensuite à partir des différents stocks de graines (de surface uniquement) le nombre de plantules de vulpin émergeant (d_i, équation (5)).

$$d_i = \delta_{new}[S_{i-1}v(1-\phi)(1-\beta)] + \delta_{old}\{SSBa_i - [S_{i-1}(1-\phi)(1-\beta)]\} \quad (5)$$

Les deux équations suivantes servent à simuler le nombre de plantules de vulpin parvenant à maturité (D_i). La proportion de plants arrivant à maturité dépendant de la décision concernant l'application d'herbicide (Herb), une équation correspondra au cas où aucun traitement herbicide n'est appliqué et où la mortalité des plantules ne sera due qu'au froid, (Herb = 0, équation (6)) et l'autre correspondra au cas où il y'a eu application d'herbicide à l'année considérée où la mortalité totale des plantules sera une combinaison de la mortalité due au froid et de celle due à l'herbicide (Herb = 1, équation (7)).

$$D_i = (1-m_c)d_i \quad \text{quand Herb} = 0 \quad (6)$$

$$D_i = (1-m_h)(1-m_c)d_i \quad \text{quand Herb} = 1 \quad (7)$$

La dernière équation est celle permettant à partir de la quantité de plants de vulpins matures calculés précédemment (D_i) et d'un rendement maximum (Y_{max}), de déterminer quel sera le rendement en blé à l'année de simulation (Y_i, équation (8)).

$$Y_i = Y_{\max}(1 - \frac{r_{\max} D_i}{1 + \gamma D_i}) \quad (8)$$

La sortie du modèle sera donc la prédiction du rendement annuel en blé pour chaque année de simulation.

- Modèle de culture du colza : AZODYN-Colza (Jeuffroy & Recous, 1999)

Ce modèle dynamique simule le rendement et la teneur en protéines des grains à la récolte d'une culture de colza. L'objectif premier du modèle était d'être utilisé comme d'outil dans le but de raisonner la fertilisation azotée en fonction d'objectifs de production quantitatifs et qualitatifs mais également de critères économiques et environnementaux. Il a été construit à partir de la variété Soissons, majoritairement cultivée en France au cours des années 1990-2002 (Prost *et al.*, 2003). Le modèle simule la croissance de la plante de la sortie-hiver à la maturité physiologique de la culture. Il se compose d'un module « sol », décrivant la dynamique de fourniture en azote du sol, et d'un module « plante », décrivant le fonctionnement de la culture, à pas de temps journalier sur l'intégralité de la période de simulation (Barbottin, 2004 ; Vocanson, 2006).

Le modèle tel qu'il a été utilisé fonctionne sur la plateforme de modélisation RECORD-VLE, résultat et instrument d'une large collaboration entre développeurs de modèles, méthodologistes de la modélisation et de la décision et enfin chercheurs et ingénieurs qui analysent, évaluent et conçoivent des systèmes de culture à l'INRA (www.inra.fr/record). Grâce à la plateforme de simulation VLE (Virtual Laboratory Environment, basée sur le formalisme de spécification des systèmes à événements discrets DEVS (Discrete Event System Specification, Quesnel *et al.*, 2009) il est possible d'utiliser de nombreux modèles codés sous formes de librairies C++ (VFL :

VLE Foundation Librairies). Cela offre la possibilité de réaliser des simulations, de développer des modèles (via l'interface graphique G-VLE) ainsi que d'analyser et de visualiser les résultats de simulation. Il est à noter qu'il existe un package R permettant d'interagir avec VLE afin d'exécuter les simulations sous VLE via R et de récupérer ensuite les résultats de cette simulation (package R-VLE, voir annexe 4). C'est par le biais de ce package R que toutes les analyses de sensibilités effectuées sur le modèle AZODYN-Colza et qui seront présentées par la suite ont été réalisées.

Le modèle AZODYN-Colza simule au jour le jour, l'activité du sol et de la culture jusqu'au stade de « maturité physiologique », qui correspond à la date pour laquelle l'accumulation d'azote et de carbone dans le grain est terminée. Au début de la simulation (sortie-hiver), le modèle a besoin de plusieurs variables d'initialisation :

- les données météo journalières entre la date de semis et la date de maturité : températures moyennes, rayonnement global, pluviométrie et évapotranspiration ;
- les caractéristiques du sol : la teneur en argile, en calcaire et en azote organique de la couche labourée, la densité apparente et l'épaisseur de cet horizon, le reliquat d'azote minéral dans le sol à la sortie de l'hiver ;
- l'histoire culturale de la parcelle : précédent cultural, gestion des résidus (enfouissement ou non et fréquence d'enfouissement) ainsi que la fréquence d'apports organiques, les dates et les doses d'engrais azoté apportées ;
- les caractéristiques de la culture : la biomasse aérienne au jour de début de simulation, les dates des différents stades clés de la culture (semis, «épi 1cm» et floraison), et la date moyenne de maturité physiologique.

Lorsque la maturité physiologique est atteinte, le modèle simule le rendement, le nombre de grains, la teneur en protéines des grains ainsi que le reliquat d'azote minéral du sol à la récolte. Toutes les variables intermédiaires utilisées par le modèle peuvent également être accessibles : la matière sèche aérienne de la culture, la quantité d'azote total, la quantité d'azote remobilisé, la quantité d'azote végétatif, etc. Les résultats des analyses de sensibilité appliquées au modèle AZODYN-COLZA sont présentés en annexes (Annexe 5).

Principe

L'objectif de l'analyse de sensibilité est de définir à quel point la sortie d'un modèle est sensible à la variabilité des paramètres ou des variables d'entrée du modèle. En ce qui concerne les modèles dynamiques, la variabilité de ces paramètres peut être associée au manque de précision concernant leur estimation. L'analyse de sensibilité est donc une étape importante dans le processus de modélisation notamment pour identifier les paramètres auxquels la sortie du modèle est la plus sensible afin de pouvoir procéder à leur estimation de manière plus précise. L'analyse de sensibilité repose sur les résultats de simulation du modèle, ainsi, le développement croissant du calcul et des simulations informatiques a permis l'essor et le développement de nombreuses méthodes d'analyse de sensibilité (Saltelli *et al.*, 2000 ; Saltelli *et al.*, 2008).

- ANOVA

Cette analyse se base sur le principe du plan expérimental factoriel afin de déterminer l'effet de plusieurs facteurs. Le principe est de définir un certain nombre de niveaux pour les paramètres à analyser et de construire un plan expérimental comprenant l'intégralité des combinaisons possibles (Frey & Patil, 2002).

Pour un cas de K facteurs ayant chacun n modalités, il y'a donc n^K scénarios d'entrée possibles et après simulation n^K sorties du modèle.

L'analyse de variance sera basée sur la décomposition de la variabilité de la réponse et les contributions de chacun des facteurs. L'avantage de cette méthode est qu'elle permet de prendre en compte facilement les interactions entre facteurs ce qui permettra d'évaluer également leur contribution. L'intérêt est alors de comparer la contribution des différents facteurs à la variabilité totale. Cela peut être fait en calculant les indices de sensibilité par ANOVA en divisant la somme des carrés des écarts obtenue pour chaque facteur par la variabilité totale.

$$S_i = \frac{SCE_i}{SCE_{Tot}}$$

A noter que SCE_i peut représenter la somme des carrés des écarts associée à un facteur mais aussi à l'interaction entre deux facteurs.

Exemple de résultats sur le modèle WEED

L'analyse de sensibilité a ici eu lieu sur le rendement en blé simulé à l'année 3 lorsqu'il n'y a pas eu de traitement herbicide cette année là. Les facteurs sélectionnés sont les 5 paramètres présentés plus haut pour lesquels ont été choisi 3 niveaux qui sont les bornes minimales et maximales et la valeur moyenne. C'est donc un plan expérimental de type 3^5 qui a été utilisé pour calculer les indices de sensibilité. Par souci pédagogique, voici un extrait du code R utilisé afin de réaliser cette analyse de sensibilité par ANOVA et de calculer les indices.

```
Yield <- weed.simul5p(param, para.mat, weed.deci);
Tab <- data.frame(Yield, mu, beta.1, beta.0, mh, Ymax);
Fit <- summary(aov(Yield~mu*beta.1*beta.0*mh*Ymax, data = Tab));
SumSq <- Fit[[1]][,2];
Total <- (nlevel^npar-1)*var(Yield);
Indices <- 100*SumSq/Total;
TabIndices <- cbind(Fit[[1]],Indices);
```

Extrait de code R permettant la mise en œuvre de l'analyse de sensibilité par la méthode de l'ANOVA et de calculer les indices de sensibilité à partir des résultats.

`para.mat` correspond à la matrice correspondant au plan d'expérience factoriel (3^5) et a été obtenue en utilisant la fonction `expand.grid`. `weed.simul5p` est la fonction permettant de simuler grâce au modèle WEED les rendements en blé à l'année 3 en prenant en compte les différentes combinaisons des 5 paramètres inclues dans `para.mat`. L'appel de l'ANOVA sous R se fait classiquement grâce à la fonction `aov`. On peut noter ici, que l'analyse de sensibilité inclut toutes les interactions entre paramètres. Une partie du tableau résultant `TabIndices` est présenté ci-dessous.

