THESE

pour obtenir le grade de

DOCTEUR

de

l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Champ disciplinaire : Statistique

présentée par

Matieyendou LAMBONI

Analyse de Sensibilité pour les modèles dynamiques utilisés en agronomie et environnement.

Soutenue le 21 Décembre 2009 devant le jury composé de :

Jean Marc AZAIS	Professeur	Université P. Sabatier	Rapporteur
Robert FAIVRE	Directeur de Recherche	INRA	Rapporteur
Liliane BEL	Professeur	AgroParisTech	présidente
Jacques-Eric BERGEZ	Directeur de Recherche	INRA	Examinateur
Stéphanie MAHEVAS	ingénieur de Recherche	IFREMER	Examinateur
Hervé MONOD	Directeur de Recherche	INRA	Directeur
David MAKOWSKI	Chargé de Recherche	INRA	Co-directeur

MIA-Jouy ; Jouy-en-Josas ; Domaine Vilvert ; F 78350

A mon Père

Au souvenir de ma Mère

A mes Frères et Sœurs

A Patience Adjivon

Remerciements

Je tiens particulièrement à exprimer mon profond respect, ma grande reconnaissance et mes remerciements à mon directeur de thèse Hervé MONOD pour sa confiance, sa disponibilité malgré son emploi du temps chargé, ses conseils avisés, ses qualités humaines, son écoute et surtout son soutien. Ma reconnaissance, mon respect et mes remerciements vont aussi à mon co-directeur David MAKOWSKI, pour son regard critique, pour avoir guidé et orienté ma thèse et pour avoir porté une lecture attentive à mon mémoire.

Je remercie également les rapporteurs (Jean Marc AZAIS et Robert FAIVRE) pour leurs lectures attentives, leurs remarques pertinentes et les membres de Jury pour avoir eu l'amabilité de participer à la soutenance de ma thèse.

Je tiens aussi à remercier Bertrand IOOSS pour son soutien, son aide et sa profonde conviction de l'intérêt de l'analyse de sensibilité fonctionnelle ou dynamique. Je remercie aussi Daniel WALLACH pour ses conseils concernant la sélection des paramètres.

Mes remerciements vont également à Benoit GABRIELLE et à Simon LEHUGER pour leurs aides et conseils pour l'usage, la compréhension et l'interprétation des résultats du modèle CERES-EGC.

Mes remerciements s'adressent également au département Environnement & Agronomie et au département de Mathématiques et Informatiques Appliquées 'MIA) de l'INRA qui ont financé mes travaux.

Merci également à toute l'équipe de l'unité MIA (chercheurs et direction, secrétariat, doctorants, stagiaires) pour leur accueil, leurs contributions à la fois professionnelle et amicale.

Table des matières

Table des figures

Liste des tableaux

NOTATIONS

– **Les différents acronymes et abréviations utilisés dans cette thèse sont :**

ACP	Analyse en Composante Principale
ANOVA	ANalyse Of VAriance
AS	Analyse de Sensibilité
$\mathbb{C}$ov	Covariance
DOP	Décomposition Orthogonale Propre
DVS	Décomposition en Valeurs Singulières
$\mathbb{E}$	Espérance
EFAST	Extended Fourier Amplitude Sensitivity Test
$\mathbb{GSI}$	Generalized Sensitivity Index
$\mathbb{I}$	Inertie
$\mathbb{IMSEP}$	Integrated Mean Square Error of Prediction
IS	Indice de Sensibilité
LAI	Leaf Area Index
LARS	Least Angle Regression Stepwise
LASSO	Least Absolute Shrinkage and Selection Operator
LHS	Latin Hypercube Sampling
MANOVA	Multivariate ANalyse Of Variance
MCO	Moindre Carré Ordinaire
$\mathbb{MSE}$	Mean Square Error
$\mathbb{MSEP}$	Mean Square Error of Prediction
OGM	Organisme Génétiquement Modifié
PCA	Principal Component Analysis
SI	Sensitivity Index
TGSI	Total Generalized Sensitivity Index
TSI	Total Sensitivity Index
$\mathbb{T}$r	Trace
$\mathbb{V}$ar	Variance
WWDM	Wheat Winter Dynamic Model

– **Les principales notations mathématiques utilisées dans cette thèse sont :**

(N.0) $\mathcal{Z}(\Omega)$ l'espace d'état ou certain des facteurs incertains

(N.1) $\mathbf{Z} = \{Z^{(1)}, Z^{(2)}, \dots Z^{(d)}\}$ est le vecteur des d facteurs incertains

(N.2) $\mathbf{z}$ une réalisation de $\mathbf{Z}$

(N.3) Le facteur $Z^{(j)}$ a n_j modalités $\{z_1^{(j)}, z_2^{(j)}, \dots z_{n_j}^{(j)}\}$, $\forall j \in \{1, 2, \dots, d\}$ dans le cas discret

(N.4) $n = \prod_{j=1}^{d} n_j$

(N.5) $\mathbf{z}(n)$ une suite de points $(\mathbf{z}_1, \mathbf{z}_2 \dots \mathbf{z}_n)$ dans $\mathcal{Z}(\Omega)$

(N.6) $\{\mathbf{z}^{(j)}\}$ un ensemble dont les éléments sont les composantes du vecteur $\mathbf{z}^{(j)}$

(N.7) $u \in \{1, 2, \dots, d\}$ est un sous-ensemble de $\{1, 2, \dots, d\}$, $-u$ est le complémentaire de u dans $\{1, 2, \dots, d\}$ et $|u|$ est le cardinal de u

(N.8) $Z^{(u)} = \{z^{(j)}, j \in u\}$ désigne l'ensemble des facteurs dont les indices figurent dans u

(N.9) $Z^{(-u)}$ désigne l'ensemble des facteurs dont les indices ne sont pas dans u

(N.10) $z_u.i_u = \{z_{i_j}^{(j)}, i_j \in \{1, 2, \dots, n_j\}, j \in u\}$

(N.11) $< \bullet \mid \bullet >$ est le produit scalaire usuel associé à l'espace de Hilbert $\mathbb{R}^N$

(N.12) $\overset{\perp}{H}_E$ est l'opérateur de projection orthogonale sur le sous espace vectoriel E à l'aide du produit scalaire usuel.

(N.13) $f(z)$ est une fonction de $\mathbb{R}^d$ à valeur réelle.

(N.14) $y(z)$ est une fonction de $\mathbb{R}^d$ à valeur réelle.

(N.15) $\forall\, u \subseteq \{1, 2, \dots, d\}$ $\mathcal{G}_u$ est l'espace des fonctions de carré intégrable de la forme $f(z^{(u)})$. $f_u(z^{(u)})$ est une fonction de $\mathbb{R}^{|u|}$ et ne dépend que des facteurs dont l'indice j figure dans u.

(N.16) $\mathcal{G} = \left\{ f(\mathbf{z}) = \sum_{u \subseteq \{1,2,\dots,d\}} f_u(z^{(u)}) \,/\, f_u(z^{(u)}) \in \mathcal{G}_u \right\}$

(N.17) $< \bullet \mid \bullet >_{L^2([01]^d)}$ est le produit scalaire associé à l'espace de Hilbert des fonctions de carré réintégrable sur $[0\,1]^d$.

(N.18) $\mathbb{V}\mathrm{ec}\,\{\mathbf{v}\}$ est l'espace vectoriel engendré par le vecteur $\mathbf{v}$

(N.19) $\overset{\perp}{\bigoplus}$ est la somme directe orthogonale

(N.20) $\mathbb{I}_{\bullet\bullet} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

$$(N.21) \quad \mathbb{1}_{z_j.i_j\bullet} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \quad \leftarrow z_j.i_j\text{ième position} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$(N.22) \quad \mathbb{1}_{z_l.i_l z_r.i_r} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \quad \leftarrow z_l.i_l, z_r.i_r\text{ième position} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Remarquons que tous les facteurs d'entrées aléatoires sont notés en majuscule contrairement aux entrées supposées déterministes. Nous utiliserons ces notations tout au long de ce manuscrit. De plus, tout résultat déjà connu sera énoncé sous l'intitulé (Théorème, Proposition, Corollaire, Lemme) suivi des noms des auteurs de ce résultat, alors que les résultats nouveaux seront énoncés sous l'intitulé (Théorème, Proposition, Corollaire, Lemme) sans aucune précision supplémentaire.

Introduction

Contexte général

La modélisation des phénomènes environnementaux, écologiques, économiques, agronomiques, physiques et chimiques s'est beaucoup développée ces dernières décennies suite au progrès des connaissances et au développement des outils informatiques. La modélisation consiste à intégrer les connaissances acquises par l'expérimentation, l'expérience et la théorie sous forme d'équations mathématiques ou de codes de calculs. Les différents modèles développés ont très souvent un objectif d'appui à la recherche appliquée ou au conseil agricole, et ils peuvent dans certains cas servir d'outil d'aide à la décision pour des décideurs économiques ou politiques, à travers une meilleure compréhension du phénomène ainsi que la prédiction et la simulation des impacts de certaines décisions. De nombreux modèles sont développés à des échelles fines pour mieux représenter le phénomène et il devient parfois complexe et difficile d'interpréter les paramètres du modèle. Les phénomènes doivent souvent être représentés par des modèles dynamiques. Par exemple, en agronomie, les modèles sont développés très souvent à des pas de temps journaliers pour simuler les effets des pratiques agricoles sur les cultures (qualité et rendements), sur l'environnement (pollution et émission de gaz à effet de serre), ou sur les flux de gènes (propagation des gènes OGM). Certains de ces modèles sont utilisés pour guider les agriculteurs dans leurs pratiques agricoles et les décideurs politiques dans la gestion et la réglementation (Brisson *et al.* 1998 [26] ; Colbach *et al.* 2001 [38] ; Meynard *et al.* 2002 [104].

La modélisation des phénomènes naturels ou des procédés humains en général et des phénomènes agronomiques en particulier est entachée de plusieurs incertitudes qu'il convient de quantifier. Pour Oreskes (1994) [109], un phénomène naturel est si variant qu'il ne peut y avoir un seul modèle pour le représenter. Pour les modèles dynamiques complexes décrivant le mécanisme d'un phénomène donné, nous distinguons globalement

deux sources d'incertitudes principales : l'incertitude sur la structure du phénomène $f_0(\bullet)$ et l'incertitude sur les entrées du modèle (paramètres et variables d'entrée). Nous ne considérons dans ce mémoire que le deuxième type d'incertitude bien que cette dernière soit liée à la structure du modèle. Nous appelons dans la suite les variables d'entrée et/ou les paramètres incertains du modèle les facteurs.

De manière générale, plus le modèle intègrera des connaissances pertinentes, mieux il représentera le phénomène étudié pour un niveau de complexité donné. Nous nous intéressons dans ce mémoire aux méthodes statistiques et informatiques permettant d'identifier les connaissances pertinentes à inclure dans un modèle lors du processus de la modélisation. Les paramètres sont utilisés dans la formulation mathématique ou dans l'implémentation informatique des phénomènes et certaines questions se posent logiquement : quel sous-ensemble de paramètres est plus important pour mieux représenter le phénomène ? Comment réduire au maximum l'ensemble des paramètres d'un code de calcul sans trop altérer la représentation du phénomène étudié ? Quels sont les sous-ensembles de paramètres qui interagissent entre eux ? Quel sous-ensemble de paramètres faudra-t-il estimer pour se rapprocher au mieux des observations ? L'incertitude sur un facteur est due au fait que soit le facteur en question est de valeur mal connue, soit il est soumis à une variabilité intrinsèque, par exemple en fonction du site et de l'année (Wallach *et al.*, 2002 [157]). Dans les deux cas, il faut l'estimer (erreur d'estimation) ou le mesurer (erreur de mesure) et évaluer l'incertitude sur sa valeur. Dans la suite, l'incertitude sur un facteur sera décrite par un intervalle dans lequel est supposée se trouver la vraie valeur inconnue du paramètre ou par une distribution de probabilité.

Pour nous, un modèle désignera des équations mathématiques ou de façon équivalente un code de calcul qui est l'implémentation informatique des équations mathématiques ou des connaissances sur le phénomène dans un langage de programmation donné. Considérons un modèle dynamique représenté par l'équation mathématique suivante :

$$y(t) = f_0(\mathbf{x}, \theta, t), \tag{0.0.1}$$

où $\mathbf{x}$ est le vecteur des variables d'entrée du modèle, θ est le vecteur de paramètres incertains et $y(t)$ est la sortie du modèle à la date t, pour $t \in \{1, 2, \ldots, T\}$. La fonction $f_0(\bullet)$ représente le phénomène étudié et elle est soit déterministe soit stochastique. L'aspect stochastique du modèle ne sera pas abordé dans ce mémoire. Cependant, il est possible de le prendre en compte en faisant des répétitions (Ginot *et al.*, 2006 [61] ; Lurette *et al.*, 2009 [95]).

En ne considérant que les variables d'entrée et/ou les paramètres incertains du modèle que nous appelons facteurs et que nous notons $\mathbf{Z}$, le modèle de l'équation (0.0.1) s'écrit :

$$y(t) = f(\mathbf{z}, t), \tag{0.0.2}$$

avec $\mathbf{z}$ une réalisation des facteurs d'entrée du modèle $\mathbf{Z}$.
Dans ce mémoire, nous travaillerons conditionnellement aux variables d'entrée. Ainsi, les facteurs ne sont que des paramètres incertains.

En général, la représentation mathématique du phénomène est si complexe qu'il devient impossible d'identifier d'une manière simple et rapide, le sous ensemble de paramètres importants. C'est le cas de l'exemple de Campbell *et al.* (2006) [31] dont l'équation mathématique s'écrit :

$$f(x) = 10 + a \exp\left(-\frac{(x-b)^2}{K_1 a^2 + c^2}\right) + (b+d)\exp\left(K_2 a x\right).$$

Cette équation représente un modèle qui prend en entrée 6 paramètres $\theta = (a, b, c, d, k_1, k_2)$ et la variable d'entrée x qui est un angle polaire et appartient à l'intervalle $[-90, 90]$. Campbell *et al.* (2006) [31] ne considèrent dans leur étude que les 4 premiers paramètres. Une complexité de ce modèle réside dans sa sortie fonctionnelle qui rend difficile la détermination des facteurs clés par exemple.

La recherche du sous-ensemble de facteurs importants se pratique dans diverses disciplines scientifiques. Citons quelques exemples :

- Considérons le modèle complexe "E level" étudié dans Saltelli *et al.* (2000) [129] qui décrit le transfert des matières radioactives dans un milieu poreux et dont la fonction réponse simule la dose totale de radioactivité que reçoit une personne dans la biosphère. C'est un modèle dynamique à un pas de temps annuel et il inclut un certain nombre de facteurs. 12 facteurs incertains furent considérés dans les travaux de Saltelli *et al.* (2000) [129].

- Un exemple en finance serait le prix d'exercice d'une option (call ou put) américaine ou européenne issu du fameux modèle de Black and Scholes (1973). Le code de calcul simulant le prix d'exercice d'une option inclut beaucoup de paramètres et fournit une sortie dynamique. Constales *et al.* (2006) [39] effectuent l'analyse de sensibilité sur l'option call américain. Dans Saltelli (2008) [127], une analyse de sensibilité bien détaillée est effectuée pour un call européen.

– Un troisième exemple en épidémiologie animale a été abordé au cours de la préparation de cette thèse. Il s'agit du modèle de transmission de la salmonelle chez les porcs de la naissance jusqu'à l'abattage qui comporte 18 facteurs incertains (Lurette *et al.*, 2009 [95]). C'est un modèle dynamique d'appréciation quantitative du risque microbiologique pour un objectif de maîtrise et de gestion du risque sanitaire.

Pour ces différents modèles, la recherche d'un sous-ensemble de facteurs importants nécessite des outils statistiques élaborés. La notion d'importance considérée ici se définit par rapport à la variabilité des réponses du modèle et sera définie plus rigoureusement dans la Section 2.1 pour une sortie statique et la Section 2.3 pour une sortie dynamique.

Problématique

Le rôle croissant des modèles dynamiques ainsi que leur grande complexité rendent indispensables la mesure et la prise en compte des différentes sources d'incertitudes de ces modèles d'une part et la variabilité de leurs diverses composantes d'autre part lors de leur mise au point ainsi que lors de leur exploitation pour la prédiction et la préconisation. Deux sources de complexités inhérentes aux modèles dynamiques vont retenir notre attention :

– Les sorties des modèles dynamiques sont essentiellement des séries temporelles ou des courbes décrivant l'évolution d'un phénomène dans le temps, telles que l'évolution d'une grandeur agronomique entre le semis et la récolte à un pas de temps journalier (évolution journalière de la biomasse du blé dans Makowski *et al.*, 2004 [96] ; émission journalière du gaz N_2O issu des parcelles agricoles dans Gabrielle *et al.*, 2006a [58]). La structure dynamique de ces modèles introduit une forte corrélation temporelle entre les différentes sorties du modèle. Associer des paramètres à une seule sortie devient pertinent si la variable d'intérêt pour l'analyse est la variable de sortie en question (Homma *et al.*, 1996 [70]). Dans ce cas de figure, la dynamique du modèle et les différentes corrélations entre les sorties du modèle sont ignorées. Par contre, si le modèle a pour vocation de prédire les sorties à plusieurs dates (c'est le cas de certains modèles dynamiques tels que les modèles de culture et agro-système de la Section 1.2.2), il est intéressant que toute analyse ou prise de décision sur l'importance d'un paramètre se fasse au moins à l'aide de toutes les sorties qui dépendent de ce paramètre. Du moment où il est difficile voire impossible d'identifier toutes les sorties qui font appel à un paramètre donné et étant donné qu'il peut exister un ou plusieurs paramètres qui gouvernent toute la dynamique du modèle, il est

intéressant de considérer toutes les sorties du modèle et d'intégrer les corrélations qui décrivent la dépendance entre les sorties du modèle dans toutes les analyses et notamment en analyse de sensibilité. Ce qui n'est possible qu'en considérant toutes les sorties du modèle.

– Par souci de modélisation plus fine pour approcher au mieux le phénomène étudié, les modèles dynamiques déterministes incluent généralement de nombreux paramètres incertains qui constituent l'une des principales sources d'incertitudes dans les sorties de modèles. L'estimation des paramètres du modèle est une étape cruciale dans le processus de la modélisation dans le sens où la performance du modèle dépend largement de l'exactitude des estimations (Butterbach-Bach *et al.*, 2004 [29] ; Gabrielle *et al.*, 2006a [58] ; Lehuger *et al.*, 2009 [93]). L'estimation des paramètres de modèles aussi complexes que certains modèles dynamiques (non linéaires en majorité) constituent un problème aussi bien en théorie qu'en pratique soit par manque d'observations pour estimer tous les paramètres (Brun *et al.*, 2001 [28] ; Tremblay, 2004 [151] ; Bechini *et al.*, 2006 [18]), soit par la présence de paramètres non identifiables statistiquement (Brun *et al.*,2001 [28]), soit par la structure du modèle (essentiellement la non régularité).

En général, le grand nombre de paramètres incertains dans un modèle ne s'accompagne malheureusement pas d'un grand nombre d'observations pour des raisons de coûts et de difficultés de mesures. Une estimation des paramètres d'un modèle avec peu d'observations est moins précise et augmente les risques d'erreurs de prévision lorsque l'on se sert de ce modèle. Une approche naturelle serait la statistique bayésienne mais cette approche reste difficile à mettre en œuvre sur les modèles dynamiques complexes en général (Brun *et al.*, 2001 [28]) et les modèles de culture en particulier (Wallach *et al.*, 2001 [156], Wallach *et al.*, 2002 [157]). Même si ce domaine évolue rapidement, l'approche bayésienne est restreinte aux modèles à faible coût de simulations avec un nombre de paramètres relativement petit. Il est aussi connu que l'estimation bayésienne en présence d'un petit nombre d'observations peut fournir des distributions a posteriori moins précises (Lehuger *et al.*, 2009 [93]). L'analyse de sensibilité (voir Section 2.1) offre un moyen de trancher cette problématique dans le cas d'un modèle à sortie scalaire (Perrin *et al.*, (2001) [114] ; Wallach *et al.*, 2002 [157] ; Tremblay 2004, [151] ; Makowski *et al.*, 2006b [97] ; Monod *et al.*, 2006 [105]). En effet, l'AS permet d'identifier les paramètres les plus importants à l'aide de l'expérimentation virtuelle du modèle sans utiliser d'observations. Cependant, l'utilisation de l'analyse de sensibilité pour sélectionner les paramètres à estimer reste une approche intuitive et nécessite une investigation approfondie.

Objectif de la thèse

La pertinence des approches de modélisation nécessite le développement de méthodes performantes pour mieux comprendre le comportement des modèles par rapport aux incertitudes des facteurs et pour mesurer l'inadéquation entre les sorties du modèle et les observations disponibles. L'un des principaux objectifs dans le processus de la modélisation est d'arriver à obtenir un modèle de sorte que les sorties de ce dernier se rapprochent au mieux des observations. Cette procédure de validation de modèles est souvent faite à l'aide de plusieurs critères statistiques tels que le $\mathbb{MSE}$, le $\mathbb{MSEP}$ pour les modèles dynamiques.

Face au manque d'observations pour estimer tous les paramètres des modèles dynamiques, une pratique courante dans la littérature consiste à sélectionner les paramètres à estimer à l'aide de l'AS (Ruget *et al.*, 2002 [122] ; Brun *et al.*, [27] ; Tremblay 2004, [151] ; Makowski *et al.*, 2006a [97] ; Wallach *et al.*, 2006 [158]). Notons que la procédure de sélection de paramètres les plus influents par analyse de sensibilité est basée sur les données simulées et se fait indépendamment de la procédure de validation du modèle qui est basée sur les données réelles. Intuitivement, ces deux procédures semblent être liées mais nécessitent une investigation formelle un peu plus poussée.

Cette thèse a pour ambition d'évaluer la pratique courante des modélisateurs en établissant une relation entre les indices de sensibilité des paramètres ou facteurs d'un modèle dynamique et les critères d'évaluation de modèles, notamment le $\mathbb{MSE}$ et le $\mathbb{MSEP}$, bien adaptés aux modèles dynamiques non linéaires. Une telle relation n'est possible que si l'on dispose d'une méthode d'AS qui nous fournit un unique indice par facteur pour les modèles dynamiques. Cette thèse contribue donc également au développement d'une méthode d'analyse de sensibilité à valeur générique qui prenne en compte non seulement l'aspect dynamique et les corrélations induites par cette dynamique, mais aussi et de façon simultanée toute la gamme de variabilité des facteurs du modèle. Cette méthode devra satisfaire les cinq principales propriétés d'une méthode d'analyse de sensibilité dite globale présentées dans la Section 2.1.1 et surtout nous fournir un indice synthétique qui mesurera l'influence de chaque facteur sur toute les sorties d'un modèle dynamique.

Organisation de la thèse

Les Chapitres 1 et 2 sont essentiellement bibliographiques alors que les Chapitres 3, 4 et 5 décrivent des travaux de recherche originaux.

Le Chapitre 1 précise le contexte de la modélisation dynamique en agronomie et en environnement. Ce chapitre décrit, en particulier des modèles dynamiques sur lesquels seront appliquées les différentes méthodes développées dans cette thèse. Nous nous intéressons aussi aux différentes méthodes statistiques qui interviennent dans les différentes étapes de la modélisation c'est-à-dire de la caractérisation du phénomène à modéliser jusqu'à l'évaluation du modèle.

Le Chapitre 2 est consacré à une synthèse originale de l'état de l'art des méthodes de l'analyse de sensibilité classiques dites globales, du lien entre l'analyse de sensibilité et l'estimation des paramètres et de l'analyse statistique multivariée.

Le lien formel entre les indices de sensibilité et les critères d'évaluation des modèles $\mathbb{MSE}$, $\mathbb{MSEP}$ est traité dans le Chapitre 3, dans le cas d'un modèle linéaire. Le problème avec les modèles dynamiques non linéaires complexes est l'impossibilité d'avoir des expressions explicites des estimateurs des paramètres, des critères de validation de modèles et des indices de sensibilité. Ce qui rend difficile toute tentative de formulation formelle du lien entre les indices de sensibilité et le $\mathbb{MSEP}$. Dans ce chapitre, nous formalisons la pratique adoptée par les modélisateurs et dégageons les principales hypothèses pour établir cette relation en se servant de la mesure de sensibilité basée sur la variance ("variance based") et des critères classiques d'évaluation de modèles notamment $\mathbb{MSE}$ et $\mathbb{MSEP}$.
Le Chapitre 4 présente le développement d'une méthode d'analyse de sensibilité multivariée pour les modèles dynamiques à l'aide de la décomposition de l'inertie. Nous considérons dans un premier temps les différents facteurs de l'analyse de sensibilité comme des facteurs qualitatifs ou quantitatifs discrets. Cette considération correspond selon les cas à la véritable nature des facteurs ou à une approximation pouvant se justifier par un lien avec des protocoles expérimentaux ou par un choix arbitraire de privilégier certains niveaux dans l'analyse. Ensuite, l'hypothèse de facteurs discrets sera relâchée pour permettre de balayer toute la gamme d'incertitudes des différents facteurs. Les facteurs peuvent alors prendre toutes les valeurs possibles dans leurs intervalles d'incertitudes. Ces facteurs continus permettront aussi de prendre en compte l'incertitude de façon plus fine à travers leurs lois de distributions.