	Df	Sum Sq	Mean Sq	Indices
Ymax	1	1.979201e+01	1.979201e+01	5.259623e+01
mh	1	5.877156e-01	5.877156e-01	1.794335e+01
beta.1	1	2.453599e-01	2.453599e-01	1.166581e+01
beta.1:mh	1	2.453599e-01	2.453599e-01	1.166581e+01
mu	1	7.256478e-02	7.256478e-02	1.000143e+01

Extrait du tableau récapitulant les résultats de l'ANOVA et listant les indices associés à chaque paramètres et interactions.

Ce type de résultat est classiquement représenté par un diagramme en bâton avec les différents indices de sensibilité calculés pour chaque facteur et chaque interaction, ici est présenté en exemple les résultats en considérant le rendement en blé de l'année 3 quand il n'y a pas eu application de traitement herbicide (Fig. 3).

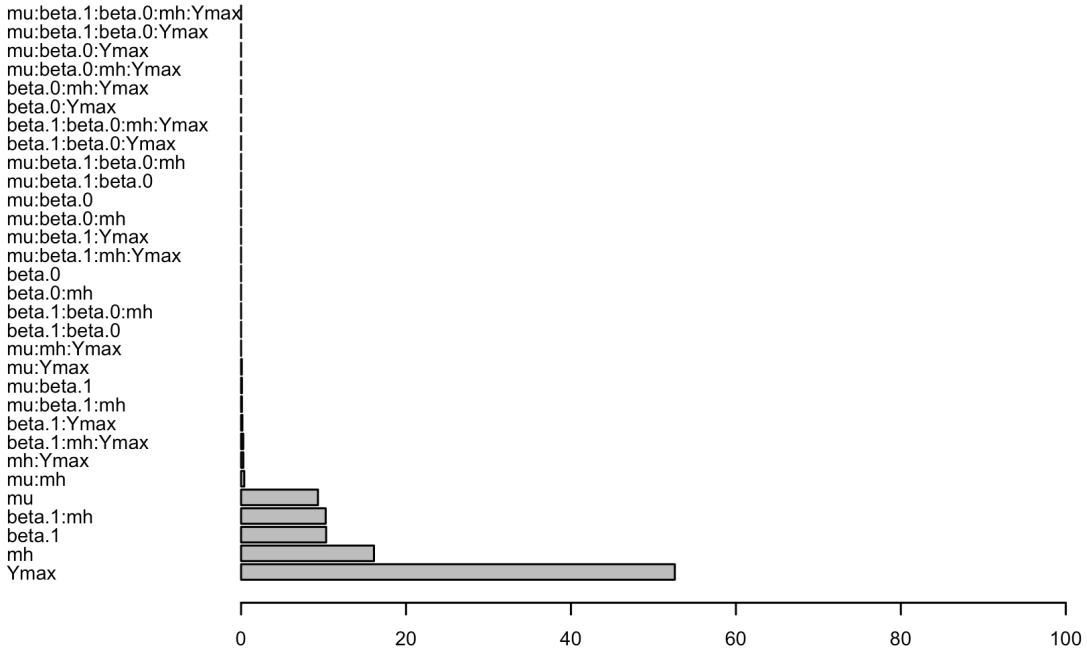

Figure 3 : *Indices de sensibilité des facteurs principaux et des interactions calculés par la méthode de l'ANOVA sur 5 paramètres du modèle WEED en l'absence de traitement herbicide à l'année 3.*

Nous pouvons voir ici que le facteur Y_{max} est le plus important, les interactions entre facteurs sont pour la plupart très faibles excepté celle entre β_1 et mh.

- Méthode de Morris

La méthode qui semble la plus évidente pour réaliser une analyse de sensibilité est de faire varier la valeur d'un facteur en laissant les autres à une valeur fixe afin

de détecter l'effet de ce paramètre uniquement sur la sortie du modèle. Ce genre de méthode est appelé One-At-a-Time (OAT), et permet de faire une analyse de sensibilité sur un seul paramètre à la fois, ce qui peut être utile mais il est généralement préférable de réaliser une analyse de sensibilité globale utilisant le principe de l'OAT sur différentes combinaisons de valeurs des différents facteurs incertains. Cela permet de plus de prendre en compte et de quantifier les interactions entre facteurs.

C'est ce que permet de faire la méthode de Morris (Morris, 1991 ; Campolongo *et al.*, 2007 ; Pujol, 2008), elle se base sur le principe que pour une combinaison de valeurs (V_i) des N facteurs incertains donnée :

$V_i = (v_{i1}, \dots, v_{in}, \dots, v_{iN})$,

l'effet élémentaire du n^{ieme} facteur est :

$$d_n(V_i) = \frac{f(v_{i1},\dots,v_{in}+\Delta,\dots,v_{iN}) - f(v_{i1},\dots,v_{in},\dots,v_{iN})}{\Delta}.$$

En construisant S différents scénarios (combinaisons de valeurs pour les facteurs), balayant toutes les valeurs possibles pour les différents facteurs et de calculer pour chaque facteur et pour chaque combinaison la valeur de $d_n(V_i)$. Ainsi, pour chaque facteur nous obtiendrons une distribution des valeurs représentant son effet dans les différents scénarios. Ces distributions sont caractérisées par la moyenne (μ_n^*) et la variance (σ_n^2) des $d_n(V_i)$ (respectivement, équations 9 et 10).

$$\mu_n^* = \frac{\sum_{i=1}^{S} \left| d_n(V_i) \right|}{S} \tag{9}$$

$$\sigma_n^2 = \frac{\sum_{i=1}^{2} [d_n(V_i) - \frac{1}{S}\sum_{i=1}^{S} d_n(V_i)]^2}{S} \tag{10}$$

Un facteur possédant une forte moyenne μ_n^* sera un facteur ayant un forte importance sur la sortie du modèle, une forte valeur de la variance σ_n^2 sera indicatrice de l'existence d'une interaction entre le facteur considéré et un autre ou alors d'un facteur dont l'effet est non linéaire.

La mise en œuvre d'une telle méthode peut être réalisée sous le logiciel R grâce à la fonction `morris` présente dans le package `{sensitivity}`. Comme indiqué dans l'extrait de script R présenté ci-dessous et appliquant la méthode au modèle WEED, il faut pour utiliser cette fonction préciser le nombre de facteurs ou comme ici leurs noms (`paraNames`), les valeurs minimales et maximales pour chaque facteur (ici présentes dans le tableau `weed.factors`, Tab. 1), la valeur de Δ (`grid.jump`), le nombre de niveau du plan expérimental (`levels`) et le nombre de répétitions (`r`).

```
> weed.factors
           mu    v   phi beta.1 beta.0 chsi.1 chsi.0 delta.new delta.old    mh  mc
nominal 0.840 0.60 0.550  0.950   0.20   0.30  0.050     0.150      0.30 0.980 0.0
binf    0.756 0.54 0.495  0.855   0.18   0.27  0.045     0.135      0.27 0.882 0.0
bsup    0.924 0.66 0.605  1.000   0.22   0.33  0.055     0.165      0.33 1.000 0.1

        Smax.1 Smax.0 Ymax   rmax  gamma
nominal  445.0  296.0  8.0 0.0020 0.0050
binf     400.5  266.4  7.2 0.0018 0.0045
bsup     489.5  325.6  8.8 0.0022 0.0055
```

Tableau 1 : *Tableau des paramètres du modèle WEED.* `nominal`, `binf` *et* `bsup` *correspondent respectivement aux valeurs nominales, de la borne inférieure et de la borne supérieure.*

```
output.morris = morris(model =  weed.simule, factors = paraNames,
r = 500, design = list(type = "oat", levels = 4, grid.jump = 2), scale=T,
binf = weed.factors["binf",], bsup = weed.factors["bsup",], weed.deci);
```

Extrait de code R détaillant la mise en œuvre de la fonction `morris {sensitivity}`

La représentation graphique typiquement associée à cette méthode consiste à représenter pour chacun des facteurs la variance σ_n^2 en fonction de la moyenne μ_n^*. Elle est obtenue sous R par la ligne de commande :

```
plot(output.morris);
```

Un exemple d'une telle représentation est présenté ci-dessous (Fig. 4). Il correspond à l'analyse effectuée sur le modèle WEED sur 10 ans avec application d'herbicide chaque année, à l'exception de l'année 3 (même cas que l'analyse par ANOVA).

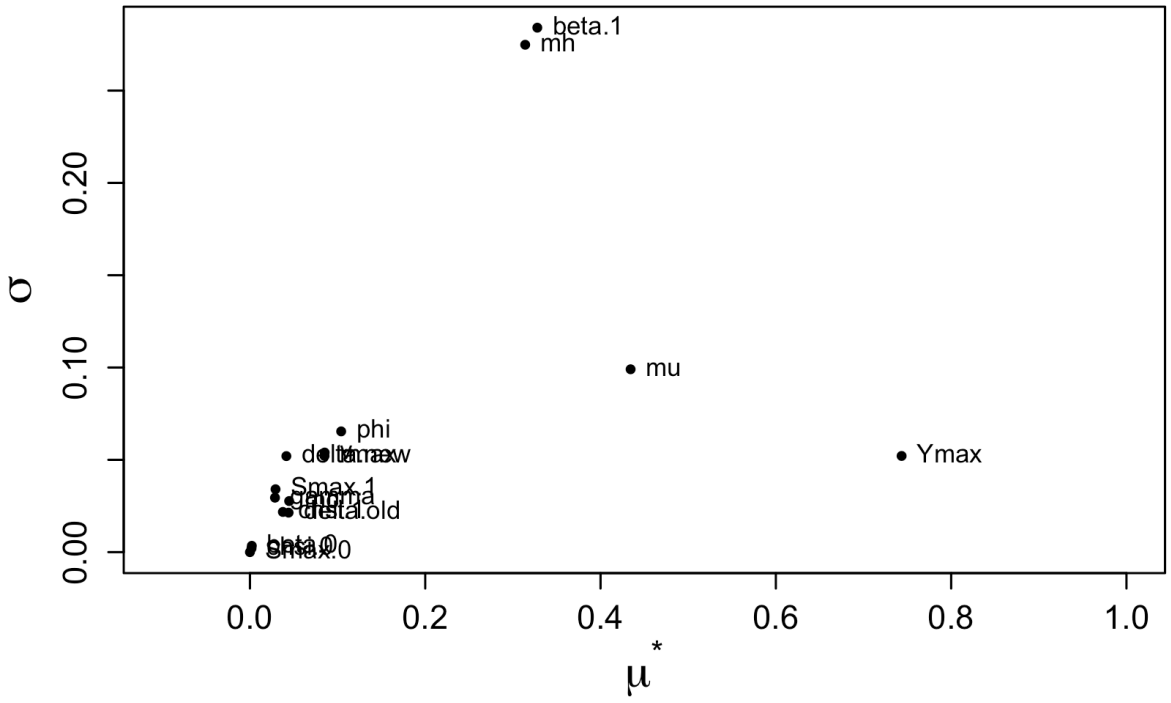

Figure 4 : *Résultats de l'analyse de Morris sur le modèle WEED simulant le rendement en blé sur 10 années (sans application d'herbicide la 3ème année).* levels = 4, $\Delta = 2$ *et* $r = 500$.

Nous pouvons voir que comme dans l'analyse par ANOVA, le paramètre Y_{max} est le plus influent sur la sortie du modèle (μ_n^* important), suivi de μ, β_1 et mh. Nous pouvons également voir que ces deux derniers paramètres sont ceux qui possèdent le σ_n^2 le plus important. Cela confirme l'existence et l'importance de l'interaction entre ces deux paramètres qui ont été observées lors de l'analyse par ANOVA. Les autres paramètres apparaissent une nouvelle fois comme ayant très peu d'importance (faibles σ_n^2 et faibles μ_n^*).

D'autres méthodes d'analyse de sensibilité ont été mises en œuvre sur ces deux modèles, notamment la méthode FAST (Cukier *et al.*, 1978 ; Saltelli *et al.*, 1999) et

la méthode de Sobol (Sobol, 1993 ; Saltelli, 2002) présentes également dans le package `{sensitivity}` de R. Dans le tableau ci-dessous est présenté un récapitulatif du coût en simulation de chaque méthode (Tab. 2).

	ANOVA	Morris	FAST	Sobol
Coût (nombre de simulations)	m^k	$r(k+1)$	$N \times k$	$N(2k+1)$ $N(k+2)$ (Saltelli, 2002)

Tableau 2 : *Récapitulatif des coûts en simulation de différentes méthodes de simulation. k : nombre de facteurs ; m : nombre de modalités par facteur ; r : nombre de scenarios ; N : taille de l'échantillon.*

Les résultats des différentes méthodes d'analyse de sensibilité présentes dans le package et appliquées au modèle AZODYN-Colza sont présentés en Annexe 5.

MÉTHODES D'ESTIMATION DE PARAMÈTRES

Ces méthodes ont pour but d'estimer la valeur de un ou plusieurs paramètres d'un modèle pour que les sorties de simulation s'ajustent le mieux possible à des données observées.

Modèle utilisé : maize.model

Pour mettre en œuvre et construire les différentes méthodes d'estimation de paramètres, nous avons utilisé un modèle de culture du maïs en situation potentielle. C'est un modèle de système dynamique reprenant un formalisme retrouvé dans de nombreux modèles de culture (Wallach *et al.*, 2001 ; Jones *et al.*, 2003 ; Brisson *et al.*, 2009) et simulant la croissance de la culture en situation potentielle, représentée par trois variables d'état : la surface foliaire par unité de sol (Leaf Area Index, LAI), la biomasse totale accumulée (B) et le temps thermique cumulé depuis l'émergence (TT). Ce modèle est basé sur les principaux concepts inclus dans la plupart des modèles de

cultures, du moins pour la partie « potentielle ». En effet, ce modèle simplifié n'inclut aucun effet de la disponibilité en eau et nutriments, bio-agresseurs, etc.

Le modèle est défini par plusieurs équations, avec au total 7 paramètres. Trois déterminent des stades phénologiques (T_{base} : Température de base pour qu'il y ait croissance, TT_M : Somme de température pour atteindre la maturité de la culture, TT_L : Somme de température à la fin de la croissance foliaire) et on les considérera comme bien connus et « fixes ». Quatre concernent les processus d'interception de la lumière (K), d'accumulation de la biomasse (RUE) et de croissance de la surface foliaire (α : taux relatif de croissance du LAI pour de faibles valeurs de LAI, LAI_{max} : LAI maximum). Ce modèle utilise comme variables d'entrées des variables climatiques (rayonnement, température minimum et maximum journalière) et les dates de semis (sdate) et de récolte (ldate). Les variables d'état sont des variables dynamiques évoluant en fonction du temps. Ainsi, on s'intéresse à ces variables en fonction de la durée en jour depuis l'émergence (j) : TT_j, B_j et LAI_j.

Les équations permettant de décrire le modèle sont :

$$TT_{j+1} = TT_j + \Delta TT_j \tag{11}$$

$$B_{j+1} = B_j + \Delta B_j \tag{12}$$

$$LAI_{j+1} = LAI_j + \Delta LAI_j \tag{13}$$

Avec :

$$\Delta TT_j = max\left[\frac{TMIN_j + TMAX_j}{2} - T_{base}, 0\right] \tag{14}$$

$$\Delta B_j = RUE.(1 - e^{-K.LAI_j}).I_j \quad \text{quand} \quad TT_j \leq TT_M \tag{15}$$
$$= 0 \quad \text{quand} \quad TT_j > TT_M$$

$$\Delta LAI_j = \alpha.\Delta TT_j.LAI_j.max\left[LAI_{max} - LAI_j, 0\right] \quad \text{quand} \quad TT_j \leq TT_L \tag{16}$$
$$= 0 \quad \text{quand} \quad TT_j > TT_L$$

Principe

Cette méthode est utilisée afin d'obtenir un estimateur pour un ou plusieurs paramètres incertains et ne pouvant pas être estimées à partir de mesures sur le processus. Ces estimateurs sont obtenus grâce à un échantillon de données mesurées correspondant à des sorties du modèle. Ici par nécessité, le but de ce travail étant pédagogique, nous n'avons pas utilisé de données mesurées réelles, mais des données issues de simulations du modèle à partir d'un très grand nombre de données climatiques (plusieurs site-années) et bruitées selon une procédure particulière (voir détail en Annexes 2 et 3) afin d'en obtenir une quantité très importante dans le but d'illustrer au mieux les méthodes lors des formations. J'en ferais toutefois référence par la suite comme des données «observées». Ici nous présenterons deux méthodes d'estimation de paramètres basées sur la méthode des moindres carrés : la méthode des moindres carrés ordinaires (Ordinary Least Squares : OLS) et la méthode des moindres carrés pondérés (Weighted Least Squares : WLS). Appliquées à notre modèle exemple, ces méthodes permettront d'obtenir des distributions d'estimateurs pour deux paramètres du modèle que nous avons choisis : les deux paramètres liés à la croissance de la surface foliaire, α et LAI_{max}.

- Moindres Carrés Ordinaires (OLS):

Si nous possédons un échantillon de N valeurs mesurées, la valeur du paramètre θ estimée par cette méthode sera celle pour laquelle la somme des carrés des écarts ($OLS(\theta)$) entre les valeurs mesurées (Y) et les valeurs obtenues par le modèle ($f(X,\theta)$) est minimale (équation (17)). Pour cela, à partir d'une valeur initiale pour le paramètre à estimer, et via l'algorithme de Gauss-Newton (fonction `nls` sous R, Bates & Watts, 1988 ; Seber & Wild, 2003 ; Bates & Chambers, 1992), la valeur du paramètre va évoluer par itérations successives (dans un espace de même dimension que le nombre de paramètres à estimer) jusqu'à atteindre la valeur pour laquelle la somme des carrés des écarts sera minimale (Fig. 5).