Le Chapitre 5 de ce mémoire analyse le lien qui existe entre les indices de sensibilité et le $\mathbb{MSEP}$ du chapitre 3 dans le cas d'un modèle dynamique non linéaire (CERES-EGC) à l'aide des indices de sensibilité développés dans le chapitre 4. Dans ce chapitre, nous montrons une relation empirique entre les indices de sensibilité et le $\mathbb{MSEP}$ en utilisant le modèle CERES-EGC. Le chapitre 5 est aussi une illustration de la méthode de

l'analyse de sensibilité (AS) développée dans le chapitre 4 sur deux modèles dynamiques agronomiques : le modèle de culture WWDM décrit dans la Section 1.2.2 et le modèle agro-système CERES-EGC présenté dans la Section 1.2.2

Chapitre 1

Modélisation dynamique en agronomie et environnement

Face aux aléas des phénomènes naturels, il est nécessaire d'utiliser la modélisation pour prévoir et tenter de maitriser le risque à travers des prises d'actions concrètes. La nature cyclique et l'aspect dynamique des phénomènes naturels en général et des phénomènes agronomiques et environnementaux en particulier conduisent au développement de modèles dynamiques. Ces modèles simulent généralement des grandeurs agronomiques et environnementales à des pas de temps donnés.

En agronomie, le développement de modèles dynamiques est d'une grande importance du fait que ces derniers servent i) d'outils d'évaluation des impacts des différentes interventions ou pratiques agricoles sur les sorties d'intérêt (effets des pratiques agricoles sur l'environnement par exemple) ii) d'outils de prévision des variables d'intérêt (la qualité et la quantité des récoltes, émission de N_2O à certaines dates) iii) d'outils d'aide à la décision (les doses et les dates d'apport d'engrais). Les modèles dynamiques sont des outils pertinents de la recherche en agronomie et en environnement dans la mesure où ils permettent de faire de nombreuses simulations à faible coût. Les diverses utilisations des modèles de culture sont décrites dans Boote *et al*. (1996) [23], Tremblay (2004) [151] et les références qui se trouvent dans ces documents.

Dans ce chapitre, nous présentons les modèles dynamiques qui sont basés sur les modèles de culture et qui serviront d'exemples d'application dans toute la suite de ce mémoire. Nous considérons également les principales étapes de la modélisation afin de replacer ce mémoire dans le contexte général de la modélisation.

1.1 Principales étapes de la modélisation

L'intégration des connaissances acquises sur un phénomène donné dans un modèle nécessite plusieurs étapes principales : intégration des connaissances sous forme d'équations ou code de calcul ; analyse de sensibilité, vérification de l'identifiabilité du modèle ; estimation et évaluation du modèle. Chacune de ces différentes étapes a déjà fait l'objet de plusieurs travaux (Wallach *et al.*, 2006 [158]). Dans cette section, nous nous limitons à faire une bibliographie des méthodes statistiques rentrant dans chacune des étapes sans prétendre être exhaustif. Nous considérons ici les méthodes les plus utilisées en modélisation dynamique.

1.1.1 Formalisme mathématique des connaissances

Tout processus de modélisation quantitative inclut une représentation mathématique des connaissances dont on dispose sur le phénomène à étudier. C'est une représentation simplificatrice mais aussi fidèle que possible de la réalité. Notons que le "vrai modèle" est inconnu et que nos connaissances sont réductrices du phénomène. Précisons que cette étape est très importante dans la mesure où elle détermine le niveau de complexité du modèle. De plus, elle est délicate à mettre en œuvre et nécessite l'avis de plusieurs experts du phénomène étudié. Cette étape introduit une approximation du "vrai modèle" inconnu et engendre des incertitudes sur le modèle notamment l'incertitude de structure du modèle (Brun *et al.*, 2001 [28] ; Goldstein (2006)[62]), l'incertitude sur l'implémentation du modèle ou de manière générale l'incertitude épistémique (Jacques, 2005 [80] ; Da-Veiga, 2007 [43]).

1.1.2 Analyse de sensibilité

L'analyse de sensibilité s'impose comme une étape à part entière dans le processus de modélisation pour des raisons évoquées dans la Section 2.1 du Chapitre 2 notamment l'identification des sources d'incertitude les plus importantes. L'AS permet de guider les experts sur le choix des parties du modèle à approfondir juste après la première étape pour simplifier le modèle ou pour réduire au maximum l'erreur de simplification du modèle. Elle intervient aussi au niveau des étapes d'identification, du choix du modèle et de l'estimation des paramètres (Brun *et al.*, 2001 [28] ; Brun *et al.*, 2002 [27] ; Tremblay, 2004 [151]).

1.1.3 Vérification de l'identifiabilité des paramètres du modèle

L'identifiabilité des paramètres d'un modèle donné est une étape cruciale (Brun *et al.*, 2001, [28]) dans la mesure où c'est l'une des hypothèses dans la modélisation statistique

qui assure la consistance des estimateurs des paramètres. En statistique mathématique, un modèle statistique est identifiable si :

$$f(\theta) = f(\theta') \Longrightarrow \theta = \theta', \tag{1.1.1}$$

avec $f(\theta)$ (resp.$f(\theta')$) la distribution de la fonction réponse quand le vecteur de paramètres est θ (resp. θ').

Notons qu'un modèle est dit identifiable si pour tout couple de vecteurs différents de valeurs des paramètres conduisent à des sorties différentes.

Jackerman et Homberger (1993) [77] proposent de réduire le modèle initial dans le but d'obtenir un nouveau modèle pour lequel les paramètres sont identifiables Cette méthode peut conduire soit à un modèle boite noire en cherchant l'identifiabilité au détriment de la compréhension du phénomène. Une méthode prometteuse est l'approche bayésienne qui ne demande pas à ce que les paramètres soient identifiables avant de les estimer (voir Brun *et al.*, 2001 [28] et les références qui s'y trouvent). Parmi ces approches nous avons la méthode GLUE (Generalized Likelihood Uncertainty Estimation) dans Beven (1992) [22], la méthode Monte Carlo Markov Chain (MCMC), la méthode Monte Carlo Filtering (MCF) dans Beck (1987) [16] .... Les applications courantes de l'approche bayésienne restent limitées aux modèles à faible coup de simulations et en grande dimension, elles souffrent du fléau de la dimension.

1.1.4 Choix de modèle et évaluation des modèles

Cette étape est cruciale et indispensable à tout type de modélisation. L'estimation des paramètres des modèles de culture fut l'objet de la thèse de Tremblay (2004) [151]. et les références qui se trouvent dans cette thèse. Les différentes méthodes d'estimation des modèles linéaires et non linéaires se trouvent dans Seber (1977) [135], Seber and Wild (1989) [136], Gallant (1987) [59], Azaïs et Bardet (2005) [15]. Dans le cas des modèles linéaires ou non linéaires, le choix de modèle se fait souvent avant l'estimation des paramètres et l'estimation des paramètres est fortement conditionnée par le modèle retenu. Par contre, La méthode LASSO (Tibshirani, 1996) [150] ou LAR (Zou *et al.*, 2005 [163] contribue simultanément à la sélection et à l'estimation des paramètres. Ces méthodes ont les propriétés de parcimonie du fait qu'elles annulent certains coefficients. La description détaillée de cette procédure est présentée dans l'annexe 2 car elle est spécifique aux modèles linéaires sparses.

Différentes méthodes de sélection de modèles furent considérées dans Georges (2000) [60], Muller (2002) [107], Draper et Smith (1981) [46], Yang (2007) [161]. Leeb (2008) [92] fournit un aperçu global de la vaste bibliographie des méthodes de sélection de modèle. Ces méthodes de sélection peuvent être regroupées en 3 grandes procédures de sélection :

⟶ Procédure de sélection basée sur les critères d'information

L'un des objectifs de la modélisation étant de chercher un modèle qui minimise une mesure de risque ou une fonction perte notamment le Mean Square Error ($\mathbb{MSE}$) ou le Mean Square Error of Prediction ($\mathbb{MSEP}$), il est naturel de sélectionner le modèle qui minimise ces critères. Pour être concis, nous ne considérons que les critères généralement utilisés par les modélisateurs et adaptés aux modèles dynamiques :

- le coefficient de détermination (R^2).

le R^2 fut utilisé comme critère de sélection de modèle par Theil (1961) [149]. Le R^2 mesure la part de variance expliquée par le modèle. C'est une mesure de la qualité d'approximation du phénomène par un modèle donné. Pour un modèle dynamique, le R^2 dynamique est calculé à différentes dates et donne une idée de la qualité des approximations faites sur le modèle (Homma *et al.*, 1996 [70] ; Lamboni *et al.*, 2008 [91]). Le coefficient de détermination est défini par :

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{1.1.2}$$

où y_i est la ième observation et $\hat{y}_i$ son estimation ou la valeur de la fonction réponse. $\bar{y}$ est la moyenne des observations.

Notons que le R^2 se calcule facilement mais qu'il n'est pas souvent utilisé pour comparer le pouvoir explicatif des modèles dans le mesure où l'objectif des chercheurs ne se ramène pas à maximiser ce critère qui augmente mécaniquement avec le nombre de paramètres.

- le Mean Square Error ($\mathbb{MSE}$)

Définition 1.1.1 *Le $\mathbb{MSE}$ mesure la qualité de l'estimation du modèle et est défini par*

$$\mathbb{MSE} = \mathbb{E}_{\mathbf{X}} \left[\left(y - f(\mathbf{X}, \hat{\theta}) \right)^2 \right], \tag{1.1.3}$$

où $\hat{\theta}$ est l'estimateur du vecteur de paramètres et y est la valeur observée de la fonction réponse.

Un estimateur standard du $\mathbb{MSE}$ est :

$$\hat{\mathbb{MSE}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\mathbf{x_i},\hat{\theta})\right)^2, \tag{1.1.4}$$

avec n le nombre total des observations utilisées pour estimer le vecteur de paramètres.

Notons que la définition 1.1.1 du $\mathbb{MSE}$ est une distance (norme L_2) entre y et son estimateur $f(\mathbf{x},\hat{\theta})$. Un bon modèle devra minimiser cette distance. En prenant la norme L_1 , nous obtenons le critère Mean Absolute Error qui n'amplifie pas les écarts entre les observations et les sorties de la fonction de réponse.

- le Mean Square Error of Prediction ($\mathbb{MSEP}$)

Définition 1.1.2 *Le $\mathbb{MSEP}$ est une mesure de la qualité de prédiction d'un modèle et est défini par :*

$$\mathbb{MSEP} = \mathbb{E}_{\mathbf{X}^*,\hat{\theta}}\left[\left(y^* - f(\mathbf{X}^*,\hat{\theta})\right)^2\right],$$

où $f(\mathbf{x}^*,\hat{\theta})$ représente la prédiction du modèle et y^* l'observation de la grandeur modélisée lorsque le vecteur de variables vaut $\mathbf{x}^*$.
Dans le cas d'un modèle dynamique, le $\mathbb{MSEP}$ intégré $\mathbb{IMSEP}$ est défini par :

$$\mathbb{IMSEP} = \mathbb{E}_{\mathbf{X}^*,\hat{\theta}}\left\{\int_{[0\,T]}\left[\left(y^*(t) - f(\mathbf{X}^*,\hat{\theta},t)\right)^2\right]dt\right\}, \tag{1.1.5}$$

Un estimateur du $\mathbb{MSEP}$ est le $\hat{\mathbb{MSE}}$ mais le calcul s'éffectue avec des observations n'ayant pas servi pour estimer les paramètres. Les équivalents de ces critères dans le cas particulier d'un modèle linéaire sont présentés dans l'annexe 1.

Remarque 1.1.1 *Ces différentes méthodes de sélection (section 1.1.4) comportent un biais : le biais de sélection (Muller 2002 [107]) du fait que les même données servent à la fois à la sélection de variables et aux estimations des paramètres du modèle.*

$\longrightarrow$ Validation croisée

La validation croisée ($\mathbb{CV}$) (Allen, 1971 [5] ; Allen, 1974 [6] ; Stone, 1974 [144]) permet d'évaluer la qualité prédictive d'un modèle en évitant d'utiliser les mêmes données pour estimer les paramètres du modèle et évaluer le modèle. Le principe de la validation croisée consiste à subdiviser les données initiales (D) en g groupes ($D_i, i \in \{1,2,\ldots,g\}$) de même taille approximativement. Ensuite il faudra utiliser une partie des données (les données des $g-1$ groupes) pour estimer les paramètres et les données complémentaires pour quantifier

la qualité d'ajustement à l'aide d'un critère notamment le $\mathbb{MSEP}$. L'algorithme suivant décrit cette procédure :

Algorithme 1.1.1 *Validation croisée*

Etape 1 :	*subdiviser les données* $D = \cup_{i=1}^{g} D_i$
Etape 2 :	*estimer les paramètres à l'aide des données* $D_{-i_1} i_1 \in \{1, 2, \ldots, g\}$ *avec* $D_{-i_1} = D \setminus D_{i_1}$
Etape 3 :	*quantifier la qualité d'ajustement* $\mathbb{MSEP}(D_{i_1})$ *à l'aide de données* D_{i_1}
Etape 4 :	*répéter les étapes 2 et 3 pour* i_1 *parcourant l'ensemble* $\{1, 2, \ldots, g\}$
Etape 5 :	*Calculer le critère moyen* $\mathbb{CV} = \frac{1}{g} \sum_{i_1=1}^{g} \mathbb{MSEP}(D_{i_1})$
Etape 6 :	*sélectionner le modèle qui a le plus petit* $\mathbb{CV}$

La validation croisée est largement utilisé (YANG, 2007 [161]) et elle s'applique aussi bien aux modèles paramétriques que non paramétriques, aux modèles linéaires que non linéaires. Les propriétés théoriques de consistance (conditions pour assurer la convergence en probabilité de sélectionner le meilleur modèle) sont discutées dans Shao (1997) [138], Shao (1993) [137] pour la régression linéaire et dans Yang (2007) [161] pour l'estimation non paramétrique. La thèse de Whittaker (2003) [160] est entièrement consacrée à la validation croisée avec une revue bibliographique des différentes versions de la validation croisée dans le Chapitre 2. Une brève revue sur la validation croisée généralisée se trouve également dans YANG (2007) [161]

La sélection par validation croisée nécessite la disponibilité d'un grand nombre d'observations afin de pouvoir subdiviser ces observations en g groupes. De plus, dans le cas des modèles sur-paramétrés, les données disponibles sont insuffisantes pour estimer tous les paramètres. Il est coûteux et parfois impossible d'augmenter le nombre de données.

Remarque 1.1.2 *Dans le cas où chaque groupe* D_i *contient une seule donnée alors le nombre de groupe* g *est égal au nombre d'observations* n*. La validation croisée dans ce cas est connue sous le nom de "Leave One Out (LOO)".*

⟶ Procédure de sélection basée sur les tests statistiques

Considérons simplement deux modèles candidats M_1 et M_2 emboités c'est - à - dire $M_1 \subseteq M_2$ (modèle M_1 est un sous modèle du modèle M_2 ou bien le modèle M_1 est une simplification du modèle M_2) et supposons sans perdre de généralité que le gros modèle M_2 est le vrai modèle. La sélection de modèle dans ce cadre peut se faire à l'aide d'un test statistique :

$$\begin{cases} Ho : M_1 \text{ est le vrai modèle ou a la bonne distribution} \\ H1 : M_2 \setminus M_1 \text{ est le vrai modèle} \end{cases} \tag{1.1.6}$$

Soient $\mathcal{D}$ la statistique de test et $\mathcal{R}$ le domaine de rejet de l'hypothèse nulle Ho. La sélection du bon modèle $\hat{M}$ est donnée par :

$$\hat{M} = \begin{cases} M_1 \text{ si } \mathcal{D} \in \mathcal{R} \\ M_2 \text{ sinon.} \end{cases} \tag{1.1.7}$$

Notons que les tests de rapport de vraisemblance, de Fisher contraint, de Wald largement utilisés en économétrie s'inscrivent dans cette procédure de sélection de modèle. Remarquons aussi que ce test peut s'étendre à d'autres critères tels que le $\mathbb{MSE}$ ou le $\mathbb{MSEP}$. L'hypothèse nulle Ho deviendra : $\mathbb{MSEP}_{M_1} = \mathbb{MSEP}_{M_2}$ ou $\mathbb{MSEP} < s$ avec s le seuil maximum d'erreur que l'on s'autorise. Pour les différentes méthodes de sélection basées sur les tests statistiques voir Anderson (1962, 1963) [8], [10] ; Leeb (2008) [92].

Les inconvénients majeurs des tests statistiques sont la difficulté de déterminer la distribution de la statistique du test, la difficulté de spécifier les hypothèses du test, le rejet systématique de l'hypothèse nulle Ho quand la taille de l'échantillon devient grand.

Les différentes méthodes de sélection de modèles présentées dans cette section sont entachées du biais de sélection de paramètres du fait que les mêmes données sont utilisées pour sélectionner et estimer les paramètres à la fois. De plus, ces méthodes d'estimation et de sélection nous donnent de bons estimateurs en terme de précision si i) nous disposons d'observations suffisantes pour estimer tous les paramètres ii) le modèle ne contient qu'un nombre raisonnable de paramètres. Il est généralement reconnu qu'estimer un grand nombre de paramètres (par exemple 10) avec peu d'observations (20 par exemple) ne garantit pas une bonne précision des estimateurs.

1.2 Modèles dynamiques basés sur les modèles de culture

1.2.1 Structure des modèles de culture

Les modèles de culture sont une représentation du système plante-sol dans son environnement physique et technique (Wallach, 2006 [158]). Ces modèles décrivent le déve-

loppement d'une culture (blé, colza, soja, ...) en parcelles agricoles sur un pas de temps généralement journalier, en intégrant des variables pédoclimatiques, des changements de phases et les effets d'interventions techniques. Les modèles de culture sont en majorité des modèles déterministes.

La complexité des modèles de culture provient des fonctions décrivant le phénomène, de la nature dynamique des sorties et du nombre de paramètres incertains (pouvant aller jusqu'à plusieurs centaines). En effet, pour mieux décrire le mécanisme des phénomènes étudiés et s'approcher au mieux de ces derniers, les modélisateurs sont amenés à introduire de nombreux paramètres dans les équations mathématiques ou dans le code de calculs modélisant le phénomène.

Les différents modèles de culture sont structurés en modules et principalement, nous distinguons quatre grands modules qui interagissent fortement entre eux :

- le module plante : ce module a pour ambition de représenter au mieux le système complexe de fonctionnement physiologique d'une plante. Il vise à intégrer les connaissances acquises sur la physiologie des plantes dans des équations mathématiques ou code de calculs. Ce module contribue par exemple à modéliser la croissance de la plante, la floraison etc.

- le module sol : la typologie du sol est un élément clé dans le développement des plantes du fait que les plantes puisent leurs ressources minérales et organiques dans le sol. Le module sol modélise le processus d'échange de flux entre la plante et le sol en se basant sur les caractéristiques du sol (composition, profondeur, etc). Ce module intègre parfois et de plus en plus "la composante environnement du sol" dans le but d'évaluer l'impact des pratiques agricoles sur l'environnement (transformation de l'azote par les microorganismes en N_2O, reliquat d'azote à la récolte, etc).

- le module atmosphère : la description du processus d'échange entre la partie aérienne de la plante et le milieu aérien est assurée par ce module notamment les échanges gazeux (CO_2, N_2O) liés à la respiration, la transpiration et la photosynthèse en fonction du rayonnement, de la température et de l'humidité.

- le module décrivant les effets des facteurs limitants : les modules précédents décrivent le fonctionnement de la plante dans des conditions normales de température, de disponibilité d'eau et de ressources. La description de l'effet des facteurs limitants

permet de prendre en compte les conditions réelles de culture telles que le stress hydrique ou thermique par exemple en contraignant les différents processus du fonctionnement normal de la plante.

1.2.2 Exemples de modèles dynamiques

Le modèle WWDM

Le modèle WWDM (Winter Wheat Dynamic model) est un modèle de culture simple qui simule la biomasse du blé d'hiver à un pas de temps journalier (Baret et Guyot, 1991 [17], Makowski *et al.*, 2004 [96]). Le modèle WWDM est composé de deux variables d'états : la biomasse cumulée $U(t)$ et l'indice de surface foliaire $\mathrm{LAI}(t)$. Le calcul de la biomasse s'effectue quotidiennement en fonction de la température cumulée (en degré jour : $^{\circ}$C/jour) en partant de 0 et du rayonnement photosynthétique actif mesuré quotidiennement. Les sorties du modèle s'étendent du semis ($t = 1$) jusqu'à la récolte. Le modèle est défini par les deux équations suivantes :

$$U(t+1) = U(t) + E_b E_{i\max} \left[1 - e^{-K.\,\mathrm{LAI}(t)}\right] \mathrm{PAR}(t) + \varepsilon(t),$$

et

$$\mathrm{LAI}(t) = L_{\max} \left\{ \frac{1}{1 + e^{-A(T(t)-T_1)}} - e^{B(T(t)-T_2)} \right\},$$

où $\mathrm{PAR}(t)$ est le rayonnement photosynthétique quotidien ; $\varepsilon(t)$ est un terme aléatoire d'espérance nulle représentant l'erreur de modélisation. $\varepsilon(t)$ correspond à l'inadéquation entre la valeur de la biomasse simulée et la "vraie" biomasse. Le modèle tel qu'il est défini est un modèle stochastique mais dans la suite de ce mémoire, par souci de simplicité, nous ne considérerons que la partie déterministe du modèle. La biomasse au semis ($t = 1$) est logiquement nulle($U(1) = 0$). En outre, la contrainte,

$$T_2 = \frac{1}{B} \log[1 + \exp(A \times T_1)],$$

est appliquée, de sorte que $\mathrm{LAI}(1) = 0$ au début.

Le modèle WWDM inclut 7 paramètres incertains qui prennent leurs valeurs dans des intervalles. La plupart d'entre eux ont une interprétation agronomique ou biologique et les différents paramètres incertains sont présentés dans la Table 1.1. En plus des paramètres, le modèle WWDM prend en entrée deux principales variables : le climat ou plus précisément la température (T) et le rayonnement (PAR).

Quelques simulations de l'évolution dynamique de la biomasse du semis à la récolte pour un scénario climatique sont présentées dans la Figure 1.1.

Paramètre	Interprétation	Valeur nominale	Intervalle d'incertitude
E_b	coefficient de conversion du rayonnement intercepté en biomasse (gm^{-2})	1.85	0.9-2.8
$E_{i\max}$	rapport maximal entre rayonnements intercepté et incident	0.94	0.9-0.99
K	coefficient d'extinction	0.7	0.6-0.8
$L_{\max}$	valeur maximale du LAI	7.5	3-12
T_1	seuil de température ($^\circ$C)	900	700-1100
A	coefficient de croissance du LAI	0.0065	0.0035-0.01
B	coefficient de décroissance du LAI	0.00205	0.0011-0.0025

TABLE 1.1 – Intervalles d'incertitudes des différents paramètres du modèle WWDM.

Le modèle AZODYN

Le modèle AZODYN (Jeuffroy et Recous, 1999 [81]) est un modèle qui simule la culture du blé et fut développé pour guider les agriculteurs dans leur itinéraire technique de fertilisation et notamment dans les dates d'apport d'engrais. Les objectifs du modèle AZODYN sont d'élaborer des stratégies de fertilisation dans un contexte pédoclimatique donné qui répondent aux objectifs de rendement, de teneur en protéines des grains et de préservation de l'environnement.

Le modèle intègre essentiellement trois modules : le module sol pour la description du sol et des composés organiques et minéraux disponibles dans le sol ; le module physiologique décrivant dans des conditions optimales la plante à travers l'indice foliaire, la biomasse, la quantité d'azote accumulée par les plantes. Le sous-module rendement décrit tous les processus permettant non seulement de quantifier le nombre de grains par m^2 mais aussi la teneur en protéines et le poids des grains. Ces différents modules permettent à AZODYN de prendre en compte les caractéristiques du sol (sol argileux, épaisseur de couche labourée), la teneur en composés organiques et minéraux (carbonate de calcium, azote, gaz carbonique), la densité apparente et de certaines variables climatiques telles

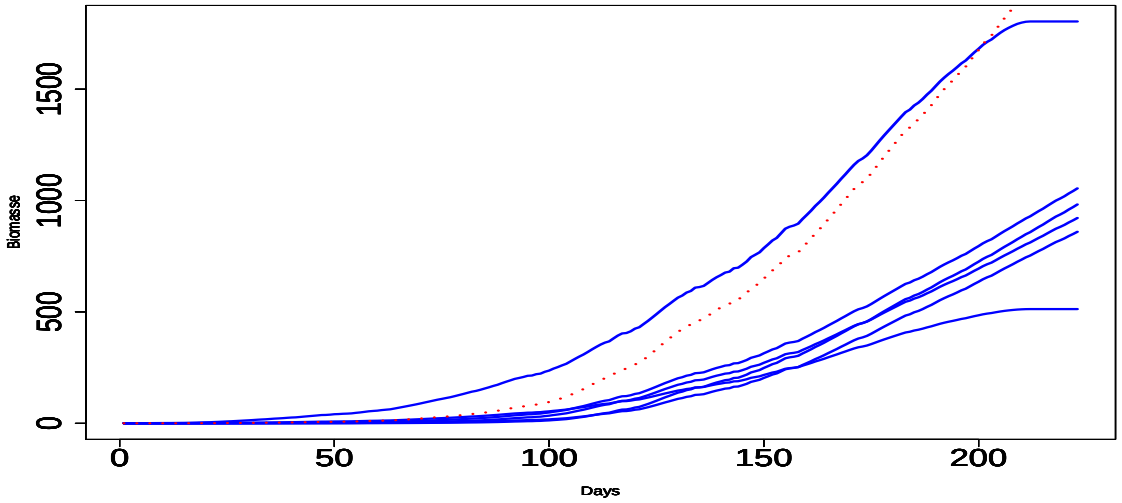

FIGURE 1.1 – Incertitude sur les sorties du modèle WWDM due à la variabilité des paramètres du modèle. La courbe en pointillé correspond aux valeurs nominales des paramètres et les autres courbes correspondent à des valeurs tirées au hasard dans les intervalles d'incertitude.

que la température moyenne, le rayonnement global et les précipitations.

Le modèle AZODYN simule différentes composantes telles que le rendement, la biomasse, la teneur en protéines des grains, le reliquat d'azote minérale dans le sol à la récolte ainsi que l'indice de nutrition azotée (INN). L'estimation de la biomasse journalière est déterminée à partir du coefficient de conversion de l'énergie lumineuse et du rayonnement intercepté par la plante qui est lui même calculé à partir de l'indice foliaire. Le calcul de l'accumulation maximale journalière d'azote se fait à partir de la production journalière de biomasse et de la teneur maximale en azote des parties aériennes. La capacité d'absorption d'azote par la plante témoin de l'indice de nitration de la plante est décrite par une vitesse d'absorption d'azote. La relation établie entre l'INN et le nombre de grains par m^2 permet le calcul du rendement. L'azote cumulé est une fonction de l'azote minéral résiduel dans le sol, des apports d'engrais et de l'INN. L'initialisation du modèle débute avec la mesure de l'azote minéral résiduel dans le sol à la sortie-hiver. Certaines sorties du modèle sont simulées à un pas de temps journalier comme INN, l'azote absorbé (Kg/ha), l'azote cumulé dans le sol (Kg/ha) tandis que le rendement et le reliquat d'azote minérale dans le sol à la récolte sont des sorties statiques.

Le modèle AZODYN inclut 69 paramètres dont 13 paramètres qui varient selon les génotypes. Nous ne considérons que les 13 paramètres génotypiques dans notre étude qui sont présentés dans la Table 1.2 avec leurs gammes d'incertitude.

La Figure 1.2 représente quelques simulations de l'INN en prenant quelques valeurs possibles des paramètres tirées dans les intervalles d'incertitude pour un scénario climatique donné.

Le modèle CERES-EGC

Le modèle CERES-EGC a été adapté de la série des modèles de culture CERES-sol en mettant plus d'accent sur des aspects environnementaux tels que le lessivage des nitrates, les émissions de gaz à effet de serre (CO_2, N_2O) et les oxydes d'azote (Gabrielle *et al.*, 2006a[58], 2006b [57]). Le modèle CERES est disponible pour un grand nombre d'espèces de culture qui partagent les mêmes composants du sol (Jones and Kiniry, 1986 [84]). CERES-EGC simule ces différentes sorties du modèle à un pas de temps journalier et prend en entrée les variables suivantes : la pluie, la température moyenne de l'air, le potentiel d'évapo-transpiration mesuré quotidiennement et la typologie du sol.

Le modèle CERES-EGC comprend plusieurs sous-modules correspondant aux principaux processus qui régissent les cycles de l'eau, du carbone et d'azote dans les systèmes

Paramètre	Interprétation	Valeur nominale	Intervalle d'incertitude
$E_{b\max}$	coefficient de conversion du rayonnement intercepté en biomasse (gm^{-2})	3	2.7-3.3
$E_{i\max}$	rapport maximal entre rayonnements intercepté et incident	0.94	0.9-0.99
K	coefficient d'extinction	0.7	0.6-0.8
Tep.flo	durée entre floraison et épiaison	150	100-200
D	rapport entre LAI et niveau d'azote critique	0.035	0.020-0.045
R	rapport entre l'azote total et l'azote des parties aériennes	1.25	1-1.5
λ	coefficient d'efficacité pour l'azote	35	25-45
μ	coefficient d'efficacité pour l'azote	0.75	0.6-0.9
DJPF	seuil de température	200	150-250
NGM2MAXVAR	nombre de grains maximal	128	107.95-146.05
P1GMAXVAR	poids maximal d'un grain	56	47-65
RDTMAXVAR	rendement maximal	118	100-137
REM2	fraction d'azote remobilisé	0.7	0.5-0.9

TABLE 1.2 – Intervalles d'incertitude des différents paramètres du modèle AZODYN

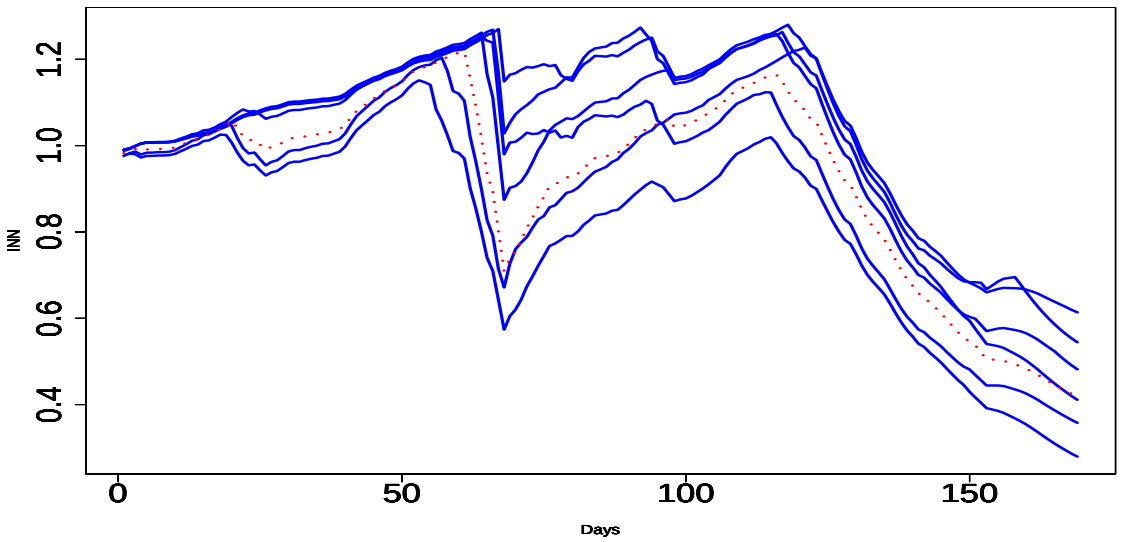

FIGURE 1.2 – Incertitude sur les sorties du modèle AZODYN due à la variabilité des paramètres du modèle. La courbe en pointillé correspond aux valeurs nominales des paramètres et les autres courbes correspondent à des valeurs tirées au hasard dans les intervalles d'incertitude.

sol-culture. Le module sol décrit les processus de transfert de chaleur, d'eau et de nitrate dans le sol ainsi que l'évaporation du sol, l'absorption de l'eau par des plantes et leur transpiration. Le module biologique du modèle CERES-EGC simule la croissance et la phénologie des cultures et la composante microbiologie décrit le processus de dégradation de la matière organique et minérale dans les parcelles agricoles par les micro-organismes : décomposition, minéralisation, immobilisation de l'azote. Enfin, le module oxyde nitreux qui nous intéresse plus particulièrement dans ce mémoire modélise l'émission quotidienne du gaz à effet de serre N_2O issu des parcelles agricoles. Le module oxyde nitreux (N_2O) a été adapté du modèle semi-empirique NOE (Hénault and Germon, 2000 [66]) et l'émission de N_2O provient de deux processus :

- la composante de dénitrification est dérivée du modèle NEMIS (Hénault *et al.*, 2000) qui calcule le taux de dénitrification (Da, en kg N ha^{-1} $jour^{-1}$) comme le produit d'un potentiel de vitesse à 20 ° C (PDR, en kg N ha^{-1} $jour^{-1}$) et de trois fonctions liées aux pores d'eau (Water Filled Pore Space : WFPS) (F_W), à la teneur en nitrates (F_N) et à la température (F_T) dans les couches arables. L'équation mathématique qui décrit ce processus de dénitrification est donnée par :

$$De = \mathrm{PDR} \times F_N \times F_W \times F_T$$

- de la même façon, le taux de nitrification quotidienne (Ni, kg N ha^{-1} $jour^{-1}$) est modélisé comme le produit d'un taux maximal de nitrification à 20 ° C (MNR, kg N ha^{-1} $jour^{-1}$) et de trois fonctions liées aux pores d'eau (N_W), à la concentration d'ammonium (N_N) et à la température (N_T) :

$$Ni = \mathrm{MNR} \times N_N \times N_W \times N_T$$

Les émissions d'oxyde nitreux résultant de ces deux processus sont spécifiques au sol considéré. Le N_2O total émis est alors une somme pondérée de ces deux modes d'émissions :

$$N_2O = \mathrm{r}De + \mathrm{c}Ni,$$

où r est la fraction de l'azote dénitrifié et c est la fraction de l'azote nitrifié.

Le modèle CERES-EGC inclut en plus des variables d'entrée 15 paramètres principaux. On distingue 4 paramètres Locaux c'est-à-dire dépendant du site : Potential Denitrification Rate (PDR), Maximal Nitrification Rate (MNR) et les fractions de l'azote nitrifié (c) et de l'azote dénitrifié (r). Les différents paramètres incertains sont présentés dans la Table 1.3.

Paramètre	Interprétation	Valeur nominale	Intervalle d'incertitude
$seuil\_wfps$	seuil de réponse "Water Field Pore Space"	0.62	0.4-0.8
Km	coefficient de demi- saturation (dénitrification) (mgN-NO_3Kg^{-1}sol)	22	5-120
$seuil\_t$	seuil de température	11	10-15
Q_dix_un	-	89	60-120
Q_dix_deux	-	2.1	1-4.8
puissance	-	1.74	0-2
opt_wfps	-	0.6	0.35-0.75
min_wfps	-	0.1	0.05-0.15
max_wfps	-	0.8	0.8-1
Km_amm	coefficient de saturation :(nitrification) : mgN-NO_3Kg^{-1}sol	10	1-50
Q_dix_nit	-	2.1	1.9-13
PDR	taux potentiel de dénitrification : KgNha^{-1}jour^{-1}	7	01-20
MNR	taux maximum de nitrification : KgNha^{-1}jour^{-1}	9	4-13
r	ratio de N_2 O (N_2O /dénitrification)	0.25	0.09-0.9
c	ratio de N_2 O (N_2O /nitrification)	0.018	0.0002-0.1

TABLE 1.3 – Intervalles d'incertitude des différents paramètres du modèle CERES.

La Figure 1.3 représente quelques simulations de l'émission du l'oxyde nitreux (N_2O) pour le site Villamblain en faisant varier les valeurs des paramètres.

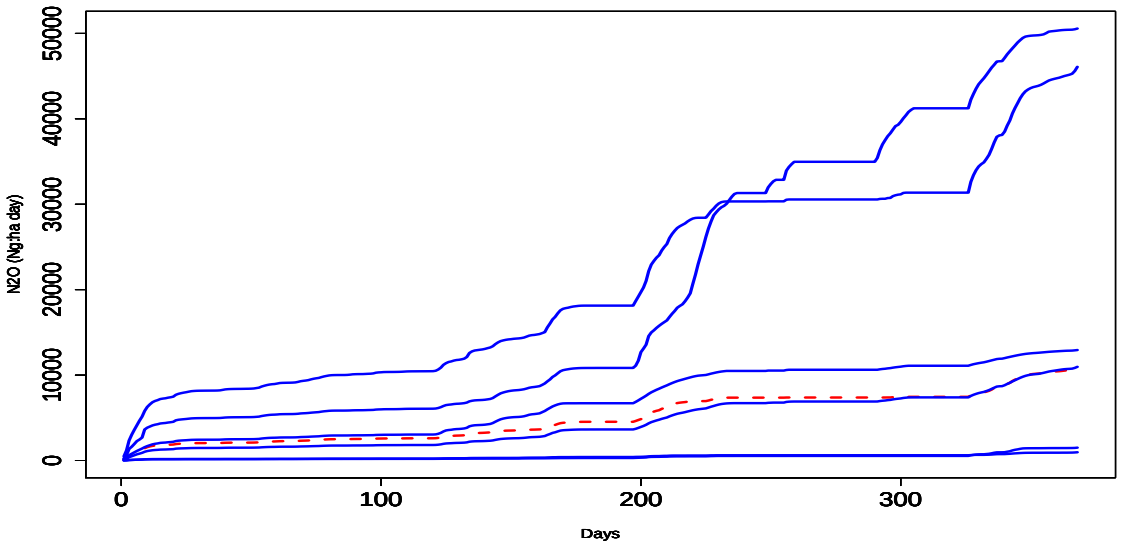

FIGURE 1.3 – Incertitude sur les sorties du modèle CERES-EGC due à la variabilité des paramètres du modèle. La courbe en pointillé correspond aux valeurs nominales des paramètres et les autres courbes correspondent à des valeurs tirées au hasard dans les intervalles d'incertitude.

1.3 Conclusion

La complexité des trois modèles dynamiques que nous venons de décrire diffère selon le nombre de paramètres inconnus et le nombre de sorties de chaque modèle. Les sorties dynamiques de ces modèles sont régulières sauf celui du modèle AZODYN qui comporte quelque irrégularités. Ainsi, nous pouvons appliquer les différentes méthodes que nous proposerons dans les chapitres suivants sur ces différents modèles pour évaluer leurs limites.

Les modèles de culture sont généralement organisés en modules. Chacun de ces modules décrit soit la physiologie de la plante soit les échanges de flux entre la plante et le sol soit les échanges gazeux entre la partie aérienne de la plante et le milieu aérien. Ces

différents modules incluent des paramètres inconnus et au final le modèle obtenu comporte de nombreux paramètres inconnus. L'analyse de sensibilité pourra permettre d'identifier les sous modules les plus importants et conduire à la simplification du modèle.

Les sorties des modèles dynamiques basés sur les modèles de culture sont fonctions des scénarios climatiques. L'évaluation de la qualité de ces modèles se fait en général par des critères tels que le $\mathbb{MSEP}$ et le $\mathbb{MSE}$.

Chapitre 2

Méthodes d'exploration numérique des modèles : lien avec l'estimation des paramètres

La modélisation des phénomènes naturels ou des procédés humains est entachée d'incertitudes qu'il est intéressant de quantifier. Dans ce chapitre, nous présentons une synthèse originale des différentes méthodes d'exploration de modèles dynamiques. Nous considérons des méthodes à valeur générique qui s'appliquent à tout type de modèles dynamiques. Le chapitre est organisé comme suit : i) l'Analyse de Sensibilité (AS) indispensable pour identifier les principales sources d'incertitudes dans un modèle complexe ii) le choix de paramètres à estimer par l'AS et les propriétés des modèles obtenus iii) les techniques d'analyses de données multivariées qui permettront d'analyser les sorties des modèles dynamiques issues des simulations.

2.1 Analyse de sensibilité et d'incertitude

L'analyse d'incertitude consiste à propager l'incertitude des facteurs (inputs incertains) sur les fonctions réponses des modèles. L'incertitude sur un facteur est soit représentée par une loi de distribution soit par la donnée d'une gamme de valeurs du facteur. Cette analyse permettra d'établir la distribution de la sortie du modèle qui mesure l'incertitude de la fonction réponse par rapport aux incertitudes des facteurs. Les différentes techniques d'analyse ou de propagation d'incertitudes sont présentées dans Saltelli *et al.* (2000) [126] ; Saltelli *et al.* , 2008 [127] ; Da-Veiga (2007) [43] ; Marrel, 2008 [100] ; Jacques (2005) [80].

L'analyse d'incertitude est quasi systématiquement accompagnée d'une analyse de sensibilité. Selon Saltelli *et al.*, (2000) [126], l'analyse de sensibilité consiste à répartir l'incertitude de la sortie du modèle entre les différentes sources d'incertitudes des entrées du modèle. L'analyse de sensibilité quantifie la part de l'incertitude de la sortie du modèle expliquée par chaque facteur du modèle conditionnellement aux intervalles d'incertitude des facteurs. Ainsi, l'analyse de sensibilité devient un pré requis pour la modélisation (Saltelli, 2002 [125]) et un outil important pour évaluer l'incertitude des modèles. Selon Kolb in Rabitz, 1989 [117] et Furbinger, 1996 [56] il serait intellectuellement malhonnête de concevoir ou d'améliorer un modèle sans faire l'analyse de sensibilité.

2.1.1 Ambitions et propriétés des méthodes d'AS

Les ambitions de l'analyse de sensibilité sont énumérées dans Saltelli *et al.*, (2000) [126] ; Saltelli *et al.*, (2008) [127] ; Jacques (2005) [80] ; Da-Veiga (2007) [43] ; Marrel(2008) [100]. L'analyse de sensibilité conduit systématiquement à identifier les facteurs les plus influents d'un modèle. Ainsi, le classement des facteurs connu sous le non de "factor prioritization" et la fixation des facteurs moins influents "factor fixing" deviennent possibles. L'analyse de sensibilité permet également d'identifier les sous gammes de variations des facteurs sur lesquelles il faut agir pour que la sortie du modèle soit dans une gamme bien précise : c'est le "factor mapping". Enfin, la "variance cuting" consiste à identifier un sous-ensemble de facteurs avec lesquels il faut travailler pour réduire la variabilité de la sortie en dessous d'un seuil. Ce dernier objectif est connu aussi sous le nom de la dimension effective développée dans Tao (2003) [146], Tao et Owen (2003) [147]. Notons que cette notion est importante dans la mesure où elle constitue une réponse au problème du fléau de la dimension mais aussi au choix de modèles.

Les différentes méthodes d'AS peuvent être organisées en trois grands groupes : le premier groupe qui ne sera pas du tout abordé dans ce mémoire concerne toutes les méthodes dites locales (analyse basée sur les dérivés). Dans le second groupe, nous retrouvons les méthodes de criblages ou "screening" telles que les bifurcations séquentielles, la méthode screening de Morris (1991) [106], la méthode OAT(One factor At Time) et l'AS à l'aide de l'ANOVA. Ces méthodes ont le mérite d'être peu coûteuses en temps de calcul et certaines sont bien adaptées aux modèles incluant des facteurs d'entrée qualitatifs. En contrepartie, les indices obtenus ne couvrent pas toute la gamme d'incertitude dans le cas des facteurs quantitatifs continus. Le dernier groupe est connu sous le nom de "variance based". Il regroupe les méthodes basées sur la décomposition de la variance en faisant varier simultanément tous les facteurs, telles que les indices de Sobol (Sobol, 1993 [142]),

la décomposition de la variance (Homma *et al.*, 1996 [70]), l'AS à l'aide de l'ANOVA (Archer *etal*, 1997 [13]), la méthode FAST (Cukier *et al.*, 1973 [40]) et EFAST (Saltelli *et al.*, 1999 [130]). Ces différentes méthodes sont décrites dans la Section 2.1.2 en s'appuyant sur les travaux de Saltelli *et al.*, (2008) [127] ; Saltelli *et al.*, (2004) [128] ; Saltelli *et al.*, (2000) [126] ; Saltelli (2002), [124] ; Monod *et al.*, (2006) [105] ; Morris (1991) [106] ; Campolongo *et al.*, (2007) [32]. Nous nous intéressons dans ce mémoire aux méthodes dites globales, indépendantes des hypothèses faites sur la forme du modèle, robustes et pertinentes pour quantifier l'incertitude inhérente aux modèles.

Les propriétés intéressantes d'une méthode d'AS sont listées ci-dessous (Saltelli, 2002 [125]) :

- "include multidimensional averaging" : le calcul des effets de chaque facteur doit se faire en faisant varier tous les facteurs simultanément contrairement à la méthode OAT.

- "model free " : la méthode doit être indépendante des hypothèses faites sur le modèle. Une méthode globale devra pouvoir identifier les interactions importantes entre les différents facteurs aussi bien pour des modèles linéaires et additifs que pour des modèles non linéaires et non additifs. Elle doit pouvoir s'appliquer à tous les modèles une fois que les facteurs d'entrée et la fonction réponse du modèle sont identifiés.

- "single index" : pour répondre aux ambitions de "factors fixing" et "factors prioritization", une mesure globale de la sensibilité des facteurs devra naturellement fournir un unique indice principal et un unique indice total pour chaque facteur. Eventuellement certaines statistiques sur la précision de l'estimation de ces indices pourront être jointes.

- "grouped factors as single factor" : une analyse de sensibilité globale devra pouvoir traiter un groupe de facteurs comme un seul facteur. Cette propriété essentielle de synthèse permet une interprétation simple des résultats d'une analyse de sensibilité. En effet, un groupe de facteurs avec un indice total suffisamment faible évite de détailler les effets de chaque facteur qui le compose.

- "shape and scale" : l'analyse de sensibilité globale doit incorporer la gamme d'incertitude des facteurs et également la manière dont on échantillonne l'espace des

facteurs (plan d'expérience ou distribution des facteurs d'entrée).

2.1.2 Méthodes classiques d'analyse de sensibilité

Méthode de Morris

La méthode de Morris (1991) [106] fait partie de la classe de méthodes dites "screening" en analyse de sensibilité. Les méthodes screening se différencient par leur faible coût en temps de calculs et sont pratiquement utilisées pour les gros modèles (modèles coûteux en temps de calculs, modèles avec un nombre important de facteurs). Ces méthodes sont basées généralement sur l'approche OAT (One factor At Time) et fournissent une première information qualitative et partielle sur l'importance des facteurs. Le principe de l'approche OAT consiste à faire varier un facteur à la fois en fixant les autres facteurs à leurs valeurs nominales. Cependant la méthode de Morris décrite dans ce mémoire ne souffre pas de cet inconvénient majeur de l'analyse OAT. De plus, cette méthode fait partie des méthodes dites "variance based".

La méthode de Morris a le mérite de fournir en plus des indices du premier ordre, l'information sur les facteurs non influents mais qui interagissent avec d'autres facteurs, la nature additive ou linéaire du modèle.

Considérons un plan factoriel complet $\mathbf{z}(n)$ des d facteurs du modèle et supposons sans perdre de généralité que chaque facteur Z_j possède p niveaux (uniformément et régulièrement repartis sur l'intervalle $[0\,1]$) et prend ces valeurs dans l'ensemble $Z^{(j)}(\Omega) = \{0, \frac{1}{p-1}, \ldots, 1\}$. Il est évident de constater que l'ensemble des scénarios $\mathbf{z}(n) = \{0, \frac{1}{p-1}, \ldots, 1\}^d$ et que $n = p^d$. Soit $\triangle$ le pas de discrétisation de l'espace d'un facteur donné ou un multiple de ce dernier. Morris définit l'effet élémentaire EE du facteur Z_j en un point $\mathbf{z}$ par :

$$EE_{z^{(j)}} = \frac{f(\mathbf{z} + \mathbb{1}_j \triangle) - f(\mathbf{z})}{\triangle} \tag{2.1.1}$$

La distribution discrète $\mathbb{F}_j$ des effets élémentaires $EE_{Z^{(j)}}$ est obtenue en tirant aléatoirement le vecteur $\mathbf{z}$ dans $\mathbf{z}(n)$ et le cardinal de l'espace certains de $\mathbb{F}_j$ vaut $r = p^{d-1}[p - \triangle(p-1)]$. En effet, il y a p^{d-1} possibilités de tirer le vecteur $\mathbf{z}$ en fixant le niveau correspondant au facteur $Z^{(j)}$ à une valeur donnée et il y a $p - \triangle(p-1)$ niveaux du facteur $Z^{(j)}$ disponibles.

Les statistiques permettant de résumer la distribution F_j telles que la moyenne μ et l'écart type σ contiennent des informations utiles sur l'importance des facteurs. Une moyenne élevée témoigne de l'importance du facteur en terme d'influence sur la sortie du modèle et une forte variabilité de la distribution $\mathbb{F}_j$ indique une interaction entre ce

dernier et les autres facteurs.