$$OLS(\theta) = \sum_{i=1}^{N} [Y_i - f(Y_i, \theta)]^2 \quad (17)$$

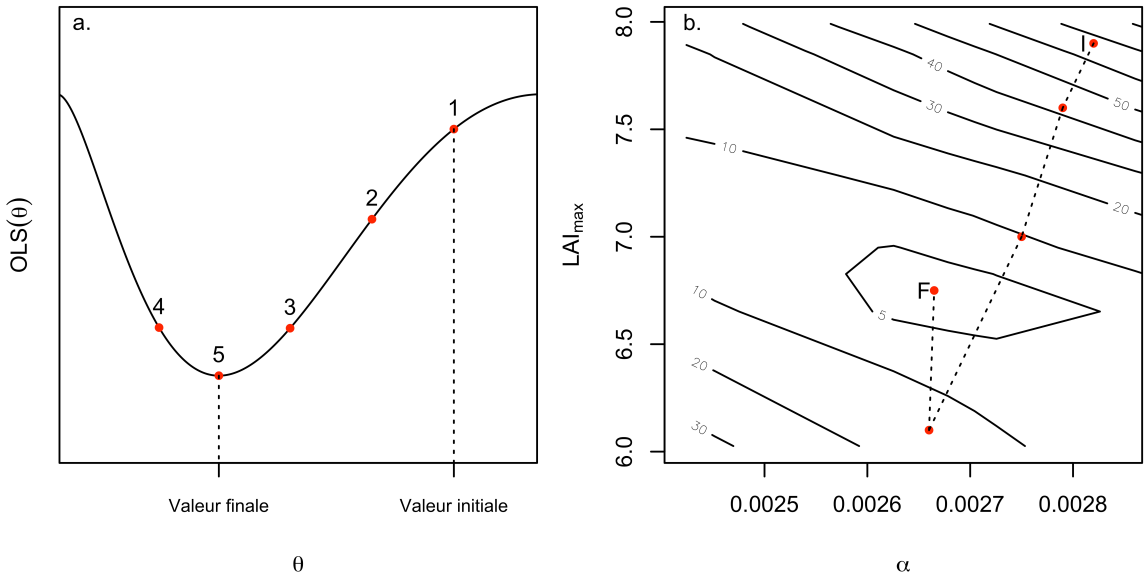

Figure 5 : *Principe de fonctionnement de l'algorithme de Gauss-Newton pour l'estimation de paramètre :*

a. Dans cet exemple à un seul paramètre à estimer, il faut 4 itérations à l'algorithme pour trouver la valeur de θ qui minimise la somme des carrés des écarts.

b. Ce graphe présente un exemple de trajectoire dans le plan pour l'estimation des deux paramètres du modèle maïs (α et LAI_{max}). Les lignes de niveaux correspondent aux valeurs de la somme des carrés des écarts dans l'espace correspondant aux valeurs des deux paramètres à estimer. I correspond au point du plan associé aux valeurs initiales et F au point du plan atteint par l'algorithme et minimisant la valeur de la somme des carrés des écarts.

Dans notre cas, pour estimer α et LAI_{max}, nous avons décidé d'utiliser des valeurs de LAI « observées » au jour 30 et au jour 60 après semis (sdate) afin de les comparer aux valeurs de LAI simulées au jour 30 et au jour 60 dans le modèle. Pour obtenir une distribution de chaque paramètre estimé, nous avons donc construit 500 échantillons de 20 valeurs : 10 valeurs de LAI au jour 30 et 10 valeurs au jour 60 obtenues dans chaque échantillon pour 10 sites-années différents.

Pour chaque échantillon, nous avons donc obtenu une valeur estimée pour chaque paramètre ce qui nous permet d'obtenir une distribution de ces estimateurs.

- Moindres carrés pondérés (WLS) :

Dans la méthode des moindres carrés ordinaires, la somme des carrés des écarts correspond à la somme de tous les termes sans pondération, cela sous-entend que l'incertitude sur la variable mesurée est constante. Si l'incertitude sur Y n'est pas constante, il est nécessaire d'utiliser la méthode des moindres carrés pondérés. En effet, il semble nécessaire et important d'accorder une importance plus forte aux valeurs mesurées dont on est « sûr » par rapport à celles sur lesquelles l'incertitude est plus grande. Pour prendre cela en compte, il faut pondérer l'écart par cette incertitude (qu'elle soit connue ou estimée à priori).

La somme à minimiser ici sera donc :

$$WLS(\theta) = \sum_{i=1}^{N} \omega_i [Y_i - f(Y_i, \theta)]^2 \tag{18}$$

où ω_i est l'inverse de l'écart-type associé à Y_i

Dans notre cas sur le modèle exemple, nous avons pris :

$$\omega_i = \frac{1}{\sqrt{f(X_i, \theta)^2 - f(0, \theta)^2}} \tag{19}$$

Résultats

L'estimation des paramètres α et LAI_{max} par ces deux méthodes nous permet d'obtenir des résultats comparables dans les deux cas. En ce qui concerne le paramètre α correspondant au taux relatif d'augmentation de l'indice foliaire lorsque celui-ci a des faibles valeurs, l'estimation par OLS sur le 500 échantillons nous donne une distribution du paramètre dont la valeur médiane est de 0,002407 et dont les valeurs minimales et maximales sont respectivement de 0,001089 et 0,003273. La distribution de ce même paramètre obtenue par la méthode WLS a pour valeur

médiane 0,002425 et pour valeurs minimales et maximales respectivement 0,002148 et 0,002721 (Fig. 6).

Pour ce qui est de l'estimation du paramètre LAI_{max} correspondant à l'index foliaire maximum, l'estimation par OLS nous donne une distribution de valeur médiane 7,055 et de bornes minimales et maximales respectivement 5,420 et 14.519. La distribution obtenue par WLS est de valeur médiane 7,014 et de valeurs minimales et maximales 6,263 et 7,937 respectivement (Fig. 7).

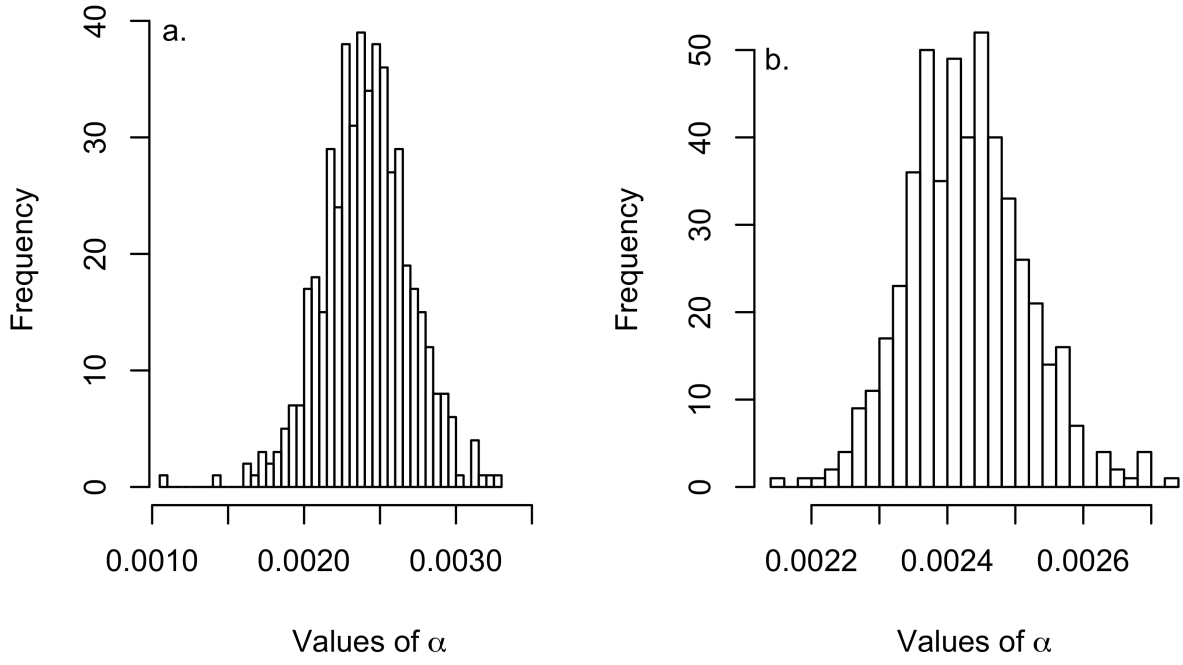

Figure 6 : *Histogrammes des distributions du paramètre α. Obtenues par : a. la méthode des moindres carrés ordinaires et b. la méthode des moindres carrés pondérés (500 échantillons à chaque fois)*

Nous pouvons voir que l'estimation de ces paramètres par ces deux méthodes nous donne des valeurs similaires, centrée sur environ 0,00242 pour α et sur 7 pour LAI_{max}. Toutefois, nous pouvons remarquer que les variances des distributions obtenues par la méthode des moindres carrés pondérés sont moins importantes que celle obtenues par la méthode des moindres carrés ordinaires. Cette première méthode semble donc plus précise que la dernière car elle présente moins de variabilité dans l'estimation, cette dernière semblant plus précise dans la méthode WLS que dans la

méthode OLS. La pondération permet donc d'augmenter la précision et la fiabilité de l'estimation de paramètres.