Notons que le nombre d'évaluations du modèle nécessaires pour le calcul des indices est une fonction exponentielle du nombre de facteurs comme tout plan factoriel complet. Morris propose de choisir p comme étant un entier pair et fixe $\triangle = \frac{p}{2(p-1)}$ pour réduire de moitié le nombre d'évaluations du modèle $r = p^{d-1}[p - \triangle(p-1)] = p^d/2$ mais le nombre obtenu reste toujours élevé. Pour réduire considérablement le nombre d'évaluations du modèle, Morris propose d'échantillonner r suites de points $\mathbf{z}^*(r)$ dans l'espace complet $\mathbf{z}(n)$ avec $r < n$. Pour chaque point de $\mathbf{z}^*(r)$, il construit une trajectoire à $d+1$ points pour calculer les effets élémentaires des d facteurs. L'algorithme décrivant cette trajectoire est le suivant :

Algorithme 2.1.1 *trajectoire de Morris (1991)[106]*

Etape 1 :	*soient* $\mathbf{z}_1^*$ *un point de* $\mathbf{z}^*(r)$ *et* $\triangle > 0$
Etape 2 :	*le second point de la trajectoire* $\mathbf{z}_2^*$ *est défini par :*
	$\mathbf{z}_2^* = \mathbf{z}_1^* + sign \triangle \mathbb{1}_j$ *et calculer*
	$EE_{z^{(j)}} = \frac{f(\mathbf{z}_2^*) - f(\mathbf{z}_1^*)}{\triangle}$ *si* $sign = +$
	$EE_{z^{(j)}} = \frac{f(\mathbf{z}_1^*) - f(\mathbf{z}_2^*)}{\triangle}$ *si* $sign = -$
Etape 3 :	*poser* $j \leftarrow k$ *et* $\mathbf{Z}_1^* \leftarrow \mathbf{Z}_2^*$
	et reprendre l'étape 2 d fois pour calculer les effets élémentaires de tous les facteurs.

Notons que cet algorithme nécessite $d+1$ évaluations du modèle pour le calcul de chaque effet élémentaire des différents facteurs et que le calcul de chaque effet élémentaire se fait toujours en augmentant d'un pas de discrétisation $\triangle$. Le nombre total d'évaluations du modèle pour calculer tous les indices est de $n = r(d+1)$.

Remarquons que la planification ainsi proposée par Morris peut ne pas être optimale dans le sens où elle ne remplit pas au mieux l'espace. Campolongo *et al.* (2007) [32] corrigent ce défaut en proposant de sélectionner les r suites de points de manière à couvrir au mieux l'espace à l'aide d'une distance euclidienne arbitraire. Leur idée est de générer plusieurs trajectoires de Morris $(500 - 1000)$ et ensuite de sélectionner les trajectoires qui maximisent cette distance. De plus, ils proposent une nouvelle mesure μ^* permettant de classer rapidement les différents facteurs du modèle sans utiliser la représentation graphique comme dans le cas de Morris. La mesure μ^* :

$$\mu^* = \frac{1}{r} \sum_{i=1}^{r} \mid EE_i \mid, \tag{2.1.2}$$

est la moyenne de la distribution de la valeur absolue des effets élémentaires de l'équation (2.1.1) et permet d'éviter la compensation des effets négatifs et positifs. Cette mesure est beaucoup plus robuste par rapport au moment d'ordre 2 centré. Pour le calcul des effets élémentaires dans l'équation (2.1.1), seul le facteur $Z^{(j)}$ varie et les autres facteurs $Z^{(-j)}$ sont fixes. Ensuite, on moyenne sur toutes les valeurs prises par $Z^{(-j)}$ pour obtenir le $\mu^*_{z^{(j)}}$. Si on approxime la variance $\mathbb{V}\mathrm{ar}(Y \mid Z^{(-j)})$ par la variation locale $\mid EE_{z^{(j)}} \mid$ (au lieu de $(EE_{z^{(j)}})^2$) alors

$$\mu^*_{z^{(j)}} \simeq \mathbb{E}_{Z^{(-j)}} \left[\mathbb{V}\mathrm{ar}_{z^{(j)}}(Y \mid Z^{(-j)}) \right],$$

est une approximation de l'indice total du facteur $Z^{(j)}$ et peut être utilisé pour fixer les facteurs non influents (Saltelli *et al.*, 2008 [127]). Il faudra tout de même comparer le signe de μ et μ^* pour avoir une idée sur l'impact négatif ou positif du facteur sur la sortie du modèle.

Malgré la correction apportée à la méthode de Morris, cette méthode reste qualitative et une analyse complémentaire et quantitative sur les facteurs retenus est nécessaire pour mieux explorer le modèle.

Méthode basée sur l'ANOVA

D'un point de vue général, *l'ANOVA consiste classiquement à identifier les effets des traitements sur une variable de sortie.* Classiquement, l'ANOVA permet de modéliser une relation linéaire entre des facteurs qualitatifs ou les facteurs quantitatifs discrets et une variable réponse quantitative dans le but d'identifier les effets des différents facteurs. Le formalisme mathématique de l'analyse de la variance fut introduit par Fisher and Yates (1934) [51] pour comparer les rendements de certaines variétés du blé issus des plans d'expérience (voir aussi Fisher, 1958 [52]). L'utilisation moderne de l'ANOVA, dans la contexte de l'AS, s'appuie sur les données simulées et permet d'identifier les facteurs les plus influents d'un modèle (Archer *et al.*, 1997 [13] ; Ginot *et al.*, 2006 [61] ; Castillo, 2007 [34]). Dans ce paragraphe, nous nous intéressons au formalisme mathématique de l'analyse de la variance classique (ANOVA) adapté aux méthodes d'analyse de sensibilité.

Supposons que le facteur $Z^{(j)}$ possède n_j modalités $\forall\, j \in \{1, 2, \ldots, d\}$ et que $n = \prod_{j=1}^{d} n_j$ est la taille du plan. Un plan factoriel fractionnaire ou d'autres plans tels que les carrés latins, les blocs complets permettent de réduire considérablement cette taille n. Le but de ce mémoire ne portant pas sur les plans d'expérience, nous considérons dans la suite un plan factoriel complet qui garantit l'identification de tous les effets factoriels

dans la décomposition de l'ANOVA. Toutefois, le développement des plans d'expérience se trouvent dans Box et Draper (1987) [25], Kobilinsky (1997) [86].

Le modèle de l'analyse de variance dans le cas des facteurs qualitatifs ou discrets et en l'absence du terme d'erreur s'écrit :

$$y_{z_{i_1}^{(1)},\ldots,z_{i_d}^{(d)}} = \eta + \eta_{z_{i_1}^{(1)}} + \eta_{z_{i_2}^{(2)}} + \ldots + \eta_{z_{i_1}^{(1)},z_{i_2}^{(2)}} + \ldots + \eta_{z_{i_1}^{(1)},\ldots,z_{i_d}^{(d)}}, \tag{2.1.3}$$

où η est l'effet moyen ; $\eta_{z_{i_1}^{(1)}}$ est l'effet de la modalité i_1 du facteur $z^{[1)}$; $\eta_{z_{i_1}^{(1)},z_{i_2}^{(2)}}$ est l'effet de l'interaction entre les modalités i_1 du facteur $z^{(1)}$ et i_2 du facteur $z^{(2)}$. L'équation (2.1.3) permet d'avoir une interprétation intuitive du modèle mais s'écrit de façon générale comme suit :

$$y_{z_{i_1}^{(1)},\ldots,z_{i_d}^{(d)}} = \eta + \sum_{j=1}^{d} \eta_{z_{i_j}^{(j)}} + \sum_{l<j} \eta_{z_{i_j}^{(j)},z_{i_l}^{(l)}} + \ldots + \eta_{z_{i_1}^{(1)},\ldots,z_{i_d}^{(d)}} \tag{2.1.4}$$

En AS, nous nous intéressons à quantifier l'effet global d'un facteur sur la fonction réponse sans chercher à identifier les effets de chaque modalité du facteur. Ceci nous conduit à considérer le modèle d'ANOVA vectoriel suivant :

$$\mathbf{y} = \eta \mathbb{1}_{\bullet\bullet} + \sum_{j=1}^{d} \eta_{z^{(j)}} + \sum_{l<j} \eta_{z^{(l)},z^{(j)}} + \ldots + \eta_{z^{(1)},\ldots,z^{(d)}},$$

ou de façon équivalente

$$\mathbf{y} = \sum_{u \subseteq \{1,2,\ldots,d\}} \eta_{z^{(u)}}, \tag{2.1.5}$$

avec $\eta_{z_\emptyset} = \eta \mathbb{1}_{\bullet\bullet}$, $\mathbf{y}$ le vecteur des réponses du modèle, $\eta_{z^{(j)}}$ le vecteur des effets du facteur $z^{(j)}$ et $\eta_{z^{(l)},z^{(j)}}$ le vecteur des effets d'interaction entre $z^{(j)}$ et $z^{(l)}$. Dans cette décomposition, les termes $\eta_{z^{(u)}}$ désignent les effets principaux si le cardinal $|u| = 1$; les effets du second ordre si $|u| = 2$ et les effets du $k$ième ordre si $|u| = k$.

Notons que le nombre de termes dans cette décomposition est de 2^d et qu'habituellement, on suppose que les effets d'ordre élevé sont négligeables pour réduire les calculs. Cette pratique est une contrainte due aux plans d'expérience qui ne permettent pas d'estimer les effets d'ordre élevé. Une approximation parcimonieuse consiste à inclure dans le modèle un certain nombre de facteurs et d'interactions permettant d'expliquer une grande part de la variabilité de la fonction réponse. Dans le cas de l'analyse fonctionnelle, Tao (2003) [146] Tao *et al.* (2003) [147] ont proposé un critère pour le choix des facteurs $Z^{(u)}$ $u \in \{1, 2, \ldots, d\}$ à considérer dans l'ANOVA à l'aide des indices de sensibilité. Le cardinal de l'ensemble des facteurs retenus est alors appelé la dimension effective (Wand

et Fang, 2003 [159]).

Rappelons que cette décomposition du vecteur de sorties n'est pas unique et que le calcul des différents effets des facteurs se fait exactement dans le cas d'un plan factoriel complet et de manière approximé dans le cas des plans factoriels fractionnaires (problème de confusion de certains effets : Box et Draper, 1987 [25] ; Kobilinsky, 1997 [86]). Notons également que l'équation (2.1.5) s'écrit comme un modèle de régression particulier dans le sens où il n'y a pas de terme d'erreur. la proposition (2.1.1) assure une décomposition unique et permet le calcul des effets des facteurs.

Proposition 2.1.1 *Sous l'hypothèse d'un plan factoriel complet et en posant :*
$E_\emptyset = \mathbb{V}ec\{\mathbb{1}_{\bullet\bullet}\}$
$E_\emptyset \overset{\perp}{\oplus} E_j = \mathbb{V}ec\left\{\mathbb{1}_{z_j.i_j\bullet};\ i_j \in \{1,2,\ldots,n_j\}\right\}$
$\left(\overset{\perp}{\bigoplus_{v\subset u}} E_v\right) \overset{\perp}{\oplus} E_u = \mathbb{V}ec\{\mathbb{1}_{z_u.i_u\bullet};\ \forall\, u \subseteq \{1,2,\ldots,d\}\}$
Alors :

(*i*) $\mathbb{R}^n = \overset{\perp}{\bigoplus_{u\subseteq\{1,2,\ldots,d\}}} E_u$

(*ii*) $\overset{\perp}{H}_{E_u}\overset{\perp}{H}_{E_v} = 0, si \quad u \neq v$

(*iii*) *Décomposition unique et orthogonale de l'équation (2.1.5) avec*

$$\eta_{z^{(u)}} = \overset{\perp}{H}_{E_u}\mathbf{y} \tag{2.1.6}$$

$$= \overset{\perp}{H}_{\mathbb{V}ec\left\{\mathbb{1}_{z^{(u)}.i_u\bullet}\right\}}\mathbf{y} - \sum_{v\subset u}\overset{\perp}{H}_{E_v}\mathbf{y}, \tag{2.1.7}$$

où $\overset{\perp}{H}_{E_v}$ *est l'opérateur de projection orthogonale sur l'espace* E_v.

Preuve 2.1.1 *La décomposition de l'espace vectoriel* $\mathbb{R}^N$ *en une somme directe orthogonale est juste une application du processus d'orthogonalisation de Gram-Schmidt. L'unicité de la décomposition résulte de l'unicité de la projection orthogonale. L'orthogonalité ou le caractère additif des termes* $\eta_{z^{(u)}}$ *est une conséquence du théorème de Pythagore.*

□

Remarquons que dans cette proposition les effets factoriels $z^{(u)}, u \subseteq \{1,2,\ldots,d\}$ sont obtenus en projetant le vecteur des sorties du modèle sur des sous espaces orthogonaux

choisis de sorte que les projections obtenues correspondent bien aux effets factoriels. Partant de l'interprétation géométrique de l'espérance conditionnelle, la projection orthogonale de $\mathbf{y}$ sur un sous espace vectoriel E_u peut s'écrire comme l'espérance conditionnelle par rapport à la mesure empirique $P_n(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{z}}(\mathbf{z}_i)$ et on a :

$$\eta_{z^{(u)}} = \overset{\perp}{H}_{E_u}\mathbf{y} \tag{2.1.8}$$

$$\simeq \mathbb{E}(\mathbf{y} \mid z^{(u)}) \tag{2.1.9}$$

Notons que l'équivalence entre la projection sur un sous espace et l'espérance conditionnelle dans l'expression (2.1.8) permet une généralisation de l'ANOVA aux facteurs quantitatifs discrets en prenant des mesures adaptées à chaque facteur et qui ne chargent pas toutes les modalités de la même façon.

La conséquence directe de l'orthogonalité de la décomposition (2.1.5) est que la variabilité ou la somme des carrés de $\mathbf{y}$ se décompose comme la somme des carrés associée aux différents termes factoriels $z^{(u)}, u \subseteq \{1, 2, \ldots, d\}$.

Proposition 2.1.2 *Si* $||\mathbf{y}||^2 < \infty$ *et en posant* $SC_{\mathbf{y}} = ||\mathbf{y} - \eta\mathbb{1}_{\bullet\bullet}||^2$ *;* $SC_{z^{(u)}} = ||\eta_{z^{(u)}}||^2$ *alors*

$$SC_{\mathbf{y}} = \sum_{u\subseteq\{1,2,\ldots,d\}} SC_{z^{(u)}}, \tag{2.1.10}$$

Preuve 2.1.2 *La démonstration est immédiate en utilisant la définition du produit scalaire et l'orthogonalité des termes factoriels.*

□

Remarquons qu'en divisant toute l'équation (2.1.10) par n (taille du plan d'expérience), on obtient une décomposition de la variance empirique de $\underline{y}$ en des parts de variances expliquées par les termes factoriels $z^{(u)}$ et les indices se définissent naturellement par :

$$\mathrm{IS}_{z^{(u)}} = \frac{\mathrm{SC}_{z^{(u)}}}{\mathrm{SC}_{\mathbf{y}}}, \quad u \subseteq \{1, 2, \ldots, d\} \tag{2.1.11}$$

$$\mathrm{IST}_{z^{(j)}} = \frac{\sum_{u\subseteq\{1,2,\ldots,d\} j\in u} \mathrm{SC}_{z^{(u)}}}{\mathrm{SC}_{\mathbf{y}}} \tag{2.1.12}$$

La méthode de Morris et celle d'AS basée sur l'ANOVA sont toutes deux basées sur les plans factoriels mais l'approche par l'ANOVA est plus générale et quantitative. L'aspect criblage (screening) de cette méthode se réalise à l'aide de plans factoriels fractionnaires de faible taille et peut conduire, dans certains cas, aux calculs des indices du premier

ordre et totaux (plans fractionnaires de résolution $R \gg 4$). Avec moins d'évaluations du modèle, cette méthode est la version empirique de l'analyse de variance fonctionnelle présentée dans la Section 2.1.2. Archer *et al.* (1997) [13] affirment que l'AS est l'ANOVA et l'ANOVA est l'AS.

Méthode de Sobol : ANOVA fonctionnelle

La méthode de Sobol (1993) [142] est largement utilisée en AS pour des modèles de faible dimension du fait qu'elle permet de balayer de façon aléatoire toute la gamme d'incertitude des facteurs et qu'elle permet d'avoir des précisions sur les indices. Les indices de Sobol sont basés sur la décomposition d'une fonction $f(z)$ comme une somme de fonctions de dimensions croissantes comme dans l'ANOVA classique présentée dans l'équation (2.1.5) de la Section 2.1.2.

L'ANOVA fonctionnelle introduite en statistique mathématique par Hoeffding (1948) [68] généralise l'ANOVA classique en l'élargissant aux facteurs quantitatifs continus. Elle fut ensuite étudiée par Efron et Stein (1981) [48] dans le contexte de l'estimation de la variance ; Antoniadis (1984) [12] dans le contexte de l'espace fonctionnel ; Sobol (1993) [142] dans le contexte de l'AS ; Stone (1994) [143] pour la formulation de l'ANOVA comme une méthode de régression ; Ramsay et Silverman (1997) [118] pour l'ANOVA temporelle ; Tao (2003)[146], Tao et Owen (2003) [147], Lemieux et Owen (2000) [94] pour la présentation de l'ANOVA comme une quasi-régression. Hooker (2007) [71] prend en compte la dépendance entre les différents facteurs en développant l'ANOVA fonctionnelle adaptée aux facteurs dépendants.

L'utilisation moderne de l'ANOVA est largement rencontrée en AS et le succès de cette décomposition vient du fait que l'ANOVA constitue une réponse alternative au fléau de la dimension grâce à sa décomposition en dimensions croissantes et à la définition de la dimension effective (Tao, 2003 [146] ; Wang et Fang, 2003 [159]).

La décomposition de Sobol s'écrit :

$$\begin{aligned} f(z) &= \sum_{j=1}^{d} f_{\{j\}}(z^{(j)}) + \sum_{j_1<j_2} f_{\{j_1,j_2\}}(z^{(j_1)}, z^{(j_2)}) + \ldots + f_{\{1,2,\ldots,d\}}(z^{(1)}, z^{(2)}, \ldots, z^{(d)}) \\ &= \sum_{u\subset\{1,2,\ldots,d\}} f_u(z^{(u)}), \end{aligned} \tag{2.1.13}$$

avec $f(\mathbf{z})$ une fonction de $\mathcal{Z}(\Omega) = [0\,1]^d$ à valeur réelle et $f_\emptyset = f_0 = \int_{[0\,1]^d} f(\mathbf{z}) d_{\mathbf{z}}$ par

définition.

Cette décomposition n'est pas unique du moment où la fonction $f(\mathbf{z})$ reste inchangée si l'on ajoute au terme $f_u(z^{(u)})$ de la décomposition une quantité α et qu'on retranche la même quantité α au terme $f_v(z^{(v)})$ avec $u \neq v$. Ceci pose le problème d'identification des termes $f_u(z^{(u)})$ avec $u \subset \{1, 2, \ldots, d\}$. Sous l'hypothèse des contraintes d'orthogonalité, nous avons la proposition suivante :

Proposition 2.1.3 *Sobol (1993) [142]*
Si $f(\mathbf{z})$ est de carré intégrable sur l'espace certain $\mathcal{Z} = [0\,1]^d$ ($\int_{[0\,1]^d} f^2(\mathbf{z}) d_{\mathbf{z}} < +\infty$) et si

$$\int_{[0\,1]} f_u(z^{(u)}) d_{z^{(j)}} = 0 \quad \forall\, j \in u,$$

alors nous avons :
(i) unicité de la décomposition de l'équation (2.1.13)
(ii) $\forall\, u \subset \{1, 2, \ldots, d\}$

$$f_u(z^{(u)}) = \int_{[0\,1]^{d-|u|}} f(\mathbf{z}) d_{z^{(-u)}} - \sum_{v \subset u} f_v(z^{(v)})$$

(iii) orthogonalité
$\forall\, u \subset \{1, 2, \ldots, d\}$, $\forall\, v \subset \{1, 2, \ldots, d\}$ *et* $u \neq v$

$$\int_{[0\,1]^d} f_u(z^{(u)}) f_v(z^{(v)}) d_{\mathbf{z}} = 0$$

Preuve 2.1.3 *La démonstration de cette proposition est immédiate en utilisant l'hypothèse $\int_{[0\,1]} f_u(z^{(u)}) d_{z^{(j)}} = 0 \quad \forall\, j \in u$. Nous renvoyons à Sobol (1993) [142] ou Antoniadis (1984) [12] ou à Tao (2003) [146] pour une preuve formelle.*

□

Partant de la définition de l'espérance et de l'espérance conditionnelle, remarquons que cette décomposition est analogue à celle présentée dans Efron et Stein (1981) [48]. En effet, la décomposition de Efron et Stein pour des variables aléatoires indépendantes s'écrit :

$$\begin{aligned} f(Z^{(1)}, \ldots, Z^{(d)}) \;=\; & \mathbb{E}\left[f(\mathbf{Z})\right] + \sum_{j=1}^{d} \mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(j)}\right] + \sum_{j_1 < j_2} \{\mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(j_1)}, Z^{(j_2)}\right] \\ & - \mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(j_1)}\right] - \mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(j_2)}\right]\}, \ldots, \end{aligned}$$

Et on a :

$$\begin{aligned} f_0 &= \mathbb{E}\left[f(\mathbf{Z})\right] \\ f_u(Z^{(u)}) &= \mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(u)}\right] - \sum_{v \subset u} \mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(v)}\right] \end{aligned}$$

Notons que cette décomposition est aussi un cas particulier du lemme de Hoeffding (1948) [68] (voir Da-Veiga, 2007 [43] pour la preuve).

En se servant de la propriété d'orthogonalité (iii) des termes $f_u(z^{[u)})$ dans la proposition (2.1.3) et de la décomposition de la fonction (i) dans l'équation (2.1.13) (idem pour celle de Efron et Stein), le théorème suivant est immédiat.

Théorème 2.1.1 *Efron et Stein (1981) [48], Antoniadis (1984) [12], Sobol (1993) [142]*

Si $Z^{(1)}, Z^{(2)}, \ldots, Z^{(d)}$ sont indépendantes alors

$$V = \sum_{j}^{d} V_j + \sum_{j_1 < j_2} V_{j_1,j_2} + \ldots + V_{\{1,2,\ldots,d\}}, \tag{2.1.14}$$

avec

$$\begin{aligned} V &= \mathbb{V}ar\left[f(\mathbf{Z})\right] \\ V_j &= \mathbb{V}ar\left[f_j(Z^{(j)})\right] \\ &= \mathbb{V}ar\left[\mathbb{E}\left(f(\mathbf{Z}) \mid Z^{(j)}\right)\right] \\ V_u &= \mathbb{V}ar\left[f_u(Z^{(u)})\right] \quad \forall u \subset \{1, 2, \ldots, d\} \\ &= \mathbb{V}ar\left[\mathbb{E}\left(f(\mathbf{Z}) \mid Z^{(u)}\right)\right] - \sum_{v \subset u} V_v \end{aligned}$$

Cette décomposition de la variance en des parts de variances expliquées par les différents facteurs conduit à la définition des indices de Sobol suivants :

$$\mathrm{IS}_{z_u} = \frac{V_u}{V} \tag{2.1.15}$$

$$\mathrm{IST}_{Z^{(j)}} = 1 - \frac{V_{-j}}{V} \tag{2.1.16}$$

Précisons que les indices de Sobol décrits ci-dessus ne sont valables que lorsque les facteurs sont non corrélés. Jacques (2005) [80], Da-Veiga [43] illustrent l'insuffisance de la méthode d'ANOVA en présence des facteurs corrélés à l'aide d'exemples analytiques.

ANOVA fonctionnelle avec facteurs corrélés

Dans cette section, nous décrivons une méthode d'analyse de la variance qui prend en compte la corrélation entre les facteurs d'entrée du modèle et qui est proposée par Hooker (2007) [71].