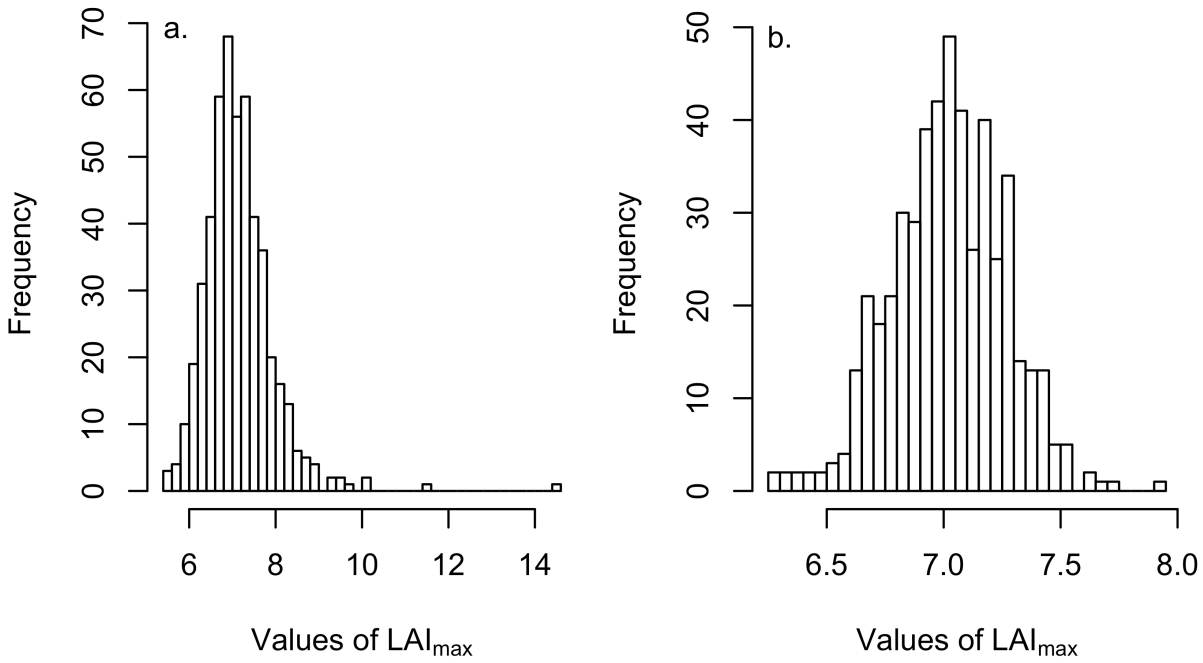

Figure 7 : *Histogrammes des distributions du paramètre LAI_{max}. Obtenues par : a. la méthode des moindres carrés ordinaires et b. la méthode des moindres carrés pondérés (500 échantillons à chaque fois).*

MÉTHODE D'ÉVALUATION DE MODÈLE

Les méthodes d'évaluation de modèles peuvent avoir plusieurs buts, que ce soit pour déterminer su le modèle répond suffisamment précisément à la question posée, pour intervenir dans le travail d'amélioration d'un modèle ou encore lorsque il est nécessaire de choisir entre différents modèles ayant la même application. Les erreurs sur un modèle peuvent être de différents types, elles peuvent être des erreurs de prédiction sur les valeurs absolues ou sur le classement. Il est nécessaire d'évaluer les erreurs associées à un modèle pour en faciliter le choix car il existe parfois plusieurs modèles pour une même application. L'intérêt ici est de déterminer le niveau d'erreur associé à un modèle. Les modèles agronomiques jouent un rôle croissant en ce qui concerne les expertises scientifiques, il est ainsi d'autant plus important d'apporter une information concernant les erreurs liées à l'utilisation d'un modèle. On utilise

classiquement l'erreur quadratique moyenne de prédiction pour évaluer la qualité d'un modèle, mais cela n'apporte pas d'indication sur l'erreur liée au classement.

La méthode ROC (Receiving Operating Characteristics ; Sing *et al.*, 2005) qui est une méthode particulièrement utilisée en médecine se base sur une règle de décision binaire. Un seuil de décision (S) est fixé et si la prédiction (P) d'un modèle est supérieure à ce seuil, on considère une variable d'intérêt servant de référence $Y = 1$ (ou VRAI), si ce n'est pas le cas, on considère $Y = 0$ (ou FAUX).

Il existe alors deux types d'erreurs de classification, les faux positifs quand $P > S$ mais $Y = 0$ et les faux négatifs quand $P < S$ mais que $Y = 1$. Le but de l'analyse ROC est d'estimer les fréquences de faux négatifs et de faux positifs pour tous les seuils de décision S. On détermine alors deux indices, la sensibilité et la spécificité tels que :

$$Sensibilité\ (S) = \frac{Nombre\ de\ vrais\ positifs}{Nombre\ total\ de\ situations\ où\ Y = 1}$$

$$Spécificité\ (S) = \frac{Nombre\ de\ vrais\ négatifs}{Nombre\ total\ de\ situations\ où\ Y = 0}$$

L'analyse se décompose ainsi en plusieurs étapes qui consistent tout d'abord à déterminer la valeur de l'indicateur I pour chaque situation, à définir ensuite un seuil S, de calculer les valeurs des indices de sensibilité et de spécificité. La représentation de l'analyse se fait grâce à la courbe ROC ($Sensibilité$ en fonction de $(1 - Spécificité)$). La statistique associée à l'analyse correspond à la valeur d'aire sous la courbe ROC (AUC). Si AUC n'est pas différent de 0,5 cela signifie que l'indicateur n'est pas meilleur qu'un classement aléatoire.

L'exemple présenté ici correspond à l'évaluation de deux modèles de culture du blé, une variation du modèle AZODYN (Barbottin *et al.*, 2006, David *et al.*, 2004, Jeuffroy & Recous, 1999) présenté plus haut adapté à la culture du blé et permettant de prédire le rendement en grain, le contenu en protéine des grains ainsi que la

quantité d'azote minéral présent dans le sol après récolte. Cela étant fait à partir des caractéristiques du sol, des données météorologiques journalières et l'état de la culture à la fin de l'hiver (Modèle 1). Le second modèle développé par Makowski *et al.* (2001) est un modèle à paramètre aléatoire basé sur un ensemble de quatre équations non linéaires. Il permet de prédire le rendement, le contenu en protéine des grains et la quantité d'azote minéral dans le sol après récolte à partir de deux variables d'entrées, la quantité totale d'azote appliqué et la quantité d'azote dans le sol à la fin de l'hiver (Modèle 2). L'analyse a été menée sur le contenu en protéine des grains mesuré dans 43 parcelles différentes et sur le contenu en protéine des grains prédit par les deux modèles. Le seuil de décision choisi pour l'analyse est de 11,5% correspondant à un seuil de qualité défini pour la farine destinée à la fabrication du pain. Ci-dessous est présenté un extrait du tableau de donnée ainsi que le code R permettant de construire le tableau de décision binaire associé en fonction du seuil choisi. Ensuite est détaillée la méthode pour appliquer la méthode ROC grâce au package `{ROCR}`, de récupérer la valeur d'aire sous la courbe `AUC`.

```
data :
    GPC   GPC.model1   GPC.model2
1  11.9         12.0         11.3
2  10.3         11.3         11.1
3  12.4         10.9         11.7
4  12.1         10.2         11.8
5  12.4          9.4         11.5
...

library(ROCR);
Ref <- data$GPC;
Seuil <- 11.5;
Ref[Ref < Seuil] <- 0;
Ref[Ref >= Seuil] <- 1;

# ROC analysis on model 1
pred.1 <- prediction(data$GPC.model1, Ref);
perf.1 <- performance(pred.1,"auc");
auc.1 <- perf.1@"y.values";
# ROC analysis on model 2
pred.2 <- prediction(data$GPC.model2, Ref);
perf.2 <- performance(pred.2,"auc");
auc.2 <- perf.2@"y.values";
```

Extrait de code R permettant de mettre en œuvre la méthode ROC

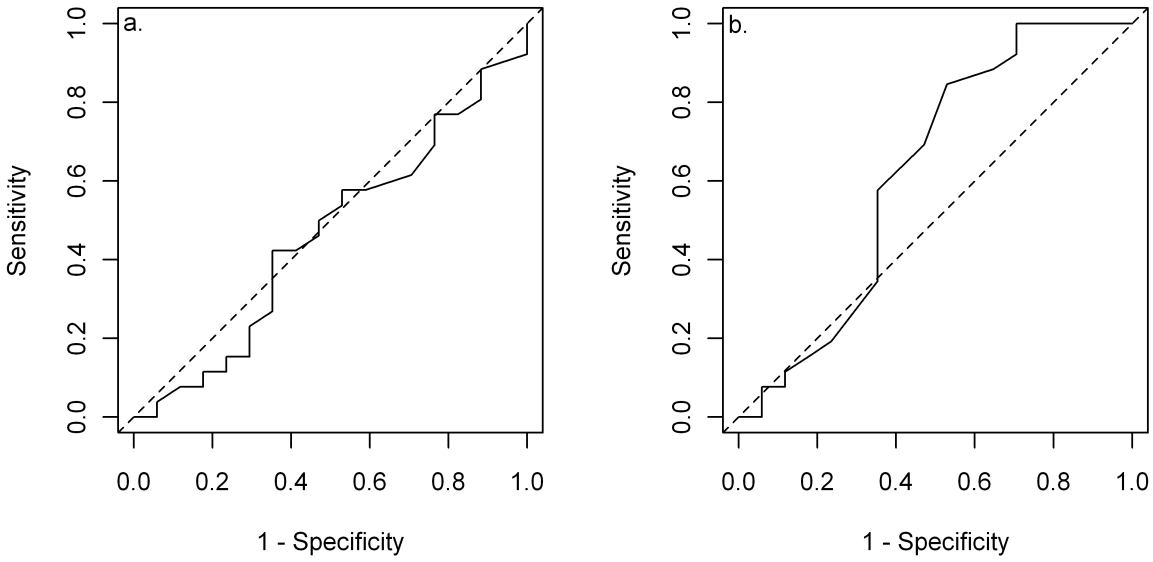

Figure 8 : *Courbes ROC obtenues pour le modèle 1 et le modèle 2 (respectivement a. et b.). Les lignes pointillées représentent le cas théorique ayant une aire sous la courbe de 0,5.*

Les résultats que nous obtenons pour cette analyse sur les deux modèles considérés (Fig. 8) nous permettent de conclure que le modèle 1 n'est pas bon en ce qui concerne le classement (valeur d'AUC de 0,46), en effet il ressort que l'indicateur de ce modèle n'est pas meilleur qu'un classement aléatoire. En ce qui concerne les résultats obtenus pour le modèle 2, nous pouvons voir que ce dernier est meilleur qu'un classement aléatoire (et donc que le modèle 1) car il possède une AUC associée de 0,623. Il faut toutefois noter que les performances des deux modèles pris en exemple ici ne sont pas excellentes.

CONCLUSIONS ET PERSPECTIVES

Le travail réalisé durant ce stage a permis d'atteindre les objectifs fixés en début de stage, c'est à dire la création de ressources pédagogiques inclues dans un package R qui servira de support lors de formations. Ces différentes ressources sont notamment basées sur des fonctions permettant de réaliser facilement et en grand nombre des simulations de différents modèles agronomiques, de comparer aisément les résultats à des données observées. Ils permettent également par le biais de fonctions

simples d'accès d'utiliser différentes techniques d'analyse de sensibilité, d'estimation de paramètres et d'évaluation de modèle. Il est à noter que pour d'autres techniques concernant l'estimation de paramètres par des méthodes Bayésiennes, l'assimilation de données ou encore l'optimisation pour la gestion de système, les outils permettant la mise en application ont été réalisés par mes co-encadrants.

Le travail que j'ai réalisé m'a permis de me perfectionner dans la programmation sous R en m'introduisant le principe de la documentation R-Oxygen. J'ai également découvert de nombreuses techniques qui m'étaient inconnues, notamment concernant les différentes méthodes d'analyses de sensibilité (Morris, FAST et Sobol notamment) et d'évaluation de modèle (méthode ROC).

Le travail n'est toutefois pas terminé, il reste de nombreuses choses à accomplir avant la mise à disposition du package R, certaines méthodes restent à y être inclues et certaines documentations sont encore incomplètes. L'étape suivante de mon travail de stage est de réaliser une étude comparative de trois modèles de culture du blé fonctionnant sur la plateforme de modélisation Record-VLE en réalisant des simulations basées sur différents scenarios de réchauffement climatique qui devrait servir de support à une formation à destination de la communauté RECORD-VLE.

BIBLIOGRAPHIE

BARBOTTIN A. (2004). Utilisation d'un modèle de culture pour évaluer le comportement des génotypes : Pertinence de l'utilisation d'Azodyn pour analyser la variabilité du rendement et de la teneur en protéines du blé tendre. Thèse, UMR d'Agronomie INRA de Grignon.

BARBOTTIN A., MAKOWSKI D., LE BAIL M., JEUFFROY M.H., BARRIER C. (2008). Comparison of models and indicators for categorizing soft wheat fields according to their grain protein contents. *European Journal of Agronomy* **28**, 175-183

BATES D.M. & WATTS D.G. (1988). Nonlinear Regression Analysis and Its Applications. *Wiley*.

BATES D.M. & CHAMBERS J.M. (1992). *Nonlinear models.* Chapter 10 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

BOOTE K.J., JONES J.W., PICKERING N.B. (1995). Potential uses and limitations of crop models. *Agronomy Journal.* Vol. **55** N°**5**, 704-716.

BRISSON N., LAUNAY M., MARY B. & BEAUDOIN N. (2009). Conceptual basis, formalisations and parametrization of the STICS crop model. *Update Sciences & Technologies, Editions Quæ.*

CAMPOLONGO F., CARIBONI J. & SALTELLI A. (2007). An effective screening design for sensitivity. *Environmental Modelling & Software*. **22**, 1509–1518.

CUKIER R.I., LEVINE H.B. & SCHULER K.E. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Comp. Physics*. **26**, 1–42.

DAVID C., JEUFFROY M.H., RECOU, S., DORSAINVILLE F. (2004). Adaptation and assessment of the Azodyn model for managing the nitrogen fertilization of organic winter wheat. *European Journal of Agronomy*. **21**, 249–266.

FREY H.C. & PATIL S.R. (2002). Identification and review of sensitivity analysis methods. *Risk analysis*. Vol. **22**, **3**, 553-578.

JEUFFROY M.H. & RECOUS S. (1999). Azodyn: a simple model simulating the date of nitrogen deficiency for decision support in wheat fertilization. *European Journal of Agronomy.* Vol. **10**, **2**, 129-144.

JONES J. W., HOOGENBOOM G., PORTER C. H., BOOTE K. J., BATCHELOR W. D., HUNT L. A., WILKENS P. W. & RITCHIE J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*. Vol. **18**, **3-4**, 235-265.

MAKOWSKI D., WALLACH D., MEYNARD J.M. (2001). Statistical methods for predicting responses to applied nitrogen and for calculating optimal nitrogen rates. *Agronomy Journal*. **93**, 531–539.

MONTEITH J.L. (1972). Solar radiation and productivity in tropical exosystems. *Journal of Applied Ecology.* **9**, 747-766.

MORRIS M.D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, **33**, 161–174.

MUNIER-JOLAIN N., CHAUVEL B. & GASQUEZ J. (2002). Long-term modelling of weed control strategies: analysis of threshold-based options for weed species with contrasted competitive abilities. *Weed research*, **42**, 107-122.

MUNIER-JOLAIN N., DEYTIEUX V., GUILLEMIN J.-P. & GABA S. (2008). Conception et évaluation multicritères de prototypes de systèmes de culture dans le cadre de la Protection Intégrée contre la flore adventice en grandes cultures. *Innovations Agronomiques*, **3**, 75-88.

PROST L., GAUFFRETEAU A. & TRISTANT D. (2003). Mais où s'en vont les nouvelles variétés de blé tendre ? DAA, Institut National Agronomique Paris-Grignon.

PUJOL G. (2008). Simplex-based screening designs for estimating metamodels. *Reliability Engineering and System Safety*.

QUESNEL G., DUBOZ R. & RAMAT E. (2009). The Virtual Laboratory Environment - An Operational Framework for Multi-Modelling, Simulation and Analysis of Complex Systems. *Simulation Modelling Practice and Theory*. **17**, 641-653.

SALTELLI A., TARANTOLA S. & CHAN K. (1999). A quantitative, model independent method for global sensitivity analysis of model output. *Technometrics*. **41**, 39–56.

SALTELLI A. (2000) Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. *John Wiley & Sons, Ltd.*

SALTELLI A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communication*. **145**, 580–297.

SALTELLI A., RATTO M., ANDRES T., CAMPOLONGO F., CARIBONI J., GATELLI D., SAISANA M. & TARANTOLA S. (2008). Global Sensitivity Analysis. The Primer. *John Wiley & Sons, Ltd.*

SEBER G. A. F. & WILD C. J. (2003). Nonlinear Regression. *Wiley-Interscience, Hoboken, N.J.*

SINCLAIR T.R, & SELIGMAN N.G. (1995). Crop modelling : From infancy to maturity. *Agronomy Journal*. Vol. **88** N°**5**, 698-704.

SOBOL I.M. (1993). Sensitivity analysis for non-linear mathematical model. *Mathematics and Modelling Computer Experiments*, **1**, 407–414.

VOCANSON A. (2006). Evaluation ex ante d'innovations variétales en pois d'hiver (Pisum sativum L.) : approche par modélisation au niveau de la parcelle et de l'exploitation agricole. Thèse, Institut National de la Recherche Agronomique Unité Mixte de Recherche INRA - INA P-G Environnement et Agronomie. Grignon.

WALLACH D., GOFFINET B., BERGEZ J.E., DEBAEKE P, LEENHART D. & AUBERTOT J. N. (2001). Parameter estimation for crop models : a new approach and application to a corn model. *Agronomy Journal*. **93**, 757-766.

WALLACH D., MAKOWSKI D. & JONES J. W. (2006). Working with dynamic crop models. *Elsevier Science Editions*.

ANNEXES

Annexe 1 : Exemple de fonction incluse dans le package ZeBook.

```
#' @title The weed model function
#' @description Model simulating the yield from a wheat culture taking in
 account the population of a weed on the same parcel.
#' @param param : vector of the 16 parameters
#' @param weed.deci : decision table for Soil, Crop et Herbicide
#' @return data.frame with annual values of yield
#' @export

weed.model <- function(param, weed.deci)
{
  # Duration of the simulation according to weed.deci (years)
  duration <- length(weed.deci$Soil);

  # Initialize variables

  # 5 states variables, as 5 vectors initialized to NA

  # Weed density (plants/m2)
  d <- rep(NA,duration+1);

  # Seed production (grains/m2)
  S <- rep(NA,duration+1);

  # Surface seed bank after tillage (grains/m2)
  SSBa <- rep(NA,duration+1);

  # Surface seed bank after tillage (grains/m2)
  DSBa <- rep(NA,duration+1);

  # Yield (tons/ha)
  Yield <- rep(NA,duration+1);

  # Initialize state variables when sowing on day "sdate"
  d[1] <- 400;
  S[1] <- 68000;
  SSBa[1] <- 3350;
  DSBa[1] <- 280;

  # compute Yield for initial year
  D0 <- (1-param["mh"]*weed.deci$Herb[1])*(1-param["mc"])*d[1];

 Yield[1]=max(0,param["Ymax"]*(1(param["rmax"]*D0/(1+param["gamma"]*D0))));
```

```
  # Integration loop
  for (year in 1:(duration))
 {
    # Update state variables
    Z <- weed.update(d[year], S[year], SSBa[year], DSBa[year],
 weed.deci$Soil[year], weed.deci$Crop[year], weed.deci$Herb[year], param);
    d[year+1] <- Z[1];
    S[year+1] <- Z[2];
    SSBa[year+1] <- Z[3];
    DSBa[year+1] <- Z[4];
    Yield[year+1] <- Z[5];
 }
# End simulation loop
return(data.frame(year=0:duration, d=d, S=S, SSBa=SSBa, DSBa=DSBa,
 Yield=Yield));
}
```

Exemple de code R inclus dans le package : la fonction `weed.model` *permettant de réaliser des simulations du modèle WEED (décrit plus bas) et documenté en langage R-Oxygen (balises* `#'` *).*

Annexe 2 : Récupération de données climatologiques en grand nombre.