Considérons le modèle d'ANOVA décrit dans l'équation (2.1.13) et modélisons la corrélation entre les facteurs par la fonction $w(\mathbf{z})$ de $[0\,1]^d$ à valeur réelle. La décomposition de l'ANOVA en présence des facteurs corrélés est décrite dans la proposition suivante :

Proposition 2.1.4 *Hooker (2007) [71]*
Si $f(\mathbf{z})$ est de carré intégrable sur l'espace $\mathcal{Z} = [0\,1]^d$ ($\int_{[0\,1]^d} f^2(\mathbf{z})w(\mathbf{z})d_{\mathbf{z}} < +\infty$) et si

$$\int_{[0\,1]} f_u(z^{(u)})w(\mathbf{z})d_{z^{(j)}} = 0 \quad \forall j \in u, \tag{2.1.17}$$

alors nous avons :
(i) unicité de la décomposition de l'équation (2.1.13)
(ii) $\forall\, u \subset \{1, 2, \ldots, d\}$

$$f_u(z^{(u)}) = \int_{[0\,1]^{d-|u|}} f(\mathbf{z})w(\mathbf{z})d_{z^{(-u)}} - \sum_{v \subset u} f_v(z^{(v)})$$

(iii) orthogonalité :
$\forall\, u \subset \{1, 2, \ldots, d\}$, $\forall\, v \subset \{1, 2, \ldots, d\}$et $u \neq v$

$$\int_{[0\,1]^d} f_u(z^{(u)})f_v(z^{(v)})w(\mathbf{z})d_{\mathbf{z}} = 0$$

Preuve 2.1.4 *La démonstration de cette proposition est analogue a celle de Sobol (1993) [142]. L'existence des fonctions $f_u(z^{(u)})$ est donnée par (ii) et l'unicité est assurée par la projection ou par l'unicité de l'intégration. L'orthogonalité est une conséquence directe de la contrainte d'orthogonalité de l'équation (2.1.17). En effet , soient $\forall\, u \subset \{1, 2, \ldots, d\}$, $\forall\, v \subset \{1, 2, \ldots, d\}$ et $u \neq v$. Supposons qu'il existe un $\{i\} \in u$ et $\{i\} \notin u$ et on a immédiatement,*

$$\int_{[0\,1]^d} f_u(z^{(u)})f_v(z^{(v)})w(\mathbf{z})d_{\mathbf{z}}d_{z^{(i)}} = \int_{[0\,1]^{d-1}} f_v(z^{(v)}) \left(\int_{[0\,1]} f_u(z^{(u)})w(\mathbf{z})d_{z^{(i)}} \right) d_{\mathbf{z}} = 0.$$

□

En se basant sur la propriété d'orthogonalité et en posant $\sigma^2 = \int_{[0\,1]^d}(f - f_0)^2(\mathbf{z})w(\mathbf{z})d_{\mathbf{z}}$; $\sigma_u^2 = \int_{[0\,1]^d} f_u^2(z^{(u)})w(\mathbf{z})d_{\mathbf{z}}$ la décomposition de la variance de f s'écrit :

$$\sigma^2 = \sum_{u \subseteq \{1,2,\ldots,d\}} \sigma_u^2$$

Remarquons que le fait de prendre w comme étant le produit des distributions de la loi uniforme sur l'intervalle $[0\,1]$ (cas d'indépendance), nous retombons sur le décomposition de Sobol. La fonction w joue le rôle de pondération entre les facteurs ou une fonction qui mesure la corrélation entre les différents facteurs. Un choix naturel de w serait la distribution jointe des facteurs. w permet aussi la prise en compte de la variabilité locale de certains facteurs notamment dans le cas d'hétéroscédasticité.

La principale difficulté pratique de cette méthode est le choix de la fonction de pondération w entre les différents facteurs du modèle. Cependant, les experts du domaine peuvent fournir des informations à priori pour modéliser w.

ANOVA fonctionnelle généralisée

L'ANOVA classique de Fisher (1958) [52], la décomposition de Sobol (1993) [142] et de Hooker (2007) [71] s'appliquent à divers modèles de différents domaines et consistent à projeter la variable réponse sur une base orthogonale définie à partir du plan d'expérience (cas de l'ANOVA classique) ou une base fonctionnelle adaptée (cas de l'ANOVA fonctionnelle). Naturellement, il est plus intéressant de projeter une fonction périodique sur une base de Fourier et d'une façon générale, de projeter la fonction réponse sur une base bien adaptée aux phénomènes.

Dans cette section nous présentons une analyse de la variance fonctionnelle généralisée de f qui est proposée par Stone, (1994) [143] et qui consiste à projeter la fonction réponse sur une base fonctionnelle bien adaptée. Cette méthode est une extension de l'ANOVA qui prend en compte la particularité de chaque phénomène étudié.

Considérons un espace de fonctions de carré intégrables et de dimensions croissantes défini par
$\mathcal{G} = \left\{ h(\mathbf{z}) = \sum_{u\in\{1,2,\ldots,d\}} h_u(z^{(u)}) \,/\, \int h^2(\mathbf{z}) d_{\mathbf{z}} < +\infty \text{ et } \int_{[0\,1]^d} h_u(z^{(u)}) h_v(z^{(v)}) w(\mathbf{z}) d_{\mathbf{z}} = 0 \right\}$.
L'ANOVA fonctionnelle généralisée de la fonction f revient à chercher une fonction $f^* \in \mathcal{G}$ qui approxime théoriquement la fonction f et s'écrit comme une régression :

$$f^* = \arg\min_{h\in\mathcal{G}} ||f - h||^2 , \tag{2.1.18}$$

et la composante $f_u^* = \overset{\perp}{H}_{\mathcal{G}_u} f$, $\forall\, u \subset \{1, 2, \ldots, d\}$ est obtenue par la projection de la fonction f sur l'espace $\mathcal{G}_u$.

Un cas particulier de cette décomposition est de se donner une base multivariée de fonctions de l'espace $\mathcal{G}$ (voir Tao, 2003 [146] ; Tao et Owen , 2003 [147]). Théoriquement, l'espace $\mathcal{G}$ est engendré par une base $\mathcal{B}$ composée d'une infinité d'éléments $\psi_r(\mathbf{z})\, r \in \mathrm{U}$ où U est un ensemble infini. L'expression de la fonction f dans la base $\mathcal{B}$ et son approximation f^* dans une sous base finie (R $\subset$ U) s'écrivent :

$$\begin{aligned} f(\mathbf{z}) &= \sum_{r\in \mathrm{U}} \beta_r \psi_r(\mathbf{z}) \\ f^*(\mathbf{z}) &= \sum_{r\in \mathrm{R}} \beta_r \psi_r(\mathbf{z}) + \varepsilon(\mathbf{z}) \end{aligned}$$

et nous avons

$$\begin{aligned} \beta_r &= < f, \psi_r > \\ \int_{[0\,1]^d} f^2(\mathbf{z}) d_{\mathbf{z}} &= \sum_{r\in \mathrm{U}} \beta_r^2) \end{aligned}$$

Précisons que l'analyse de la variance dans ces conditions revient à projeter la fonction f sur les différentes fonctions orthogonales ψ_r pour obtenir les coefficients β_r. La somme des carrés de ces coefficients servira à la décomposition de la variance de la fonction f en somme des parts de variances expliquées par des facteurs $Z^{(u)}\, u \in \{1, 2, \ldots, d\}$. Le choix de la base fonctionnelle est crucial dans cette analyse. Différentes bases fonctionnelles usuelles sont présentées dans la Section 2.3.5.

Méthode FAST

La méthode FAST (Fourier Amplitude Sensitivity Test) introduite par Cukier *et al.* (1973) [40], (1975) [42], (1978) [41] et la version EFAST (Extended FAST) introduite par Saltelli *et al.* (1999) [130] sont des méthodes d'analyse de sensibilité robustes en terme de stabilité des indices pour de petites tailles d'échantillonnage. La méthode FAST permet de passer d'une intégrale de dimension d à une intégrale unidimensionnelle pour le calcul des moments du modèle grâce à un changement de variables. Le choix de la fonction de transformation permet d'utiliser la décomposition en série de Fourier pour le calcul des indices.

Considérons l'espérance du modèle $f(\mathbf{Z})$

$$\mathbb{E}[f(\mathbf{Z})] = \int_{[0\,1]^d} f(\mathbf{z}) d_{\mathbf{z}}, \tag{2.1.19}$$

une intégrale multiple à évaluer. A l'aide de la transformation suivante,

$$z^{(j)} = G_j[\sin(\omega_j s)], \quad j \in \{1, 2, \ldots, d\}, \tag{2.1.20}$$

où $s \in [-\pi\, \pi]$, Cukier propose d'approximer l'équation (2.1.19) par :

$$\mathbb{E}[f(\mathbf{Z})] \simeq \frac{1}{2\pi} \int_{[-\pi\, \pi]} f(s) d_s, \tag{2.1.21}$$

avec $f(s) = f[G_1(\sin(\omega_1 s)), G_2(\sin(\omega_2 s)), \ldots, G_d(\sin(\omega_d s))]$ et ensuite utilise la décomposition en série de Fourier de la fonction périodique $f(s)$ pour calculer l'espérance et la variance des sorties du modèle.

La décomposition en série de Fourier de la fonction $f(s)$ s'écrit :

$$f(s) \simeq \sum_{j=-\infty}^{+\infty} A_j \cos(js) + B_j \sin(js)$$

avec

$$A_j = \frac{1}{2\pi} \int_{[-\pi\, \pi]} f(s) \cos(js) d_s,$$

et

$$B_j = \frac{1}{2\pi} \int_{[-\pi\, \pi]} f(s) \sin(js) d_s,$$

En utilisant la propriété d'orthogonalité de la décomposition en séries de Fourier, la variance de $f(\mathbf{Z})$ est approximée par :

$$\begin{aligned} \mathbb{V}\mathrm{ar}[f(\mathbf{Z})] &= \mathbb{E}\left[f^2(\mathbf{Z})\right] - \left[\mathbb{E}\left(f(\mathbf{Z})\right)\right]^2 && (2.1.22) \\ &\simeq \sum_{j=-\infty}^{+\infty} (A_j^2 + B_j^2) - (A_0^2 + B_0^2) && (2.1.23) \\ &\simeq 2\sum_{j=1}^{+\infty} (A_j^2 + B_j^2). && (2.1.24) \end{aligned}$$

En posant $V = 2\sum_{j=1}^{+\infty}(A_j^2 + B_j^2)$ la valeur approchée de la variance de $f(\mathbf{Z})$, remarquons que V peut se décomposer comme suit :

$$V = \sum_{\omega_j \in \mathbb{N}_p} 2\sum_{m=1}^{+\infty} (A_{m\omega_j}^2 + B_{m\omega_j}^2), \tag{2.1.25}$$

avec $V_{z^{(j)}} = 2\sum_{m=1}^{+\infty}(A_{m\omega_j}^2 + B_{m\omega_j}^2)$ la variance expliquée par le facteur $Z^{(j)}$ et $\mathbb{N}_p$ l'ensemble des entiers premiers pour éviter le chevauchement de certaines fréquences qui conduiraient à la confusion des effets des différents facteurs. Les indices de sensibilité sont définis par :

$$\begin{aligned} \mathrm{IS}_{z^{(j)}} &= \frac{2\sum_{m=1}^{+\infty}(A_{m\omega_j}^2 + B_{m\omega_j}^2)}{V} && (2.1.26) \\ \mathrm{IST}_{z^{(j)}} &= 1 - \frac{2\sum_{m=1}^{+\infty}(A_{m\omega_{\sim j}}^2 + B_{m\omega_{\sim j}}^2)}{V}, && (2.1.27) \end{aligned}$$

où $\omega_{\sim j}$ regroupe l'ensemble de fréquences n'appartenant pas à l'ensemble $\{m\omega_j \,/\, m = \{1, 2, \ldots, \}\}$ (Saltelli *et al.*, 1999 [130]).

Seul un petit nombre de fréquences par facteur est considéré pour l'estimation des indices par souci d'économie du nombre d'évaluations du modèle. D'où les indices approchés suivants :

$$\begin{aligned} \tilde{\mathrm{IS}}_{z^{(j)}} &= \frac{2\sum_{m=1}^{M}(A^2_{m\omega_j} + B^2_{m\omega_j})}{V} \\ \tilde{\mathrm{IST}}_{z^{(j)}} &= 1 - \frac{2\sum_{m=1}^{M}(A^2_{m\omega_{\sim j}} + B^2_{m\omega_{\sim j}})}{V}, \end{aligned}$$

où M ($M = 4$ ou $M = 6$) harmoniques sont utilisées selon Cukier *et al.* (1975) [42]. Pour un M donné, le nombre minimum d'évaluations du modèle nécessaire pour calculer les indices du facteur $Z^{(j)}$ vaut

$$n_0 = 2M\omega_{max} + 1,$$

avec ω_{max} la fréquence maximale des fréquences associées aux différents facteurs. Dans ces conditions, le nombre total d'évaluations du modèle pour calculer tous les indices est $n = \prod_{j=1}^{d} n_j = n_0 \times d$.

Notons qu'il est important que les fréquences ω_j $j \in \{1, 2, \ldots, d\}$ assignées à chaque facteur soient distinctes dans la transformation de l'équation (2.1.20). Si deux facteurs $Z^{(j_1)}, Z^{(j_2)}$ ont la même fréquence alors ils ont forcement les mêmes indices de sensibilité selon la définition (2.1.26). Pour le calcul des indices totaux, Saltelli *et al.* (1999) [130] proposent d'attribuer les hautes fréquences aux ω_j et des petites fréquences aux $\omega_{\sim j}$.

Plusieurs fonctions G_j furent proposées dans la littérature (voir Saltelli *et al.*, 2000 [126]) mais nous retenons celle proposée par Saltelli *et al.*, (1999)[130] qui recouvre d'une manière uniforme et au mieux l'espace des facteurs. De plus Saltelli rend sa méthode d'échantillonnage aléatoire en introduisant un paramètre de phase (φ) supplémentaire. Cette fonction s'écrit :

$$z_j(s) = G_j[\sin(\omega_j s)] = \frac{1}{2} + \frac{1}{\pi} arcsin[\sin(\omega_j s + \varphi_j)] \tag{2.1.28}$$

Remarque 2.1.1 *La méthode FAST consiste à projeter la fonction de réponse sur la base de Fourier (voir Section 2.3.5) et à utiliser la variabilité des coefficients de Fourier pour calculer les indices de sensibilité. La variance des coefficients associés à tous les éléments de la base de Fourier de même fréquence ω_j correspond à la variance expliquée par le facteur $Z^{(j)}$.*

Des efforts sur le choix des fréquences ω_j et par conséquent sur le nombre d'évaluations du modèle pour le calcul des indices furent considérés dans Tarantola *et al.* (2006) [148]. Ils proposent un plan d'expérience "Random Balance Design (RBD)" qui permet le calcul des indices principaux pour une valeur quelconque n d'évaluations du modèle. L'idée principale du plan consiste à assigner la même fréquence ω à tous les facteurs et ensuite à faire une permutation aléatoire des points pour évaluer les indices. Pour le calcul des indices principaux du facteur $Z^{(j)}$, la permutation est faite de sorte que les valeurs correspondantes à $Z^{(j)}$ soient ordonnées par ordre croissant.

Remarque sur les différentes méthodes

Les différentes méthodes d'analyse de sensibilité présentées dans cette section sont toutes basées sur la décomposition de la variance. Partant d'une fonction de réponse $f(\mathbf{Z})$ et en se basant sur le fait que la meilleure approximation de la fonction f à un seul facteur $Z^{(j)}$ par rapport à la norme quadratique est $\mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(j)}\right]$, Homma et Saltelli (1996) [70] proposent une définition générale des indices de sensibilité. Intuitivement, le facteur $Z^{(j)}$ est influent sur la sortie du modèle, si la variance de $\mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(j)}\right]$ est importance. Les indices principaux et totaux se définissent :

Définition 2.1.1 *En utilisant la décomposition de la variance suivante*

$$\mathbb{V}ar[f(\mathbf{Z})] = \sum_{u \subseteq \{1,2,\ldots,d\}} \left\{ \mathbb{V}ar\left[\mathbb{E}\left(f(\mathbf{Z}) \mid Z^{(u)}\right)\right] - \sum_{v \subset u} \mathbb{V}ar\left[\mathbb{E}\left(f(\mathbf{Z}) \mid Z^{(v)}\right)\right] \right\},$$

les indices sont définis par :

$$IS_{z^{(u)}} = \frac{\mathbb{V}ar\left[\mathbb{E}\left(f(\mathbf{Z}) \mid Z^{(u)}\right)\right] - \sum_{v \subset u} \mathbb{V}ar\left[\mathbb{E}\left(f(\mathbf{Z}) \mid Z^{(v)}\right)\right]}{\mathbb{V}ar[f(\mathbf{Z})]}$$

$$IST_{z^{(j)}} = 1 - \frac{\mathbb{V}ar\left[\mathbb{E}\left(f(\mathbf{Z}) \mid Z^{(-j)}\right)\right]}{\mathbb{V}ar[f(\mathbf{Z})]}$$

2.1.3 Estimation des indices de sensibilité

L'estimation des indices de sensibilité revient à évaluer les intégrales multiples. Le calcul explicite des intégrales n'étant pas possible dans la majorité des cas, il est indispensable de les approximer par des méthodes de calcul numérique (approximation par la méthode de rectangle, trapèze, polynômes) largement rencontrées dans le calcul d'intégrales. Une meilleure exploration de l'espace et un échantillonnage intensif réduisent largement l'erreur d'approximation et assurent la convergence vers la vraie valeur de l'intégrale.

L'estimation des indices dans le cas de la méthode EFAST revient à évaluer des intégrales unidimensionnelles. Les intégrales unidimensionnelles sont moins coûteuses à évaluer numériquement. Elles nécessitent moins d'évaluations du modèle et sont plus robustes que l'évaluation d'intégrales multidimensionnelles.

Pour l'ANOVA fonctionnelle, Sobol (1993) [142] évalue les $V_u = \mathbb{V}\text{ar}[f_u(Z^{(u)})]\, u \in \{1, 2, \ldots, d\}$ du théorème 2.1.1 par des calculs numériques. En constatant que V_j peut s'écrire aussi comme :

$$\begin{aligned} V_j &= V - \mathbb{E}\left\{\mathbb{V}\text{ar}[f(\mathbf{Z}) \mid Z^{(j)}]\right\} \\ &= \mathbb{E}[f^2(\mathbf{Z})] - \mathbb{E}^2[f(\mathbf{Z})] - \mathbb{E}\left\{\mathbb{E}[f^2(\mathbf{Z}) \mid Z^{(j)}]\right\} + \mathbb{E}\left\{\mathbb{E}^2[f(\mathbf{Z}) \mid Z^{(j)}]\right\} \\ &= \mathbb{E}\left\{\mathbb{E}^2[f(\mathbf{Z}) \mid Z^{(j)}]\right\} - \mathbb{E}^2[f(\mathbf{Z})], \end{aligned}$$

Sobol propose les estimateurs suivants :

$$\begin{aligned} \hat{f}_0 &= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) \\ \hat{V} &= \frac{1}{n}\sum_{i=1}^{n} f^2(\mathbf{z}_i) - \hat{f}_0^2 \\ \hat{V}_j &= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z^{(-j)}}_i, z_i^{(j)}) f(\mathbf{z'^{(-j)}}_i, z_i^{(j)}) - \hat{f}_0^2 \\ \hat{V}_u &= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z^{(-u)}}_i, z_i^{(u)}) f(\mathbf{z'^{(-u)}}_i, z_i^{(u)}) - \hat{f}_0^2 \end{aligned}$$

La convergence de ces estimateurs vers leurs vraies valeurs respectives est prouvée dans Sobol (1993) [142]. L'estimation des indices peut se faire par l'échantillonnage intensif du type Monte Carlo ou les hyper cube latin qui assurent un recouvrement acceptable de l'espace des facteurs. Notons que les estimateurs ainsi proposés nécessitent une évaluation intensive du modèle pour le calcul des indices principaux et totaux. Homma *et al.* (1996) [70] ont modifié certains estimateurs et ont contribué à réduire le nombre de simulations du modèle à $n \times (2d + 2)$ pour le calcul de tous les indices en proposant d'estimer la part de variance expliquée par tous les facteurs sauf le facteur $Z^{(j)}$ V_{-j} par

$$\hat{V}_{-j} = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z^{(j)}}_i, z_i^{(-j)}) f(\mathbf{z'^{(j)}}_i, z_i^{(-j)}) - \hat{f}_0^2$$

Dans la même logique, Saltelli (2002) [124] introduit deux matrices $\mathcal{M}_1$ et $\mathcal{M}_2$ de type $n \times d$ issues de l'échantillonnage de l'espace des facteurs. Il construit ensuite la matrice

$\mathcal{N}_u$ à partir de la matrice $\mathcal{M}_1$ en remplaçant les colonnes de la matrice $\mathcal{M}_1$ dont l'indice figurent dans u par celles de la matrice $\mathcal{M}_2$. Ainsi la matrice $\mathcal{M}_1$ ne diffère de celle de $\mathcal{N}_u$ qu'à travers les colonnes dont l'indice figure dans u. Ceci permet de calculer $\hat{V}_{-u}$. Ainsi $\mathcal{M}_1$ et $\mathcal{N}_j$ permet de calculer $\hat{V}_{-j}$ du facteur $Z^{(j)}$; $\mathcal{M}_1$ et $\mathcal{N}_{-j}$ celui de $\hat{V}_j$.

Remarquons qu'il y a une certaine symétrie entre le couple $(\mathcal{M}_1, \mathcal{N}_{-j})$ et celui $(\mathcal{N}_{\{1,2,\ldots,d\}}$, $\mathcal{N}_j)$ en terme de permutation des matrices $\mathcal{M}_1$ et $\mathcal{M}_2$ pour l'estimation des indices. En se basant sur cette symétrie, Saltelli (2002) [124] propose une méthode de calcul des indices de sensibilité du premier ordre, du second ordre et totaux à un coût de $n \times (d+2)$ en terme d'évaluations du modèle.

L'inconvénient majeur de cette méthode est que les indices peuvent être négatifs et que les indices totaux peuvent être inférieurs aux indices principaux si l'espace des facteurs n'est pas bien couvert par l'échantillonnage utilisé.

Da-Veiga (2007) [43] propose deux méthodes d'estimation des indices de sensibilité : la première méthode d'estimation est basée sur l'estimation non paramétrique de $\mathbb{E}\left[f(\mathbf{Z}) \mid Z^{(j)}\right]$ à l'aide des polynômes locaux. Cette méthode d'estimation des indices permet la prise en compte des facteurs corrélés. L'inconvénient de cette méthode résulte du fait que seules les simulations $(z^{(j)}, f(\mathbf{z}))$ interviennent dans l'estimation des indices du facteur $Z^{(j)}$. La seconde méthode est basée sur l'estimation des opérateurs d'intégrale de densité.

2.1.4 Analyse de sensibilité multivariée

L'analyse de sensibilité sur les modèles dynamique peut se faire sur les différentes sorties du modèle (Saltelli *et al.*, 2000 [129]) et permet d'avoir l'évolution de l'importance d'un facteur tout au long de la dynamique. Pour avoir un indice unique par facteur, il faudra prendre la moyenne ou la moyenne pondérée (Passo *et al.*, 2003 [110] ; Kontoravdi, 2005 [87]). Cette pratique ne tient pas compte de la corrélation qui existe entre les différentes dates. Campbell *et al.*, 2006 [31] proposent d'utiliser l'ACP et d'autres bases pour capter la corrélation qui existent entre les différentes dates. Ils projettent les sorties dynamiques sur les différents éléments de la base et conduisent l'AS sur les coefficients obtenus. Sur chaque composante de la base, ils définissent un indice principal et un indice total par facteur. Cependant, ils ne proposent pas un indice unique par facteur qui tient compte de toute la dynamique.

2.2 Lien entre la qualité du modèle et l'analyse de sensibilité

2.2.1 Sélection des paramètres clés par les indices de sensibilité

L'incertitude sur les facteurs du modèle (inputs $\mathbf{X}$ et paramètres θ) et sur les observations affecte la qualité du modèle. Le nombre relativement élevé de paramètres dans un modèle dynamique par rapport aux observations disponibles nécessite des outils pour sélectionner un certain nombre de paramètres à estimer. La complexité des modèles dynamiques et le manque d'observations rendent difficile la mise en œuvre des techniques habituelles de sélection de modèles ou d'estimation bayésienne.

Les modélisateurs partitionnent le vecteur de paramètres en deux groupes selon des critères subjectifs ou objectifs. Par exemple, Sievanen et Burk (1993) [141] n'estiment pas les paramètres directement mesurables par expérimentation. Des méthodes rigoureuses de sélection des paramètres des modèles sur-paramétrés par rapport aux observations consistent à hiérarchiser les paramètres selon leur degré d'importance à l'aide d'un critère rationnel. Une approche prometteuse pour le choix des paramètres est l'analyse de sensibilité du fait que les paramètres les plus importants sont ceux qui contribuent largement à la réduction de la variabilité de la fonction réponse. Brun *et al.* (2001) [28] proposent une mesure basée sur les indices de sensibilité locaux pour les modèles dynamiques qui permet de sélectionner et de ne garder que les paramètres identifiables. Brun *et al.* (2001) [28], Brun *et al.* (2002) [27] définissent deux mesures :

- le coefficient de colinéarité qui est l'inverse de la plus petite valeur singulière de la sous-matrice des indices de sensibilité associée à un sous-ensemble de paramètres. Plus ce coefficient est grand, moins le sous-ensemble de paramètres est identifiable.

- la seconde mesure est le produit de toutes les valeurs singulières de la sous-matrice des indices pondérée par le cardinal du sous-ensemble de paramètres. Plus cette mesure est grande, plus le sous groupe de paramètres est important et identifiable.

Ruget *et al.* (2001) [122], conduisent une analyse de sensibilisé sur le modèle de culture STICS pour identifier les paramètres clés à estimer. D'autres applications utilisant l'analyse de sensibilité pour identifier les paramètres clés se trouvent dans Makowski *et al.* (2006) [97] ; Ginot *et al.* (2006) [61] ; Lurette *et al.* (2009) [95] ; Lehuger (2009) [93] etc.