Pour générer des valeurs « observées » de LAI à partir de données simulées bruitées, nous avons dû réaliser de très nombreuses simulations ce qui implique d'avoir à notre disposition de très nombreuses données météorologiques. Il a donc été décidé d'utiliser la NASA Climatology Resource for Agroclimatology (power.larc.nasa.gov/cgi-bin/cgiwrap/solar/agro.cgi?email=agroclim@larc.nasa.gov) qui nous permet d'obtenir en indiquant la latitude et la longitude d'un lieu les données journalières météorologiques correspondant depuis 1984. En sélectionnant aléatoirement 40 sites dans une zone théorique où l'on peut trouver des cultures de maïs en France (Fig. Ann. 1), nous avons codé un programme permettant à partir des coordonnées GPS de ces lieux d'obtenir les données météorologiques correspondantes de 1984 à 2011.

```
#' @title The fetch_weather function
#' @description This function allows fetching climatic data from the NASA
 database.
#' @param GPSlatitude : Latitudes coordinates corresponding to the site
 from which we want to fetch the climatic data
#' @param GPSlongitude : Longitudes coordinates corresponding to the site
 from which we want to fetch the climatic data
#' @param YearBeg : Integer corresponding to the year from which we want to
 get climatic data (between 1984 and 2011)
#' @param YearEnd : Integer corresponding to the year to which we want to
 get climatic data (between 1984 and 2011)
#' @return data.frame with daily climatic data for the considered site from
 YearBeg to YearEnd
#' @export

fetch_weather <- function(GPSlatitude, GPSlongitude, YearBeg, YearEnd)
{
 url=paste("http://power.larc.nasa.gov/cgi-
 bin/cgiwrap/solar/agro.cgi?email=agroclim%40larc.nasa.gov&step=1&lat=",
 GPSlattitude, "&lon=", GPSlongitude,"&ms=1&ds=1&ys=", YearBeg,
 "&me=12&de=31&ye=", YearEnd, "&submit=Yes", sep="");
 download.file(url, "data.dat", method = "auto", quiet = FALSE, mode = "w",
 cacheOK = TRUE);
}
```

Extrait de code R présentant la fonction permettant de récupérer le fichier de données météorologiques pour un site donné

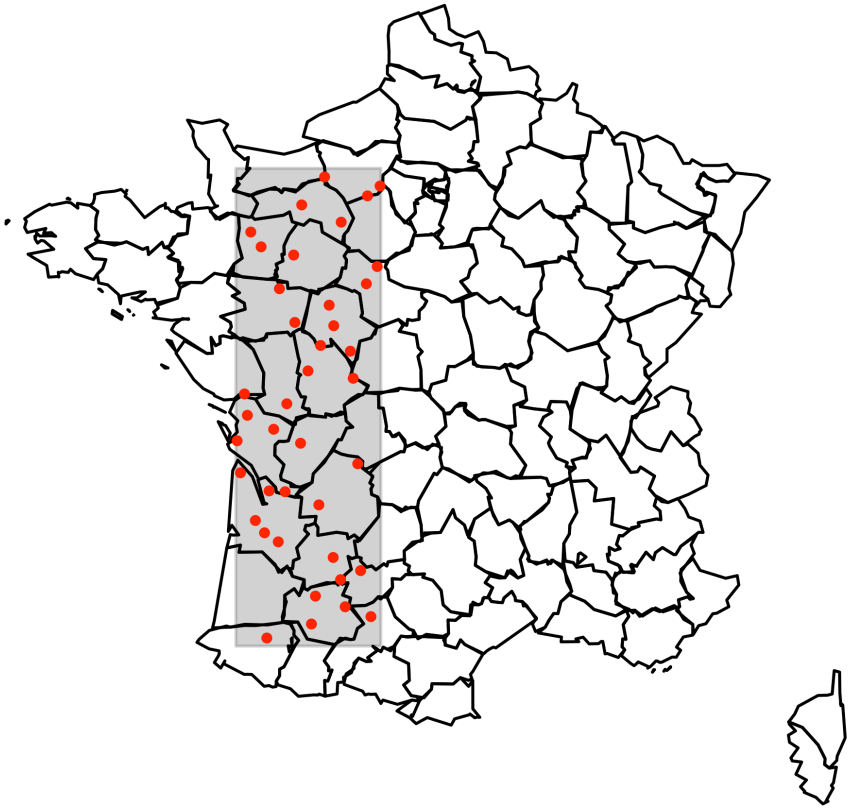

Figure Annexe 1 : *Carte présentant (en grisé) la zone de culture de maïs potentielle et les lieux sélectionnés aléatoirement (points rouges).*

Annexe 3 : Génération de données « observées » bruitées à partir du modèle maize.

L'étape suivante une fois la grande quantité de données météorologiques disponibles a été de réaliser un grand nombre de simulations pour ensuite les bruiter afin d'obtenir des jeux de données que nous avons ensuite considéré par souci de pédagogie comme des données observées. Y_i est la valeur de LAI observée au jour i et f_i est la valeur donnée par le modèle, $Y_i - Y_{i-1}$ et $f_i - f_{i-1}$ sont respectivement l'augmentation de LAI observées et selon le modèle.

Nous avons fait plusieurs hypothèses à propos de l'erreur $\varepsilon_{i,i-1}$ (différence entre $Y_i - Y_{i-1}$ et $f_i - f_{i-1}$) :

$$(Y_i - Y_{i-1}) = (f_i - f_{i-1}) + \varepsilon_{i,i-1}$$

- Aux conditions initiales, l'erreur est nulle ($Y_0 = f_0$),
- Toutes les erreurs sont distribuées normalement avec une espérance de 0 et une variance $\sigma^2_{i,i-1}$:

  $$\varepsilon_{i,i-1} \sim N\left(0, \sigma^2_{i,i-1}\right)$$

- La variance dépend de la valeur de LAI donnée par le modèle aux temps i et $i-1$:

  $$\sigma^2_{i,i-1} = [a(f_i^2 - f_{i-1}^2]^2$$

- Toutes les erreurs $\varepsilon_{i,i-1}$ sont indépendantes (pour chaque jour dans une parcelle et entre parcelles).

Ainsi, pour chaque jour (et chaque valeur de LAI), nous tirons aléatoirement une valeur d'erreur dans la loi normale définie plus haut $\varepsilon_{i,i-1} \sim N\left(0, \sigma^2_{i,i-1}\right)$ et ensuite nous calculons la valeur de LAI « bruitée » au jour i selon l'équation :

$$LAI_i = LAI_{i-1} + (f_i - f_{i-1}) + \varepsilon_{i,i-1}$$

Nous avons ainsi pu créer d'énormes jeux de données pour pouvoir pratiquer l'estimation de paramètres sur le modèle de culture du maïs présenté dans ce rapport.

Annexe 4 : Utilisation du modèle AZODYN-Colza via Rvle

Pour réaliser des simulations du modèle AZODYN-Colza, il est possible de passer par l'interface graphique G-VLE qui permet à l'aide de boutons et de fenêtres à renseigner de modifier les valeurs de certains paramètres, les fichiers de données météorologiques, etc. Le lancement des simulations se fait également via un système fenêtre-bouton. Les résultats peuvent être enregistrés directement dans différents fichiers (.txt, .csv ou .rdata) ou encore être stockés en mémoire sur la machine. C'est par ce biais qu'il devient possible de récupérer les résultats d'une simulation sous R. Le package `{rvle}` permet d'accéder à toutes les options de VLE via R. Il est donc possible de lancer des simulations en modifiant les paramètres initiaux pour ensuite récupérer les résultats directement sous R pour pouvoir les traiter. Ci-dessous est présenté un extrait de code R permettant de réaliser une simulation d'AZODYN-Colza sous VLE via R et d'en récupérer les résultats. Attention, la nature de la sortie récupérable sous R (résultats du modèle) doit être configurée au préalable sous G-VLE.

```
# Loading the needed packages
library(rvle);
# Setting the model to be used
Azodyn.VLE <- new("Rvle", pkg = "Omegasys", file = "azodyn.vpz");
# Running a simulation of the model.
Simulation <- run(Azodyn.vle);
# Getting the needed results
Yield <- (results(Simulation)[[1]][,"Omegasys:Eco2.rendement"]);
```