2.2.2 Evaluation de la qualité de la procédure

Cette approche largement repondue (sélection des paramètres importants à estimer par analyse de sensibilité décrite dans la Section 2.2.1) pour palier les insuffisances du nombre d'observations pour estimer les paramètres a donné lieu à des travaux pour évaluer leur performance. Tremblay (2004) [151] confronte cette approche avec les procédures classiques de sélection des paramètres tels que l' AIC (Akaike, 1973 [3] ; 1974, [4]) et le BIC (Schwartz, 1978 [134] ; Cavanaugh et Neith, 1997 [36] ; Cavanaugh et Neith, 1999 [35] ; Haughton, 1991 [65] ; Pauler, 1998 [111] ; Mcquarrie, 1999 [102]). L'AS du modèle mini-STICS (LAI ($t = 20$)) détermine 4 paramètres influents. L'étude montre que la sélection de paramètres par ces critères permet de réduire beaucoup plus le $\mathbb{MSEP}$ que l'approche par l'analyse de sensibilité. Remarquons que ce résultat n'est pas étonnant dans la mesure où on estime un nombre différent de paramètres (1 paramètre pour BIC et AIC et 4 pour l'AS) avec le même nombre d'observations (14). De plus, c'est au modélisateur de fixer un seuil en fonction du nombre d'observations pour déterminer le nombre de paramètres influents à estimer. Pour terminer, cette comparaison ne peut avoir lieu que si les données disponibles permettent d'estimer un certains nombre de paramètres (4).

Dans le but d'explorer un biais potentiel de cette pratique, Brun *et al.* (2002) [27], étudient l'effet de la variabilité des paramètres fixés sur les paramètres estimés et montrent que les paramètres estimés dépendent fortement des valeurs des paramètres fixés (20% de variabilité des paramètres fixés entrainent 70% de variation des paramètres estimés pour le modèle considéré dans leur cas d'étude). De plus, ils étudient l'impact des paramètres fixés sur la somme de carré des résidus. Wallach *et al.* (2002) [157] étudient la stabilité des paramètres estimés suite aux variations des paramètres fixés. Ils quantifient aussi l'effet de l'incertitude des paramètres fixés sur le $\mathbb{MSEP}$. Wallach et Goffinet (1987) [155],Wallach et Genard (1998) [154] ; Wallach *et al.* (2002) [157] décomposent l'erreur quadratique de prédiction $\mathbb{MSEP}$ comme étant la somme de trois termes. Avant de présenter ces trois termes, précisons quelques formalismes nécessaires à leur compréhension.

Modélisation de l'incertitude sur les facteurs

Nous distinguons deux types de facteurs : les variables et les paramètres. Considérons $\mathbf{x}$, le vecteur de variables d'entrée du modèle dont les valeurs sont indispensables pour l'usage du modèle. C'est l'exemple en agronomie des variables pédoclimatiques qui sont soit estimées soit calculées soit mesurées et ensuite utilisées dans des modèles. Le vecteur de variables est en réalité inconnu et ces valeurs notées $\mathbf{X}$ sont entachées d'incertitude et

nous avons :

$$\mathbf{X} = \mathbf{x} + \varepsilon_x,$$

où ε_x est le terme d'erreur aléatoire d'espérance nulle et $\mathbf{x}$ est le vrai vecteur de valeurs inconnues.

Considérons θ le vecteur de d paramètres inconnus et supposons que seul un certain nombre p de paramètres peut être estimé. Nous supposons sans perdre de généralité que le vecteur θ_e de p ($p < d$) paramètres doit être estimé et que le reste de paramètres θ_f doit être fixé. La fixation du vecteur θ_f à un certain vecteur de valeurs $\hat{\theta}_f$ est une pré-estimation qui se fait dans la littérature ou par des mesures ou par des experts du domaine. Ceci introduit une erreur que nous modélisons par :

$$\hat{\theta}_f = \theta_f + \varepsilon_f,$$

avec ε_f est le terme d'erreur aléatoire d'espérance nulle.

L'estimation du vecteur θ_e se fait conditionnement aux valeurs du vecteur de paramètres fixées $\hat{\theta}_f$. L'estimateur $\hat{\theta}_e(\hat{\theta}_f)$ de θ_e s'écrit :

$$\hat{\theta}_e(\hat{\theta}_f) = \theta_e(\hat{\theta}_f) + \varepsilon_e,$$

avec ε_e est le terme d'erreur aléatoire d'espérance nulle. En négligeant l'incertitude liée aux variables d'entrée du modèle les facteurs désignent alors las paramètres incertains.

Décomposition de $\mathbb{MSEP}$

Dans ce contexte particulier, le $\mathbb{MSEP}$ est donné par :

$$\mathbb{MSEP} = \mathbb{E}_{\mathbf{X}^*,\hat{\theta}_f,\hat{\theta}_e(\hat{\theta}_f)}\left[\left(y^* - f(\mathbf{X}^*,\hat{\theta}_f,\hat{\theta}_e(\hat{\theta}_f)\right)^2\right],$$

où $f(\mathbf{X}^*,\hat{\theta}_f,\hat{\theta}_e(\hat{\theta}_f)$ représente la prédiction du modèle au point $\mathbf{X}^*$ et y^* l'observation de la grandeur modélisée lorsque le vecteur de variables inconnues vaut $\mathbf{x}^*$.

Wallach et Genard (1998) [154] décomposent le $\mathbb{MSEP}$ comme suit :

$$\begin{aligned}\mathbb{MSEP} &= \mathbb{E}\left[(y^* - \mathbb{E}(y^* \mid \mathbf{X}^*))^2\right] + \mathbb{E}\left[\left(\mathbb{E}(y^* \mid \mathbf{X}^*) - \mathbb{E}(f(\mathbf{X}^*,\hat{\theta}_f,\hat{\theta}_e(\hat{\theta}_f) \mid \mathbf{X}^*)\right)^2\right] \\ &\quad + \mathbb{E}\left[\left(\mathbb{E}(f(\mathbf{X}^*,\hat{\theta}_f,\hat{\theta}_e(\hat{\theta}_f) \mid \mathbf{X}^*) - f(\mathbf{X}^*,\hat{\theta}_f,\hat{\theta}_e(\hat{\theta}_f)\right)^2\right] \qquad (2.2.29)\end{aligned}$$

Le premier terme de la décomposition de l'équation (2.2.29) est connu sous le nom de la variance de la population qui caractérise la réduction du $\mathbb{MSEP}$ si la vraie valeur du vecteur de variables $\mathbf{x}^*$ correspondant à l'observation y^* était connue. Le second terme représente le biais du modèle et devrait en principe s'annuler si le modèle était linéaire et si les estimateurs étaient non biaisés. Le dernier terme qui nous intéresse plus particulièrement dans cette section correspond à toute la variabilité du modèle. A l'aide du développement de Taylor de la fonction $f(\hat{\mathbf{X}}^*, \hat{\theta}_f, \hat{\theta}_e(\hat{\theta}_f))$, Wallach et Genard (1998) [154] affinent cette décomposition en montrant que la variabilité du modèle peut s'écrire comme la somme de variabilités expliquées par le vecteur de variables, le vecteur des paramètres fixés et celui des paramètres estimés. Cette décomposition montre l'existence du biais dû non seulement à la fixation des paramètres mais aussi à la non linéarité du modèle. L'incertitude sur les paramètres fixés et l'incertitude sur le point où l'on souhaite faire la prédiction augmentent le *msep*. La réduction du $\mathbb{MSEP}$ par une telle pratique est alors compliquée et dépend de plusieurs quantités complexes.

2.2.3 Indices de sensibilité prenant en compte la qualité du modèle

Kanso *et al.* (2006) [85] utilisent la distribution a posteriori des paramètres pour modéliser l'incertitude sur les différents paramètres et conduisent l'analyse de sensibilité dans le but de déterminer des paramètres qui permettent de mieux rapprocher les observations aux sorties du modèle. Cette approche associe les données dans l'AS pour déterminer les paramètres les plus influents. Les mêmes données sont utilisées à la fois pour sélectionner les paramètres et pour les estimer. Une approche analogue est présentée dans Ratto *et al.* (2001) [120]. Ratto *et al.* (2001) [120] conduisent l'AS à l'aide de la vraisemblance obtenue par la méthode GLUE (Generalized Likelihood Uncertainty Estimation). Pour un scénario de valeurs des paramètres, la vraisemblance associée à ce scénario est obtenue en prenant l'inverse du $\mathbb{MSEP}$ calculé lorsque le vecteur de paramètres du modèle vaut ce scénario.

Intuitivement la procédure de sélection de paramètres à estimer par l'AS devrait améliorer la qualité du modèle. Bien que ces différentes études commencent par formaliser l'évaluation de cette pratique en étudiant l'effet de l'incertitude des paramètres fixés sur la qualité du modèle $\mathbb{MSEP}$ d'un côté et en faisant l'AS sur des distributions qui tiennent compte du $\mathbb{MSEP}$ de l'autre coté, il est intéressent de pouvoir déterminer dans quelles conditions la sélection basée sur l'analyse de sensibilité contribue à la réduction du $\mathbb{MSEP}$.

2.3 Analyse des données multivariées et réduction de la dimension

L'expression des modèles dynamiques à pas de temps discret représentée dans l'équation (0.0.2) est vue statistiquement comme un vecteur aléatoire et se prête bien à diverses analyses multivariées. La réduction de la dimension est essentielle lorsque nous souhaitons analyser des phénomènes dans de grandes dimensions soit pour éviter des coûts exorbitants de simulations soit pour l'utilisation des simulations en temps réels soit pour une meilleure compréhension et représentation des phénomènes étudiés. Nous nous intéressons dans cette section à la description des méthodes d'analyses de données et des techniques prometteuses de réduction de la dimension utilisées en modélisation dynamique et qui conservent au maximum les propriétés du phénomène d'intérêt.

Le caractère aléatoire est dû aux incertitudes sur des facteurs d'entrée du modèle. Nous distinguons dans cette section deux types de présentation d'un facteur aléatoire à savoir facteur discret et facteur continu. Bien que la nature continue d'un facteur permet de mieux balayer son espace certain, il arrive, pour des raisons économiques (coût en temps d'évaluations de modèles) ou politiques (protocoles expérimentaux) de ne considérer que certaines valeurs des facteurs continus. Ceci conduit à approximer les facteurs continus par des facteurs discrets quantitatifs ou qualitatifs. Il est naturellement possible de mélanger les deux types de facteurs comme le souligne Stone (1994) [143] mais cette configuration ne sera pas traitée ici.

Dans le cas de la modélisation dynamique avec facteurs discrets, théoriquement, il est possible d'explorer tout l'espace des facteurs par une simple combinaison de tous les niveaux des différents facteurs. Les différents scénarios obtenus permettent de réaliser toutes les simulations du modèle qui sont ensuite stockées sous forme d'une matrice de type $N \times T$ si l'on suppose que le nombre de scenarios est N.

En présence des facteurs continus, chaque facteur prend une infinité de valeurs et ceci nécessite un traitement particulier. Néanmoins, un certain nombre N de simulations du vecteur aléatoire est largement utilisé en statistique pour extraire des informations (Dauxois *et al.*, 1982 [44] ; Besse (1992) [21] ; Bosq (2000) [24] ; Hall (2006a) [63]). Cette approximation empirique des réalisations du vecteur aléatoire est une matrice de type $N \times T$.

Les simulations des modèles dynamiques à pas de temps discrets sont stockées comme

suit :

$$\mathcal{Y} = \begin{pmatrix} y_1(1) & \dots & y_1(t) & \dots & y_1(T) \\ \vdots & & \vdots & & \vdots \\ y_i(1) & \dots & y_i(t) & \dots & y_i(T) \\ \vdots & & \vdots & & \vdots \\ y_N(1) & \dots & y_N(t) & \dots & y_N(T) \end{pmatrix},$$

Cette matrice de simulations est adaptée aux différentes analyses multivariées (Anderson *et al.*, 2003 [11], (Saporta (2006) [131]). Un individu ou un scenario ou encore une simulation $\mathcal{Y}_i$ évolue dans l'espace temporelle $\mathbb{R}^T$ des variables et une colonne $\mathcal{Y}_t$ ou une sortie du modèle à la date t évolue dans l'espace des individus $\mathbb{R}^N$.

2.3.1 Mesure de variabilité

La mesure de variabilité dépend naturellement de l'espace dans lequel les analyses se feront. Dans l'espace des individus, chaque sortie du modèle est un vecteur de $\mathbb{R}^N$ et généralement la variabilité des observations $\mathcal{Y}'$ est mesurée par le volume de l'hyper cube construit sur les vecteurs $\mathcal{Y}_t\, t \in \{1, 2, \dots, T\}$ et vaut $\det(\mathcal{Y}'\mathcal{Y})$ (Escoufier, 1973 [49]). Cet espace est plus adapté à la caractérisation des corrélations entre les différentes sorties du modèle $\mathcal{Y}_t\, t \in \{1, 2, \dots, T\}$ grâce au coefficient de corrélation vectorielle introduit par Hotelling (1936) [72]. Dans la suite de ce mémoire, nous ne considérons que l'espace des variables du fait que les sorties des modèles dynamiques sont corrélées d'avance d'une part et du fait que nous cherchons à identifier les effets des différents facteurs à travers toute la dynamique d'autre part.

Inertie et norme matricielle

Considérons $\mathcal{M}_{N,T}(\mathbb{R})$ l'espace des matrices de type $(N \times T)$ à coefficients réels muni du produit scalaire canonique suivant :

$$\forall \mathcal{A} \in \mathcal{M}_{N,T}(\mathbb{R}), \forall \mathcal{B} \in \mathcal{M}_{N,T}(\mathbb{R}), < \mathcal{A} \mid \mathcal{B} >= \mathbb{T}\mathrm{r}(\mathcal{A}'\mathcal{B}),$$

et, la norme canonique ou la norme de Frobenius associée à ce produit scalaire sur l'espace $\mathcal{M}_{N,T}(\mathbb{R})$ est donnée par $||\mathcal{A}||_F^2 = \mathbb{T}\mathrm{r}(\mathcal{A}'\mathcal{A})$.

Classiquement, $\forall\ \mathcal{Y} \in \mathcal{M}_{N,T}(\mathbb{R})$. la mesure de la variabilité de la matrice $\mathcal{Y}$ se fait à l'aide de l'inertie qui mesure la dispersion des individus par rapport à leur centre de gravité et qui est notée $\mathbb{I}(Y)$. C'est une mesure empirique de la variabilité largement utilisée en analyse de données.

Définition 2.3.1 *Soient* $\mathbf{g} = \left[\sum_{i=1}^{N} p_i \mathcal{Y}_{i1}, \sum_{i=1}^{N} p_i \mathcal{Y}_{i2}, \ldots, \sum_{i=1}^{N} p_i \mathcal{Y}_{iT}\right]'$ *le centre de gravité du nuage de points, et* p_i *le poids de la l'individu ou de la simulation* i*, l'inertie de* $\mathcal{Y}$ *par rapport au centre de gravité* $\mathbf{g}$ *vaut :*

$$\mathbb{I}_{\mathbf{g}} = \sum_{i=1}^{N} p_i \left|\left|\mathcal{Y}_i - \mathbf{g}\right|\right|^2$$

Notons que dans cette définition, l'inertie est évaluée au point $\mathbf{g}$ et qu'il est possible de l'évaluer en un point quelconque $\mathbf{a}$. Dans la suite de ce mémoire, nous ne considérons que l'inertie évaluée au centre de gravité du fait de la relation suivante :

$$\mathbb{I}_{\mathbf{a}} = \mathbb{I}_{\mathbf{g}} + \left|\left|\mathbf{a} - \mathbf{g}\right|\right|^2 .$$

Dans le but de construire les indices de sensibilité, nous accordons le même poids à chaque simulation $p_i = \frac{1}{N}$. Ainsi, il est évident de constater que l'inertie de la matrice $\mathcal{Y}$ ($\mathbb{I}$) est la norme de la matrice $\frac{\mathcal{Y}_c}{\sqrt{N}}$ avec $\mathcal{Y}_c$ la matrice obtenue en centrant les colonnes de $\mathcal{Y}$:

$$\begin{aligned} \mathbb{I} &= \left|\left|\frac{\mathcal{Y}_c}{\sqrt{N}}\right|\right|_F^2 \\ &= \mathbb{T}\mathrm{r}\left(\frac{\mathcal{Y}_c}{\sqrt{N}}' \frac{\mathcal{Y}_c}{\sqrt{N}}\right) \end{aligned}$$

Cette équivalence entre la définition de l'inertie et la norme matricielle canonique permet une décomposition de l'inertie en des parts d'inerties associés à des sous espaces orthogonaux comme le montre la proposition suivante :

Proposition 2.3.1 *Saporta (2006) [131]*
Soient $P_1\, P_2\, \ldots\, P_d$ d *matrices de projection de type* $(N \times N)$ *vérifiant* $P_{j_1} P_{j_2} = P_{j_1} \delta_{j_1 = j_2}$ *;* $\forall\, j_1\, j_2 \in \{1, 2, \ldots, d\} et P_j^2 = P_j$ *;* $P_j' = P_j \,\forall\, j \in \{1, 2, \ldots, d\}$ *et* $I_N = \sum_{j=1}^{d} P_j$ *alors on a :*

$$\begin{aligned} \mathbb{I} &= \left|\left|\frac{\mathcal{Y}_c}{\sqrt{N}}\right|\right|_F^2 \\ &= \sum_{j=1}^{d} \left|\left| P_j \frac{\mathcal{Y}_c}{\sqrt{N}}\right|\right|_F^2 \\ &= \sum_{j=1}^{d} \mathbb{I}_{P_j}. \end{aligned}$$

Preuve 2.3.1 *La preuve est une conséquence directe de l'orthogonalité des matrices de projection et du théorème de Pythagore.*

□

Notons que dans cette proposition, l'inertie se décompose comme la somme des parts d'inerties expliqués par les sous espaces associés aux différents projecteurs P_j $j \in \{1, 2, \ldots, d\}$ que l'on pourra affecter aux différents facteurs du modèle.

2.3.2 Analyse en Composante Principale : ACP

L'Analyse en Composante Principale (ACP) introduit par Pearson (1901) [112] est un outil largement utilisé dans l'analyse de données multidimensionnelles (voir Jolliffe 2002 [82] ; Anderson 2003 [11] ; Saporta 2006 [131]) et surtout comme une technique optimale de réduction de la dimension en terme de perte minimale de l'information sur le phénomène étudié. L'ACP est la première approche d'exploration de données en vue d'analyses approfondies qui permet d'identifier les principaux types de courbes que constituent les données. Elle apporte une meilleure information sur la structure de la matrice de variance covariance qui, toute seule, est difficile à interpréter. La réduction de la dimension permet la visualisation des données dans un plan par exemple.

Le principe de l'ACP consiste à déterminer un sous-espace de dimension réduite qui soit optimal au sens où le sous espace retenu déforme le moins possible la projection des données sur ce dernier. Statistiquement, l'ACP revient à chercher une combinaison linéaire (produit scalaire usuelle) des variables ou caractères qui maximisent l'inertie. Formellement, l'ACP revient à chercher une nouvelle base de telle sorte que les premiers axes expliquent le maximum de l'inertie.

Considérons $\mathcal{Y}$ la matrice de données et supposons sans perdre de généralité que ces colonnes sont centrées. Considérons la matrice $\mathcal{V}$ de type $(\mathcal{I} \times \mathcal{I})$ dont les vecteurs colonnes $\mathbf{v}_j$ $j \in \{1, 2, \ldots, T\}$ constituent une base $\mathcal{B}$ de l'espace $\mathbb{R}^T$. La matrice de projection sur la base $\mathcal{B}$ se définit par $\overset{\perp}{H}_{\mathcal{V}}^{\perp} = \mathcal{V}\mathcal{V}'$. La projection de l'individu $\mathcal{Y}_i$ de $\mathbb{R}^T$ sur $\mathcal{B}$ vaut $\mathcal{V}\mathcal{V}'\mathcal{Y}_i$ et la perte d'information associée à cette projection est $\mathcal{Y}_i - \mathcal{V}\mathcal{V}'\mathcal{Y}_i$. La fonction perte $l(\mathcal{Y}, \mathcal{B})$ occasionnée lors de la projection de tous les individus sur la nouvelle base $\mathcal{B}$ s'écrit :

$$\begin{aligned} l(\mathcal{Y}, \mathcal{B}) &= \sum_{i=1}^{N} ||\mathcal{Y}_i - \mathcal{V}\mathcal{V}'\mathcal{Y}_i||^2 \\ &= ||\mathcal{Y}' - \mathcal{V}\mathcal{V}'\mathcal{Y}'||_F^2 \end{aligned} \tag{2.3.30}$$

La recherche de la base $\mathcal{B}$ ou la sous base qui minimise cette perte d'information s'écrit

alors comme un problème d'optimisation ou de régression suivant :

$$\begin{aligned} \hat{\mathcal{V}} &= \arg\min_{\mathcal{V}} ||\mathcal{Y}' - \mathcal{V}\mathcal{V}'\mathcal{Y}'||_F^2 \\ & s.c \quad \mathcal{V}'\mathcal{V} = I \end{aligned} \tag{2.3.31}$$

La matrice $\hat{\mathcal{V}}$ est l'estimateur de la matrice habituelle de vecteurs propres et la $k$ième composante principale est définie par $\mathbf{h}_k = \mathcal{Y}\mathcal{V}_k$. La norme au carré de $\mathbf{h}_k$ représente la $k$ième valeur propre.

Notons que les composantes principales correspondent à la projection des données initiales dans la nouvelle base définie par $\hat{\mathcal{V}}$ et qu'elles constituent de nouvelles variables non corrélées qui résument au maximum l'information. Dans le cas des variables continues ou "sampled" ACP selon Jolliffe (2002) [82], les propriétés de convergence et de consistance des estimateurs des valeurs propres et des vecteurs propres furent étudiées par Dauxois *et al.*, 1982 [44] ; Bosq (2000) [24] ; Hall and Nasab (2006) [64]. Hall *et al.* (2006) [63] proposent un développement du genre Taylor des valeurs propres et des vecteurs propres.

Choix de la dimension

La principale difficulté dans la réduction de la dimension par l'ACP est justement le choix de la dimension acceptable pour mieux conserver l'information. Plusieurs critères de moins élaborés ou heuristiques au plus rigoureux furent considères dans la littérature :

- Les procédures heuristiques couramment utilisées consistent i) à choisir toutes les composantes dont les valeurs propres sont supérieures à un ($\lambda_k > 1$, $\forall\, k \in \{1, 2, \ldots, d\}$) selon le critère de Kaiser-Guttman ii) à représenter les valeurs propres et ensuite sélectionner les valeurs propres qui ne semblent pas être alignées sur une droite (méthode basé sur la rupture de pente ou scree plot de Zebra and Collins, 1992 [162]) iii) à retenir les valeurs propres qui sont supérieures à leurs valeurs critiques fournies par la distribution des bâtons (modèle de la rupture de bâton de Frontier, 1976 [54]) iv) à choisir les k premières composantes dont la somme des valeurs propres est supérieure à un seuil fixé (critère de proportion d'inertie de Jolliffe, 1986 [83]). Généralement les 5 premières composantes expliquent plus de 95% de l'inertie.

- Les approches statistiques sont nombreuses (voir Jolliffe, 2002 [82] ; Jackson, 1991 [79] ; Peres-Neto *et al.*, 2005 [113]) et nous ne considérons ici que des méthodes qui tentent de formaliser les approches heuristiques mentionnées ci-dessus. Anderson (1963) [9] proposa un test de rapport de vraisemblance pour identifier les valeurs propres suffisamment petites pour être négligées. L'hypothèse nulle suppose

une égalité entre les valeurs propres négligées. Ce développement fut élaboré en considérant une distribution multi-normale pour le vecteur de variables.

Besse (1992) [21] proposa un critère plus rigoureux qui garantit la stabilité de la projection des individus et qui permet de choisir la dimension sans aucune hypothèse sur la distribution des données. Le critère proposé est l'écart entre la matrice de projection théorique $\mathcal{V}'\mathcal{V}$ et celle estimée $\hat{\mathcal{V}}'\hat{\mathcal{V}}$. Cette fonction de risque est inversement proportionnelle à l' écart entre les valeurs propres. Ainsi, plus la différence entre les valeurs propres est grande mieux on conserve l'information. Dans ces conditions, le choix de la dimension revient à minimiser le critère de perte de Besse.