Extrait de code R présentant les bases pour l'utilisation d'un modèle via RVLE. Il faut d'abord définir le modèle à utiliser en précisant le package VLE auquel il appartient (ici `"Omegasys"` *») et le nom du fichier vpz correspondant au modèle (ici* `"azodyn.vpz"`*. la fonction* `run` *permet de réaliser une simulation. Si comme ici, aucun autre argument que le modèle n'est précisé, le modèle va tourner avec les paramètres originaux tels que définis dans le fichier vpz. La fonction* `result` *permet de récupérer les résultats tels que définis sous G-VLE, ici une seule valeur en fin de simulation : le rendement* (`Yield`)*.*

Modifier par exemple les paramètres du modèle se fait ensuite très simplement en rajoutant la nouvelle valeur du paramètre en argument lors de l'appel de la

fonction `run` comme indiqué ci-dessous dans l'exemple. C'est ainsi que nous avons pu réaliser les analyses de sensibilité sur le rendement en faisant varier la valeur de nombreux paramètres du modèle AZODYN-Colza.

```
Simulation2 <- run(model.vle, condParam_Rendement.ANGPOT = 250,
                               condParam_Enracinement.Velong = 7);

Yield <- (results(Simulation2)[[1]][,"Omegasys:Eco2.rendement"]);
```

Extrait de code R montrant comment modifier la valeur de deux paramètres pour réaliser une simulation. Il faut connaître le nom du paramètre à modifier (ici `ANGPOT`*, le nombre de grains potentiel produit et* `Velong`*, la vitesse d'élongation racinaire) ainsi que les « chemins » associés, c'est à dire à quels modules ces paramètres appartiennent (ici* `condParam_Rendement` *et* `condParam_Enracinement`*).*

Annexe 5 : Résultats des analyses de sensibilité sur le modèle AZODYN-COLZA.

Par manque de place dans le corps principal, nous n'avons pas présenté les exemples de résultats d'analyse de sensibilité appliqués au modèle de culture du colza fonctionnant sur la plateforme RECORD-VLE : AZODYN-COLZA. Ces résultats n'ont ici qu'un but illustratif, ce qui explique qu'ils sont placés en annexe. L'analyse a été menée sur le rendement annuel en sélectionnant 16 paramètres parmi la cinquantaine inclus dans le modèle. Les simulations ont été obtenues à partir d'un jeu de données météorologiques correspondant au site d'Auzeville en 2004 (source NASA).

Résultat de l'analyse par la méthode de Morris :

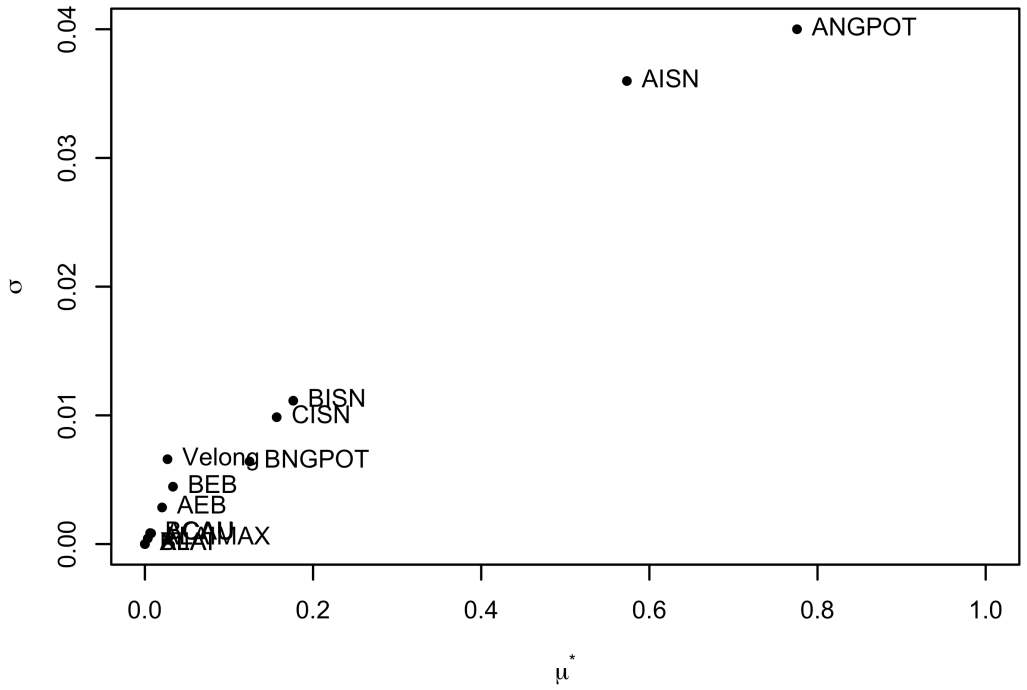

Figure Annexe 2 : *Résultats de l'analyse de sensibilité par la méthode de Morris sur le modèle AZODYN-COLZA.*

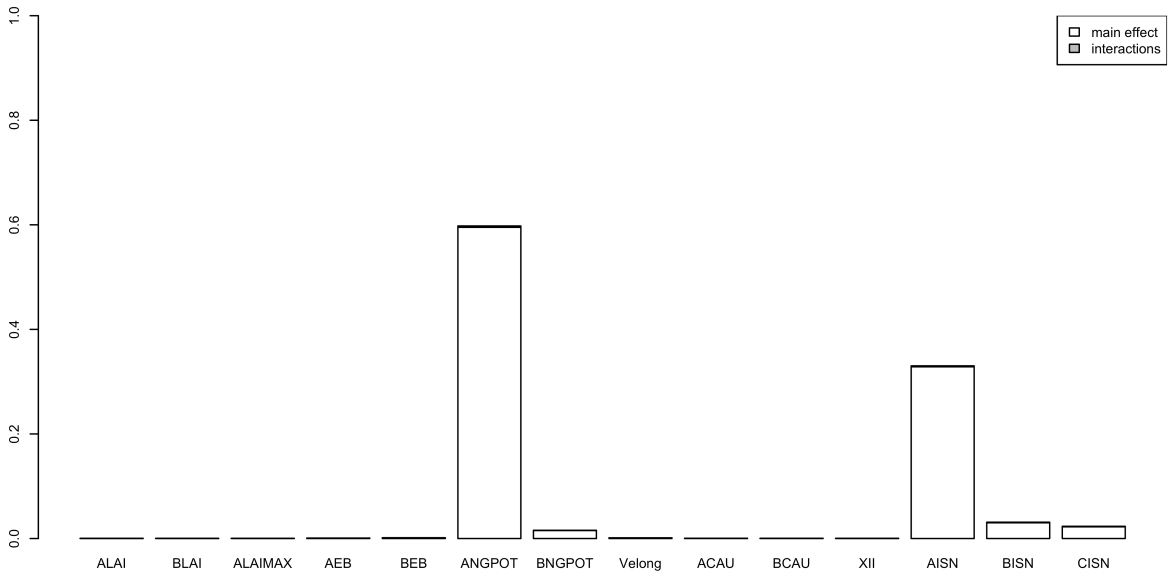

Figure annexe 3 : *Résultats de l'analyse de sensibilité par la méthode FAST appliquée au modèle AZODYN-COLZA. Pour chaque paramètre, la hauteur du diagramme en blanc indique la valeur de l'effet principal ; la hauteur en grisé représente la valeur de l'interaction*

Nous pouvons voir ici que les résultats des deux analyses vont dans le même sens, les paramètres ayant un effet important sur le rendement en graines de colza sont les paramètres ANGPOT qui intervient dans le calcul du nombre potentiels de grains pouvant être produits et AISN qui intervient dans le calcul du rendement. Nous pouvons noter qu'ici, aucune interaction n'est observée que ce soit lors de l'analyse de Morris (aucun paramètre n'est associé à une valeur de σ importante) ou lors de l'analyse FAST (où aucun paramètre ne présente une zone grisée importante sur le graphe Fig. Ann. 3).

RÉSUMÉ

Les modèles dynamiques sont largement utilisés dans la recherche en agronomie, mais la construction et l'analyse de tels modèles requièrent l'utilisation de méthodes mathématiques et statistiques qui manquent souvent aux modélisateurs. Le travail présenté ici a permis la création d'un package R permettant de mettre en œuvre différentes méthodes d'analyse appliquées à divers modèles agronomiques. Ce package par le biais de fonctions permet la réalisation facile d'un grand nombre de simulations ou encore la comparaison des sorties des modèles à des données réelles pour par exemple appliquer différentes méthodes d'analyse de sensibilité, d'estimation de paramètre ou d'évaluation de modèles. Le package ainsi créé permettra l'application de ces méthodes lors de formations organisées par le Réseau Modélisation et Agriculture (www.modelia.org) et permettra de servir d'outil à tous les modélisateurs en agronomie grâce à la documentation et les nombreux exemples associés à chaque fonction.

ABSTRACT

Dynamic models are vastly used in agronomy, but creating such models implies the use of mathematical and statistical methods with which researchers are mainly unfamiliar. The work presented here led to the creation of a R package allowing the use of several analysis methods applied to various examples of dynamic models for agriculture. This package contains many functions that allow running numerous model simulations or comparing model outputs to observed data in order to illustrate and implement several sensitivity analyzes, parameter estimations or model evaluation. The package created will allow the utilization of these methods during courses organized by French Networg for Modelling in Agriculture and can be used as a tool for all agronomy modellers thanks to the documentation and the examples associated with each function.