La méthode de proportion d'inertie facile à mettre en œuvre est jugée peu fiable par Jackson (1991) [79] pour déterminer le nombre de composantes à retenir sans perdre de l'information ni ajouter du bruit. Cependant, dans le cas de l'AS, ce critère est pertinent dans la mesure où nous cherchons à expliquer la variabilité totale des données. Il est alors essentiel de savoir quel pourcentage de variabilité est négligé. Une étude comparative de l'efficacité et de la robustesse de plus de 20 approches pour le choix de la dimension fut considérée dans Peres-Neto *et al.* (2005) [113], Jackson (1993) [78]. Il ressort de ces études que l'efficacité de ces critères dépend largement du niveau de corrélation des variables et de la taille de la matrice. Les techniques heuristiques semblent être peu performantes.

Variante d'ACP

Les sorties des modèles dynamiques peuvent être considérées comme des séries temporelles particulières et faire l'ACP sur ces sorties peut conduire parfois à des composantes chaotiques difficiles à interpréter. Particulièrement, les sorties des modèles de cultures sont affectées par le climat en générale et la température en particulier. Il est intéressant d'introduire des contraintes de régularités pour lisser les composantes. Ramsay *et al.* (1997 [118], 2002 [119]) proposent diverses méthodes de l'ACP lissée. L'ACP lissée consiste à ajouter une pénalité de régularisation (dérivabilité et continuité) sur les vecteurs propres dans la régression présentée dans l'équation (2.3.31). De même en remplaçant la pénalité L_2 sur les vecteurs propres par la pénalité L_1 ou une combinaison des deux pénalités, Zou *et al.* (2006) [164] proposent le "sparse" PCA qui permet de négliger entièrement la faible contribution de certaines variables dans la construction des composantes principales.

Un autre inconvénient de l'ACP est son incapacité à capter ou à résumer la non linéarité entre les variables initiales. L'astuce pour introduire la non linéarité est de rem-

placer les composantes linéaires usuelles ou de façon équivalente les fonctions linéaires des variables par des fonctions quelconques. En particulier, le "kernel" ACP est obtenu en remplaçant le produit scalaire par un noyau (Dong et McAvoy, 1996 [45]). La difficulté pratique de cette méthode demeure le choix de le fonction ou du noyau pour décrire la non linéarité.

Dans la recherche de bases $\mathcal{V}$ adaptées sur lesquelles il faut projeter les données, une extension de l'ACP consiste à se donner une base puissante et efficace pour approximer les données ou une base sur laquelle il faut faire la régression (Campbell *et al.*, 2006 [31]). Ceci permet de capter la non linéarité et de tenir compte de la vraie structure des données. Si le phénomène étudié est cyclique et périodique, la base de Fourier serait bien adaptée pour capter le maximum de variabilité. L'ACP "spline" en est un exemple dans le cas où les bases splines sont considérées.

Considérons $\psi_1, \dots \psi_d$ une base multivariée quelconque, la projection des sorties de modèles dynamiques sur cette base revient à approximer $f(\mathbf{z}, t)$ par :

$$f(\mathbf{z}, t) \approx \sum_{k=1}^{d} a_k(t)\psi_k(\mathbf{z}). \tag{2.3.32}$$

Cette approximation ouvre la voie vers la régression à coefficients variables ou à coefficients fonctionnels Cai *et al.* (2000) [30]. Les différentes bases et leurs propriétés sont présentées dans la Section 2.3.5. Le grand inconvénient de cette approche est qu'il faudrait une dimension un peu élevée pour capter le maximum de variabilité.

2.3.3 Décomposition en Valeurs Singulières : DVS

La Décomposition en Valeur Singulière (DVS) est une méthode de réduction de la dimension introduite dans les années 1870-1875 (Jolliffe (2002) [82]). C'est une méthode proche de l'ACP mais plus générale que L'ACP. Elle s'applique à tout type de matrices (carrées ou rectangulaires) et est utilisée comme un outil d'inversion de matrices, d'approximation d'une matrice par une autre ayant moins de colonne, de résolution des problèmes d'optimisation sous contraintes (ACP par exemple). La DVS d'une matrice $\mathcal{Y}$ de type $(N \times T)$ consiste à factoriser cette dernière comme étant le produit de trois matrices :

$$\mathcal{Y} = \mathcal{U}\mathcal{S}\mathcal{V}', \tag{2.3.33}$$

où $\mathcal{U}$ est une matrice de type $(N \times T)$, $\mathcal{S}$ et $\mathcal{V}$ sont des matrices carrées $(T \times T)$. Les colonnes de $\mathcal{U}$ sont les vecteurs singuliers dits de gauche et mutuellement deux à deux

orthogonaux. La matrice $\mathcal{S}$ est diagonale et les éléments non nuls de cette matrice sont des valeurs singulières et sont toutes positives. Les colonnes de $\mathcal{V}$ sont des vecteurs singuliers de droite et forment une base orthogonale pour l'expansion de la matrice $\mathcal{Y}$. Dans le cas d'une matrice singulière $\mathcal{Y}$, certaines valeurs singulières sont nulles et ceci permet la réduction de la dimension en supprimant les colonnes nulles de la matrice $\mathcal{S}$. Généralement les valeurs singulières sont ordonnées par ordre décroissant.

Notons que la DVS d'une matrice carrée revient à diagonaliser la matrice en question et que faire une ACP de $\mathcal{Y}$ revient à faire la DVS de $\mathcal{Y}'\mathcal{Y}$. Dans cette analogie, la matrice $\mathcal{V}$ est équivalente à la matrice de vecteurs propres en ACP, $\mathcal{US}$ est équivalente à la matrice des composantes principales et les valeurs singulières sont équivalentes à des valeurs propres. Plus de détails se trouvent dans Wall *et al.* (2003) [153] ainsi qu'aux références mentionnées dans ce papier. La DVS souffre aussi des mêmes insuffisances que l'ACP mentionnées dans la Section 2.3.2

2.3.4 Décomposition Orthogonale Propre (DOP)

La DOP introduite par Kosambi (1943) [88] est un outil puissant d'analyse de données non linéaire qui consiste à déterminer un sous-espace de dimension réduite qui approxime le mieux au sens de l'erreur de projection des processus ou des phénomènes évoluant dans des espaces de grandes dimensions. La DOP est une extension de la DVS qui prend en compte les phénomènes non linéaires (voir Kunsisch et Volkwein, 2001 [89] ; Volkwein, 2008 [152] ; Chatterjee, 2000 [37]).

Considerons $\{\psi_1, \dots \psi_T\}$ une base multivariée quelconque et $\mathcal{Y}$ la matrice des sorties d'un modèle dynamique. La DOP a pour principe de déterminer les k éléments de la base sur lesquels il faudra projeter la matrice $\mathcal{Y}$ et conserver le maximum d'information. Elle se présente sous forme d'un problème d'optimisation :

$$\begin{aligned} \{\hat{\psi}_{i_1} \dots \hat{\psi}_{i_k}\} &= \arg \min_{\psi_{i_1} \dots \psi_{i_l}} \sum_{i=1}^{N} \left|\left| \mathcal{Y}_i - \sum_{t=1}^{k} (\mathcal{Y}_i' \psi_t) \psi_t \right|\right|^2 \\ & s.c \quad \psi_{j_1}' \psi_{j_2} = \delta_{j_1 j_2} \end{aligned} \tag{2.3.34}$$

Notons que le DOP est un voisin proche de l'ACP non linéaire. Dans le cas de l'ACP ou de la DVS, la DOP revient à retenir les k premières composantes principales ou vecteurs singuliers. La DOP devient extrêmement utile dans la régression à coefficient variable du fait que nous disposons souvent d'une base de grandes dimensions et nous ne savons pas à priori quels sont les éléments de la base qui sont les plus pertinents.

2.3.5 Différentes Bases

Cette section est entièrement consacrée à la description des bases fréquemment utilisées et qui peuvent être intéressantes pour modéliser certains phénomènes agronomiques. La biomasse cumulée peut être modélisée par les polynômes de Legendre ; l'émission du N_2O qui comporte des pics comme du signal serait bien décrite par des ondelettes et les phénomènes cycliques et périodiques par des bases de Fourier. Sans perdre de généralité, la description des bases se fera sur l'intervalle $[0\,1]$. Nous ne considérons que des bases univariées dans la mesure où les bases multivariées sont construites en prenant le produit tensoriel des bases univariées (Stone, 1994 [143] ; Huang, 1998 [73] ; Tao, 2003 [146]).

Bases adaptées

Les bases adaptées aux données sont majoritairement obtenues par l'ACP ou la DVS (voir Section 2.3.2, 2.3.3) et correspondent à la matrice des vecteurs propres. Ces bases permettent généralement de résumer le maximum de l'information en faible dimension.

Bases de Fourier

Les bases de Fourier sont bien reconnues pour décrire les phénomènes cycliques, périodiques et réguliers en termes de dérivation. La base de Fourier considérée est :

$$\begin{aligned} \phi_0(x) &= 1 \\ \phi_k(x) &= \sqrt{2}\sin(2k\pi x) \quad \forall\, k > 0 \\ \phi'_k(x) &= \sqrt{2}\cos(k\pi x) \quad \forall\, k > 0 \end{aligned}$$

et la Figure 2.1 illustre les 5 premiers éléments de la base.

Polynômes Orthogonaux

Les polynômes orthogonaux sont des bases régulières "smooth" (Shumaker, 1981[133]) qui permettent naturellement de capter les effets linéaires et quadratiques des phénomènes. En pratique, ces bases permettent de décomposer un phénomène donné comme une somme d'une tendance linéaire et d'une parabole. Parmi les polynômes orthogonaux, nous distinguons les polynômes de Legendre qui sont définis sur l'intervalle $[-1\ 1]$. Nous nous

intéresserons ici aux polynômes définis sur l'intervalle [0 1] et qui s'écrivent :

$$\begin{aligned}
\phi_0(x) &= 1 \\
\phi_1(x) &= \sqrt{3}(2x-1) \\
\phi_2(x) &= \frac{3}{2}\sqrt{5}\left((2x-1)^2 - \frac{1}{3}\right) \\
\phi_3(x) &= \frac{5}{2}\sqrt{7}\left((2x-1)^3 - \frac{3}{5}(2x-1)\right) \\
\phi_{k+1}(x) &= \frac{\sqrt{(2k+3)(2k+1)}}{k+1}(2x-1)\phi_k(x) - \frac{k}{k+1}\sqrt{\frac{2k+3}{2k-1}}\phi_{k-1}(x)
\end{aligned}$$

et la Figure 2.2 illustre les 4 premiers éléments de la base.

Les bases polynômiales peuvent être utilisées pour identifier les caractéristiques grossières des phénomènes et l'interpolation par un modèle polynômial est parfois inconsistante (Schumaker, 1981 [133])

Splines polynômiales

Différentes bases de spline existent notamment les B-splines, les splines polynômiales .... (voir Schumaker, 1981 [133]). Les splines polynômiales considérées dans cette section,

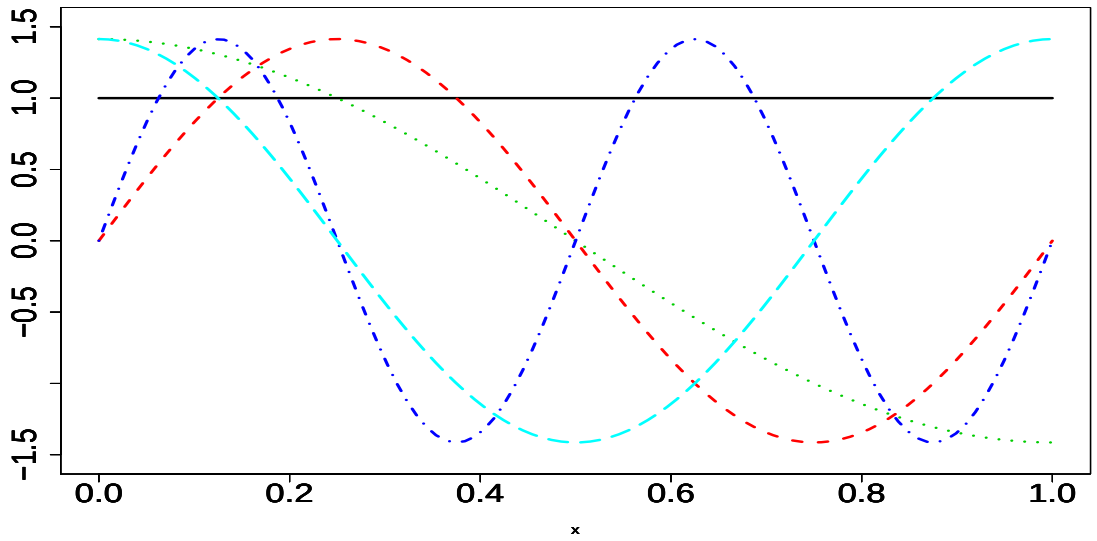

FIGURE 2.1 – Base de Fourier.

sont des polynômes par morceau et furent introduites pour éviter les problèmes d'oscillations rencontrés dans la modélisation avec des polynômes orthogonaux sur des intervalles un peu large.

Considérons m un entier, $[0\,1]$ un intervalle, $t_0 = 0 < t_1 < t_2 < \ldots < t_s = 1$ des réels positifs et $I_j = [t_{j-1}\, t_j] \quad j \in \{1, 2, \ldots, s\}$, une partition de l'intervalle $[0\,1]$. Une fonction f est une spline polynômiale de degré m sur $[0\,1]$ si elle vérifie les deux propriétés suivantes :

(i) f est un polynôme de degré m au plus sur $I_j = [t_{j-1}\, t_j]$, $\forall\, j \in \{1, 2, \ldots, s\}$

(ii) f est de classe C^{m-1} sur $[0\,1]$, $\forall\, m > 1$

Bases d'ondelettes

Les bases d'ondelettes à support compact sont bien adaptées au traitement du signal en particulier et à tout phénomène dont les sorties sont chaotiques, irrégulières avec des pics. Les bases de Haar (bases d'ondelettes spécifiques) sont des fonctions en escaliers continues par morceau mais pas du tout continues (voir Ruskai *et al.*, 1992 [123]). Elles

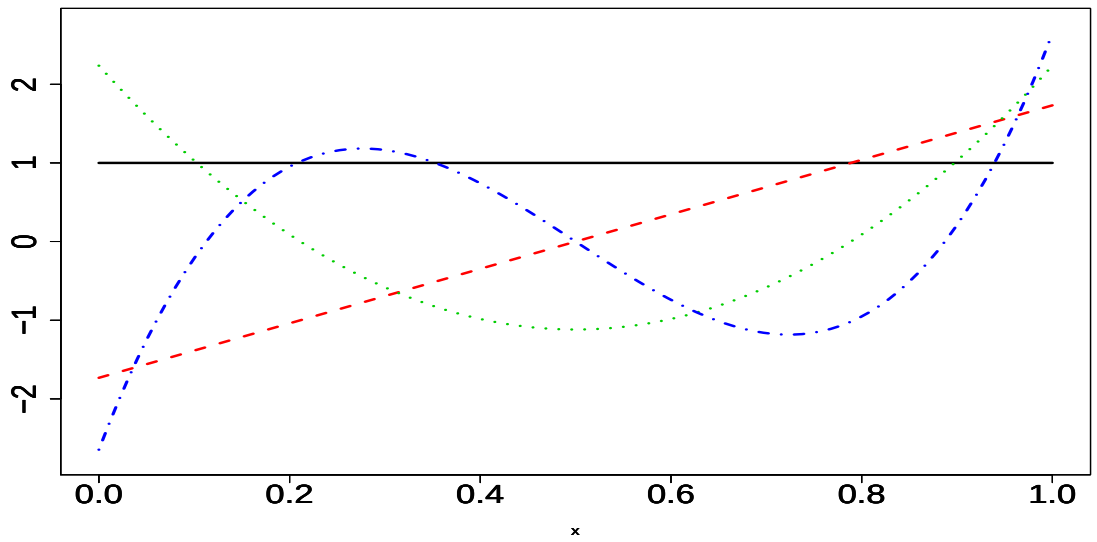

FIGURE 2.2 – Base de polynômes orthogonaux.

sont définies par :

$$\begin{aligned} W_0(x) &= 1 \\ W_{1,0}(x) &= \begin{cases} 1 & \text{si } 0 \leq x \leq 1/2 \\ -1 & \text{si } 1/2 < x \leq 1 \\ 0 & \text{sinon} \end{cases} \\ W_{j,k}(x) &= 2^{(j-1)/2} W_{1,0}\left(2^{j-1}x - k\right) \quad 0 \leq k < 2^{j-1}, \quad j \geq 1 \end{aligned}$$

La Figure 2.3 montre quelques éléments de la base de Haar sur l'intervalle [0 1].

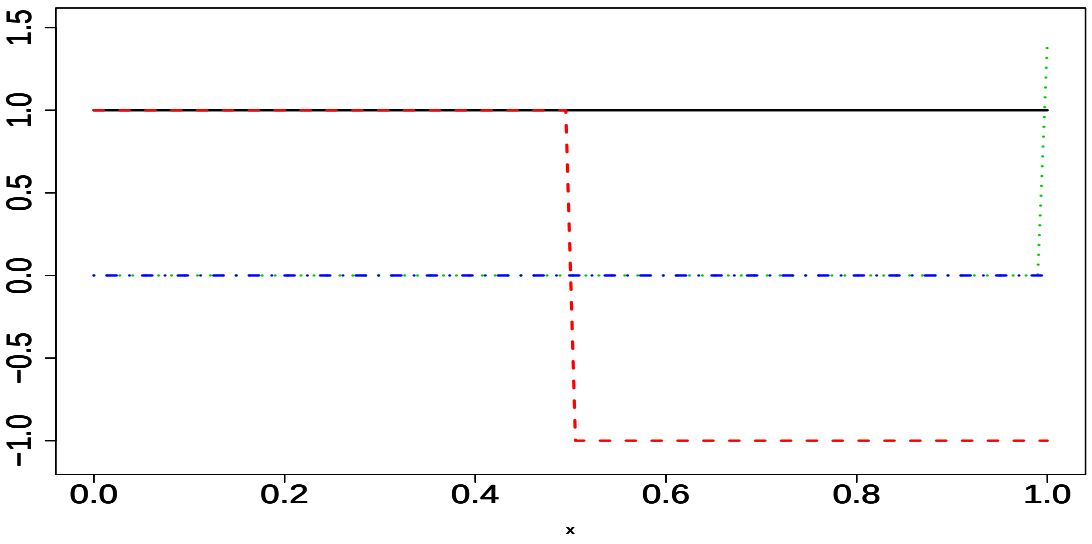

FIGURE 2.3 – Bases de Haar.

2.4 Conclusion

Dans le cas des modèles sur-paramétrés par rapport au nombre d'observations, la sélection du sous groupe de paramètres à estimer se fait soit par des critères objectifs soit par des critères subjectifs. La fixation du reste de paramètres non estimés introduit un biais potentiel sur les estimations. L'incertitude sur les paramètres fixés impacte négativement la qualité du modèle en contribuant à augmenter le $\mathbb{MSEP}$. L'analyse de sensibilité est utilisée comme un critère objectif et puissant de sélection de paramètres dans la mesure où elle consiste à déterminer les paramètres qui influencent le plus la variabilité du modèle. Du fait que l'incertitude sur les paramètres fixés augmente le $\mathbb{MSEP}$

et que les indices de sensibilité cherchent à déterminer les paramètres les plus incertains, il semble logique (intuitif) que la sélection de paramètres à estimer par des indices de sensibilité réduise le $\mathbb{MSEP}$.

Cependant, la complexité de la décomposition du $\mathbb{MSEP}$ et le fait que la mesure de la qualité du modèle se base sur des observations réelles alors que les indices de sensibilité sont issus des simulations rendent moins intuitifs la relation entre les indices de sensibilité et la qualité du modèle. L'analyse de sensibilité n'est pas explicitement reliée à la qualité prédictive du modèle (Wallach et Genard, 1998 [154]) et il est intéressant de pouvoir montrer dans quelles mesures la sélection basée sur l'analyse de sensibilité contribue à la réduction du $\mathbb{MSEP}$.

L'indice de sensibilité dynamique (Saltelli *et al.*, 2000 [129]) permet de voir l'évolution de l'importance d'un paramètre et l'indice moyen ou l'indice moyen pondéré (Passo *et al.*, 2003 [110]) permet d'avoir un indice unique qui ne tient pas compte de la corrélation entre les sorties. Les techniques d'analyses multivariées et de réduction de la dimension permettent la prise en compte des corrélations en construisant de nouvelles sorties non corrélées sur lesquelles seront effectuée l'analyse de sensibilité (Campbell *et al.*, 2006 [31]). La variabilité des dynamiques des modèles est mesurée par l'inertie dans l'espace des variables et est l'équivalent de la variance dans le cas d'un modèle à une seule sortie. En se basant sur les propriétés et les ambitions de l'AS, il intéressant d'utiliser cette métrique pour décomposer la variabilité des modèles dynamiques en une somme de variabilités expliquées par les différents facteurs.

Chapitre 3

Lien entre indices de sensibilité et critères $\mathbb{MSE}$, $\mathbb{MSEP}$ dans le cas d'un modèle linéaire

3.1 Introduction

De nombreux travaux utilisent l'analyse de sensibilité pour sélectionner les paramètres clés à estimer (Ruget *et al.*, 2002 [122] ; Brun *et al.*, 2001 [28] et 2002 [27]). Il est intéressant de quantifier le gain de cette pratique courante en termes du $\mathbb{MSEP}$ (Mean Square Error of Prediction) et du $\mathbb{MSE}$ (Mean Square Error). Nous nous intéressons dans ce chapitre aux liens formels qui existent entre les indices de sensibilité de paramètres et ces critères d'évaluation de modèles.

Afin de formaliser la présentation de ces relations, nous nous plaçons dans un cadre méthodologique proche de celui adopté par les modélisateurs en agronomie et environnement défini par les postulats (que nous ne mettons pas en cause) et les différents concepts suivants :

– **P**1 : le phénomène à prédire est représenté par la fonction de réponse suivante :

$$m = f(\mathbf{x}, \beta), \tag{3.1.1}$$

où $\mathbf{x}$ est le vecteur de variables d'entrée et β le vecteur de paramètres.

Par souci de simplification, le phénomène est supposé connu à l'exception de ses paramètres. Autrement dit, la fonction $f()$, et les variables d'entrée du modèle $\mathbf{x}$ sont connues par les modélisateurs.

- $\mathbf{P}2$: le vecteur de paramètres β est inconnu des modélisateurs mais ces derniers ou des experts du domaine disposent de connaissances leur permettant de fournir des distributions de probabilité pour modéliser l'incertitude sur les différents paramètres. Les experts fournissent les valeurs les plus plausibles pour remplacer les valeurs inconnues des paramètres et ils quantifient le degré d'incertitude associé à chacune de ces valeurs. Dans la suite, la variable aléatoire B_j représente l'incertitude sur le paramètre β_j, $\forall\, j \in \{1, 2, \ldots, d\}$; μ_j est la valeur la plus plausible pour remplacer β_j et σ_j représente le degré d'incertitude sur la valeur μ_j autrement dit σ_j^2 est la variance de B_j. Les variables B_j sont supposées indépendantes d'espérance μ_j et on note $\mathbf{B}$ le vecteur aléatoire dont les composantes sont les B_j.

- $\mathbf{P}3$: le phénomène à prédire est observable selon le modèle statistique suivant :

$$y = f(\mathbf{x}, \beta) + \epsilon, \tag{3.1.2}$$

  où y est une observation du phénomène au point $\mathbf{x}$ et ϵ représente une erreur d'observation. Les erreurs d'observation sont supposées indépendantes et identiquement distribuées, d'espérance nulle, et d'écart type σ_ϵ.

- $\mathbf{P}4$: le modélisateur cherche à prédire la réponse du modèle en un point particulier des variables d'entrée que nous notons $\mathbf{x}^*$. Ceci le contraint à estimer les paramètres et à évaluer son modèle au préalable.

- $\mathbf{P}5$: nous nous plaçons dans le cas où le modélisateur dispose d'un jeu de données $(\mathbf{x}_i, y_i), i \in \{1, 2, \ldots, n\}$ considéré insuffisant pour estimer tous les paramètres inconnus et où il utilise les indices de sensibilité pour sélectionner les paramètres clés à estimer. Nous notons β_e, le vecteur de paramètres à estimer avec les observations disponibles et β_f, les paramètres fixés à des valeurs plausibles notées $\mathbf{b}_f$. Les valeurs $\mathbf{b}_f$ sont supposées être tirées dans la loi des variables aléatoires $\mathbf{B}_f$. Les paramètres β_e seront toujours supposés estimables avec le jeu de données $(\mathbf{x}_i, y_i), i \in \{1, 2, \ldots, n\}$.

La stratégie décrite par les postulats $\mathbf{P}1$-$\mathbf{P}5$ et exposée dans Brun *et al.* (2002) [27]) est largement utilisée dans la littérature pour résoudre le problème d'estimation des modèles

sur-paramétrés et l'un des problèmes est l'évaluation de son efficacité. L'évaluation des modèles dynamiques se faisant généralement à l'aide des critères $\mathbb{IMSEP}$, $\mathbb{MSE}$ et $\mathbb{MSEP}$ définis dans la Section 1.1.4 du Chapitre 1 (Wallach et Genard, 1998 [154] ; Wallach *et al.*, 2002 [157] ; Wallach *et al.*, 2006 [158]), nous utiliserons dans la suite de ce chapitre ces critères pour mesurer la qualité d'un modèle.

L'objectif du Chapitre 3 consiste à évaluer la pertinence de la stratégie décrite par les postulats **P**1-**P**5 dans le contexte très simplifié du modèle linéaire multiple, ou plus précisément à déterminer les conditions dans lesquelles cette stratégie permet d'améliorer la qualité globale du modèle. Nous posons donc le postulat supplémentaire suivant :

P6 : la fonction de réponse est un modèle linéaire multiple défini par :

$$f(\mathbf{x}, \beta) = \mathbf{x}'\beta, \tag{3.1.3}$$

avec $\beta = (\beta_1, \ldots, \beta_d)'$ le vecteur de d vrais paramètres inconnus et $\mathbf{x} = (x_1, \ldots, x_d)'$ le vecteur de variables d'entrée. Pour réaliser une simulation du modèle $f()$ le modélisateur doit fixer le vecteur de paramètres inconnus β à un vecteur de valeurs connues noté $\mathbf{b} = (b_1, b_2, \ldots, b_d)'$, supposé être une réalisation de la distribution $\mathbf{B}$.

Dans un premier temps, nous établissons la relation théorique entre les indices de sensibilité et le $\mathbb{MSE}$, $\mathbb{MSEP}$. Ensuite, une investigation approfondie de cette relation théorique suivra à l'aide de simulations pour mieux préciser les conditions qui garantissent une réduction systématique du $\mathbb{MSEP}$ lorsque les paramètres les plus influents sont estimés c'est-à-dire ceux qui ont les plus grands indices de sensibilité.

Notation

$$\text{Posons } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \text{ et } \mathcal{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}.$$

Supposons sans perte de généralité que les $q < d$ premiers paramètres sont ceux estimés par des observations ($\beta_e = (\beta_1, \beta_2, \ldots, \beta_q)'$) et que ce sont les q paramètres les plus influents, c'est-à-dire qui ont les q plus grands indices de sensibilité. Nous allons estimer le vecteur de paramètres β_e et fixer le reste des paramètres noté $\beta_f = (\beta_{q+1}, \beta_{q+2}, \ldots, \beta_d)'$ à des valeurs $\mathbf{b}_f = (b_j,\ j = q+1, q+2, \ldots, d)$ tirées dans les distributions gaussiennes

décrivant l'incertitude sur des paramètres. Nous utiliserons dans toute la suite l'indice "e" pour désigner les quantités relatives aux paramètres estimés et l'indices "f" pour les termes liés aux paramètres fixés. Nous avons alors les notations suivantes :

$$\beta = (\beta_e, \beta_f)',$$

$$\mathbf{x} = (\mathbf{x}_e, \mathbf{x}_f)' \quad \text{avec} \quad \mathbf{x}_e = (x_1, x_2, \ldots, x_q)' \quad \text{et} \quad \mathbf{x}_f = (x_{q+1}, x_{q+2}, \ldots, x_d)',$$

$$\mathbf{x}^* = (\mathbf{x}_e^*, \mathbf{x}_f^*),$$

$$\mathcal{X} = (\mathcal{X}_e, \mathcal{X}_f).$$

3.2 Indices de sensibilité

Nous utilisons la définition probabiliste des indices de sensibilité globale, basée sur la décomposition de la variance des sorties du modèle. La seule nuance est que, pour des raisons de simplicité, les indices ne sont pas normalisés par la variance marginale des sorties.

Définition 3.2.1 *L'indice de sensibilité ($\mathbb{IS}$) principale globale non normalisé des différents paramètres de la fonction réponse (3.1.3) calculé au point $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_d^*)'$ de variables d'entrée s'écrit :*

$$\mathbb{IS}_{\beta_j} = \mathbb{V}ar\left[\mathbb{E}(\mathbf{x}^{*\prime}\,\mathbf{B}|B_j)\right], \quad \forall\, j \in \{1, 2, \ldots, d\} \tag{3.2.4}$$

Propriété 3.2.1 *Les indices de sensibilité dans le cas particulier du modèle linéaire vérifient les relations suivantes :*

$$\begin{aligned} \mathbb{IS}_{\beta_j} &= \mathbb{V}ar\left(x_j^* B_j + \sum_{k \neq j} x_k^* \mu_k\right) \\ &= x_j^{*2} \mathbb{V}ar(B_j) \\ &= \sigma_j^2 x_j^{*2}. \end{aligned} \tag{3.2.5}$$

Remarquons que les indices de l'équation (3.2.5) sont des fonctions croissantes de σ_j^2 et x_j^*. Le choix des degrés d'incertitudes (σ_j^2) détermine les indices de sensibilité et donc la pertinence des paramètres du modèle. Moins on connaîtra précisément la valeur d'un paramètre, plus on augmentera son degré d'incertitude et plus il fera parti du sous-groupe de paramètres les plus influents. Il faudra avoir recours aux experts du domaine dans le choix des valeurs des paramètres σ_j et μ_j $\forall\, j \in \{1, 2, \ldots, d\}$. Il faudra aussi définir avec précision le point de prédiction pour lequel nous nous intéresserons dans le contexte de l'étude.

3.3 Estimation

Considérons le modèle de régression multivariée :

$$\begin{aligned} y_i &= f(\mathbf{x}_i, \beta) + \epsilon_i, \\ &= \mathbf{x_i}'\beta + \epsilon_i, \quad \text{pour} \quad i \in \{1, 2, \ldots, n\}, \end{aligned} \tag{3.3.6}$$

où $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ est le ième vecteur de valeurs des variables explicatives ou variables d'entrée ; $f(\mathbf{x}_i, \beta)$ est la fonction réponse du modèle correspondant au ième vecteur de variables d'entrée ; y_i désigne l'observation expérimentale associée aux mêmes variables $\mathbf{x}_i$ et ϵ_i représente l'inadéquation entre y_i et $f(\mathbf{x}_i, \beta)$. Le modèle de régression (3.3.6) s'écrit sous la forme matricielle suivante :

$$\mathbf{y} = \mathcal{X}\beta + \varepsilon \tag{3.3.7}$$

L'estimation consiste à chercher les valeurs du vecteur de paramètres β de façon à ce que les sorties de la fonction réponse se rapprochent au mieux des observations. En prenant comme fonction objectif, la fonction perte (Err) ($Err(\beta) = ||\mathbf{y} - \mathcal{X}\beta||^2$) qui mesure l'écart quadratique entre les observations et les sorties de la fonction réponse, l'estimation s'écrit comme un problème d'optimisation :

$$\hat{\beta} = \arg\min_{\beta} Err(\beta). \tag{3.3.8}$$

Sous l'hypothèse que la matrice $\mathcal{X}$ est de plein rang, l'estimateur par la méthode des Moindres Carrés Ordinaires (MCO) de β vaut :

$$\hat{\beta} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathbf{y} \tag{3.3.9}$$

Les différentes hypothèses et propriétés des estimateurs MCO se trouvent dans Azaïs et Bardet (2005) [15].

3.4 Relation entre la qualité du modèle et les indices

3.4.1 Qualité d'estimation

La qualité d'estimation de paramètres est généralement évaluée par le Mean Square Error $\mathbb{MSE}$ et mesure l'écart quadratique moyen entre le vecteur de paramètres estimés et celui des vraies valeurs. Les propriétés décrites ci dessous fournissent une relation entre cette mesure de qualité d'estimation et les indices de sensibilité.

Lemme 3.4.1 *Sous les postulats* ***P***1-***P***6*, l'erreur quadratique moyenne de l'estimateur* $\hat{\beta}_e$ *conditionnelle au fait que* β_f *est fixé à* $\mathbf{b}_f$*, notée* $\mathbb{MSE}[\hat{\beta}_e(\mathbf{b}_f)]$*, vérifie :*

$$\mathbb{MSE}\left[\hat{\beta}_e(\mathbf{b}_f)\right] = \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + ||\mathbf{A}(\beta_f - \mathbf{b}_f)||^2, \tag{3.4.10}$$

avec $\mathbf{A} = \left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\mathcal{X}_e'\mathcal{X}_f$ *une matrice d'ordre* $(q, d-q)$*.*

Preuve 3.4.1

Selon les postulats **P**3, **P**5 et **P**6 Le modèle de régression s'écrit,

$$\mathbf{y} = \mathcal{X}_e\beta_e + \mathcal{X}_f\beta_f + \varepsilon, \tag{3.4.11}$$

et l'estimateur des Moindres Carrés Ordinaires (MCO) de β_e conditionnellement à $\mathbf{b}_f$ est donné par

$$\hat{\beta}_e(\mathbf{b}_f) = \left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\mathcal{X}_e'\left(\mathbf{y} - \mathcal{X}_f\mathbf{b}_f\right). \tag{3.4.12}$$

Le biais d'estimation dû à la fixation de β_f aux valeurs $\mathbf{b}_f$ vaut :

$$\begin{aligned}
\mathbb{B}\text{iais}[\hat{\beta}_e(\mathbf{b}_f)] &= \mathbb{E}[\hat{\beta}_e(\mathbf{b}_f)] - \beta_e \\
&= \left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\mathcal{X}_e'\mathcal{X}_f\left(\beta_f - \mathbf{b}_f\right) \\
&= \mathbf{A}\left(\beta_f - \mathbf{b}_f\right),
\end{aligned} \tag{3.4.13}$$

avec $\mathbf{A} = \left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\mathcal{X}_e'\mathcal{X}_f$ une matrice d'ordre $(q, d-q)$. L'erreur quadratique moyenne ($\mathbb{MSE}$) de l'estimateur $\hat{\beta}_e$ vaut :

$$\begin{aligned}
\mathbb{MSE}\left[\hat{\beta}_e(\mathbf{b}_f)\right] &= \mathbb{E}\left[\left|\left|\hat{\beta}_e(\mathbf{b}_f) - \beta_e\right|\right|^2\right] \\
&= \mathbb{E}\left[\left|\left|\hat{\beta}_e(\mathbf{b}_f) - \mathbb{E}\left(\hat{\beta}_e(\mathbf{b}_f)\right)\right|\right|^2 + \left|\left|\mathbb{E}\left(\hat{\beta}_e(\mathbf{b}_f)\right) - \beta_e\right|\right|^2\right] \\
&= \mathbb{T}\text{r}\left\{\mathbb{E}\left[\left(\hat{\beta}_e(\mathbf{b}_f) - \mathbb{E}\left(\hat{\beta}_e(\mathbf{b}_f)\right)\right)\left(\hat{\beta}_e(\mathbf{b}_f) - \mathbb{E}\left(\hat{\beta}_e(\mathbf{b}_f)\right)\right)'\right]\right\} + ||\mathbf{A}(\beta_f - \mathbf{b}_f)||^2 \\
&= \mathbb{T}\text{r}\left[\mathbb{C}\text{ov}\left(\hat{\beta}_e(\mathbf{b}_f)\right)\right] + ||\mathbf{A}(\beta_f - \mathbf{b}_f)||^2 \\
&= \sigma_\epsilon^2\mathbb{T}\text{r}\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + ||\mathbf{A}(\beta_f - \mathbf{b}_f)||^2
\end{aligned}$$

□

Pour prendre en compte le degré d'incertitude sur $\mathbf{b}_f$, les postulats **P**2, **P**3 et **P**5 permettent de considérer le $\mathbb{MSE}[\hat{\beta}_e(\mathbf{b}_f)]$ de l'équation (3.4.10) comme une réalisation d'une variable aléatoire ($\mathbb{MSE}[\hat{\beta}_e(\mathbf{b}_f)] = \mathbb{E}\left\{\mathbb{MSE}[\hat{\beta}_e(\mathbf{B}_f)] \mid \mathbf{B}_f = \mathbf{b}_f\right\}$). Le $\mathbb{MSE}$ devient une quantité aléatoire et son espérance nous donne la moyenne potentielle des erreurs quadratiques liées à la stratégie suivie. La proposition suivante donne cette espérance en fonction des indices de sensibilité.

Proposition 3.4.1 *Sous les postulats* ***P***1 **-** ***P***6*, l'espérance du critère* $\mathbb{MSE}$ *vaut :*

$$\mathbb{E}\left\{\mathbb{MSE}\left[\hat{\beta}_e(\mathbf{B}_f)\right]\right\} = \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + \sum_{j=q+1}^{d} \gamma_j \mathbb{IS}_{\beta_j} + ||\mathbf{A}\left(\beta_f - \mu_f\right)||^2 \qquad (3.4.14)$$

avec $\mathbf{A} = \left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\mathcal{X}_e'\mathcal{X}_f = (a_{ij})_{1\leq i\leq q, q+1\leq j\leq q-d}$, $\gamma_j = \sum_{i=1}^{q} \frac{a_{ij}^2}{(x_j^*)^2}$.

Preuve 3.4.2

En posant Σ_f la matrice de variance-covariance de $\mathbf{B}_f$ et μ_f son espérance et $\mathbf{A} = \left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\mathcal{X}_e'\mathcal{X}_f = (a_{ij})_{1\leq i\leq q, q+1\leq j\leq q-d}$, nous avons :

$$\begin{aligned}
\mathbb{E}\left\{\mathbb{MSE}\left[\hat{\beta}_e(\mathbf{B}_f)\right]\right\} &= \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + \mathbb{T}r\left\{\mathbb{E}\left[\left(\beta_f - \mathbf{b}_f\right)\left(\beta_f - \mathbf{b}_f\right)'\right]\mathbf{A}'\mathbf{A}\right\} \\
&= \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + \mathbb{T}r\left\{\left[\mathbb{C}\mathrm{ov}\left(\beta_f - \mathbf{b}_f\right) + \mathbb{E}\left(\beta_f - \mathbf{b}_f\right)\mathbb{E}\left(\beta_f - \mathbf{b}_f\right)'\right]\mathbf{A}'\mathbf{A}\right\} \\
&= \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + \mathbb{T}r\left(\mathbf{A}\Sigma_f\mathbf{A}'\right) + \mathbb{T}r\left[\mathbf{A}\left(\beta_f - \mu_f\right)\left(\beta_f - \mu_f\right)'\mathbf{A}'\right] \\
&= \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + \mathbb{T}r\left(\mathbf{A}\Sigma_f\mathbf{A}'\right) + ||\mathbf{A}\left(\beta_f - \mu_f\right)||^2 \\
&= \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + \sum_{i=1}^{q}\sum_{j=q+1}^{d} a_{ij}^2\sigma_j^2 + ||\mathbf{A}\left(\beta_f - \mu_f\right)||^2 \\
&= \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + \sum_{i=1}^{q}\sum_{j=q+1}^{d} \frac{a_{ij}^2}{(x_j^*)^2}\mathbb{IS}_{\beta_j} + ||\mathbf{A}\left(\beta_f - \mu_f\right)||^2 \\
&= \sigma_\epsilon^2 \mathbb{T}r\left[\left(\mathcal{X}_e'\mathcal{X}_e\right)^{-1}\right] + \sum_{j=q+1}^{d} \gamma_j \mathbb{IS}_{\beta_j} + ||\mathbf{A}\left(\beta_f - \mu_f\right)||^2
\end{aligned}$$

□

Remarquons que $\mathbf{A}$ mesure le degré d'orthogonalité entre les variables explicatives associées aux paramètres estimés et celles associées aux paramètres non estimés. Les γ_j sont des pondérations des indices de sensibilité qui dépendent du jeu de données.

3.4.2 Qualité de prédiction

Rappelons que la qualité de la prédiction est souvent évaluée par le critère Mean Square Error of Prediction ($\mathbb{MSEP}$) et que nous nous plaçons dans l'optique d'utiliser le modèle de régression (3.3.6) pour prédire le phénomène au point $\mathbf{x}^*$ c'est à dire au point où l'analyse de sensibilité a été effectuée. Il est en effet important de faire la prédiction dans les mêmes conditions que fut conduite l'analyse de sensibilité du fait que les indices

de sensibilité sont dépendants des variables d'entrée du modèle $\mathbf{x}$.

Les deux propriétés suivantes formalisent la relation entre les indicateurs de pertinence des paramètres et la qualité de prédiction du modèle.

Lemme 3.4.2 *Sous les postulats* ***P***1-***P***6*, l'erreur de prédiction conditionnelle lorsque* β_f *est fixé à* $\mathbf{b}_f$*, notée* $\mathbb{MSEP}[\hat{y}^*(\mathbf{b}_f)]$*, vérifie :*

$$\mathbb{MSEP}\left[\hat{y}^*(\mathbf{b}_f)\right] = \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \left[\mathbf{w}'(\beta_f - \mathbf{b}_f)\right]^2 \tag{3.4.15}$$

où $\mathbf{w}$ *est un vecteur de longueur* $d-q$ *vérifiant* $\mathbf{w} = \left[\mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathcal{X}_e'\mathcal{X}_f - \mathbf{x}_f^{*\prime}\right]'$.

Preuve 3.4.3

La prédiction $\hat{y}^*$ au point x^* conditionnellement aux valeurs de $\mathbf{b}_f$ est définie par

$$\hat{y}^*(\mathbf{b}_f) = \mathbf{x}_e^{*'}\hat{\beta}_e + \mathbf{x}_f^{*'}\mathbf{b}_f, \tag{3.4.16}$$

et le biais de prédiction conditionnel $\mathbb{B}\text{iais}[\hat{y}^*(\mathbf{b}_f)]$ vaut

$$\begin{aligned} \mathbb{B}\text{iais}[\hat{y}^*(\mathbf{b}_f)] &= \mathbb{E}[\hat{y}^*(\mathbf{b}_f)] - y^* \\ &= \left[\mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathcal{X}_e'\mathcal{X}_f - \mathbf{x}_f^{*\prime}\right](\beta_f - \mathbf{b}_f), \end{aligned} \tag{3.4.17}$$

où y^* est la vraie valeur du phénomène que nous cherchons à prédire. Dans le cas particulier du modèle linéaire, le $\mathbb{MSEP}$ s'écrit :

$$\begin{aligned} \mathbb{MSEP}\left[\hat{y}^*(\mathbf{b}_f)\right] &= \mathbb{E}(\hat{y}^*(\mathbf{b}_f) - y^*)^2 \\ &= \mathbb{V}\text{ar}[\hat{y}^*(\mathbf{b}_f)] + \left[\mathbb{B}\text{iais}\left(\hat{y}^*(\mathbf{b}_f)\right)\right]^2 \\ &= \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \left\{\left[\mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathcal{X}_e'\mathcal{X}_f - \mathbf{x}_f^{*\prime}\right](\beta_f - \mathbf{b}_f)\right\}^2 \\ &= \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \left[\mathbf{w}'(\beta_f - \mathbf{b}_f)\right]^2, \end{aligned}$$

avec $\mathbf{w} = \left[\mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathcal{X}_e'\mathcal{X}_f - \mathbf{x}_f^{*\prime}\right]'$

□

Remarquons que la qualité de la prédiction ainsi obtenue dépend des valeurs auxquelles sont fixées le reste des paramètres non estimées $\mathbf{b}_f$. Elle est une fonction de ces valeurs et s'écrit $\mathbb{MSEP}\left[\hat{y}^*(\mathbf{b}_f)\right] = \mathbb{E}\left\{\mathbb{MSEP}\left[\hat{y}^*(\mathbf{B}_f)\right] \mid \mathbf{B}_f = \mathbf{b}_f\right\}$. En faisant varier $\mathbf{b}_f$ aléatoirement selon les distributions $\mathbf{B}_f$ considérées dans le postulat $\mathbf{P}2$ pour prendre en compte le degré d'incertitude sur $\mathbf{b}_f$, le $\mathbb{MSEP}$ devient une variable aléatoire et son espérance est décomposée dans la proposition suivante :

Proposition 3.4.2 *Sous les postulats* ***P***1 **-** ***P***6*, l'espérance du* $\mathbb{MSEP}$ *vaut :*

$$\mathbb{E}\left\{\mathbb{MSEP}\left[\hat{y}^*(\mathbf{B}_f)\right]\right\} = \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \sum_{j=q+1}^{d} \lambda_j \mathbb{IS}_{\beta_j} + \left[\mathbf{w}'(\beta_f - \mu_f)\right]^2, \qquad (3.4.18)$$

avec $\lambda_j = \frac{w_j^2}{(x_j^*)^2}$.

Preuve 3.4.4

Les B_j, $j \in \{1, 2, \ldots, d\}$, étant supposées indépendantes et en posant $\mathbf{w} = (w_{q+1}, w_{q+2}, \ldots, w_d)$, l'espérance de $\mathbb{MSEP}$ vaut :

$$\begin{aligned}
\mathbb{E}\left\{\mathbb{MSEP}\left[\hat{y}^*(\mathbf{B}_f)\right]\right\} &= \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \mathbb{E}\left[\mathbf{w}'(\beta_f - \mathbf{b}_f)\right]^2 \\
&= \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \mathbb{V}\mathrm{ar}\left[\mathbf{w}'(\beta_f - \mathbf{b}_f)\right] + \left[\mathbb{E}\left(\mathbf{w}'(\beta_f - \mathbf{b}_f)\right)\right]^2 \\
&= \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \mathbf{w}'\Sigma_f\mathbf{w} + \left[\mathbf{w}'(\beta_f - \mu_f)\right]^2. \\
&= \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \sum_{j=q+1}^{p} w_j^2\sigma_j^2 + \left[\mathbf{w}'(\beta_f - \mu_f)\right]^2 \\
&= \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \sum_{j=q+1}^{p} \frac{w_j^2}{(x_j^*)^2}\mathbb{IS}_{\beta_j} + \left[\mathbf{w}'(\beta_f - \mu_f)\right]^2 \\
&= \sigma_\epsilon^2 \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathbf{x}_e^* + \sum_{j=q+1}^{p} \lambda_j \mathbb{IS}_{\beta_j} + \left[\mathbf{w}'(\beta_f - \mu_f)\right]^2,
\end{aligned}$$

avec $\lambda_j = \frac{w_j^2}{(x_j^*)^2}$

□

Lorsque certains paramètres des modèles sont fixés à des valeurs nominales, la proposition 3.4.2 montre que l'erreur quadratique de la prédiction se décompose en trois termes :

- le premier terme de l'équation (3.4.18) est le terme d'erreur dû à la variance d'estimation des paramètres. Ce terme d'erreur, qui dépend du choix des paramètres à estimer au travers du jeu de données, est par contre irréductible par l'analyse de sensibilité effectuée au point $\mathbf{x}^*$.

- le second terme est une fonction croissante des indices de sensibilité des paramètres non estimés, toutes choses égales par ailleurs. Moins les paramètres fixés sont importants, plus le modèle est acceptable pour faire de la prédiction. Cette tendance semble confirmer la pratique courante qui consiste à estimer les paramètres les plus influents déterminés grâce à l'analyse de sensibilité. Mais les indices dans cette expression sont pondérés par la qualité des variables explicatives (λ_j) que nous

décrivons dans la Section 3.4.3.

– le dernier terme provient de l'écart entre les vraies valeurs des paramètres non estimés et l'espérance des distributions utilisées pour décrire l'incertitude sur les facteurs. Ce terme représente le biais que nous commettons en fixant les paramètres non estimés à leurs valeurs les plus plausibles. Plus cet écart est important, plus la qualité du modèle se dégrade. En pratique, l'avis des experts permet de réduire ce biais car logiquement, plus nous avons de l'incertitude sur un paramètre, plus nous augmentons sa chance d'être sélectionné et d'être estimé.

3.4.3 Cas particuliers

Le second terme de l'erreur moyenne de prédiction décrite dans l'équation (3.4.18) dépend non seulement des indices de sensibilité mais aussi des variables explicatives utilisées pour estimer β_e et du point où la prédiction est faite. Notons que les pondérations λ_j, $j \in \{q+1, \ldots, d\}$, varient en fonction des observations et il est intéressant de savoir ce qu'ils représentent. Pour mieux comprendre et interpréter ces pondérations, nous distinguons plusieurs cas :

Cas 1 : confusion d'effets

Si les valeurs $\mathcal{X}_e$ sont très proches des valeurs $\mathcal{X}_f$ (fortement corrélées), et si $\mathbf{x}_e^*$ est également proche de $\mathbf{x}_f^*$, alors l'erreur moyenne de prédiction dépend très faiblement des indices de sensibilité associés aux paramètres non estimés d'une part et du biais introduit en fixant les paramètres non estimés à leurs valeurs les plus plausibles d'autre part. En effet, dans le cas où, $\mathbf{x}_e^* \simeq \mathbf{x}_f^*$, et $\mathcal{X}_f \simeq \mathcal{X}_e$ on a par conséquent, $\mathbf{w} \simeq 0$. De même si $\mathbf{x}_f^{*\prime} \simeq \mathbf{x}_e^{*\prime}(\mathcal{X}_e'\mathcal{X}_e)^{-1}\mathcal{X}_e'\mathcal{X}_f$ alors $\mathbf{w} \simeq 0$. Il s'agit en quelque sorte d'un mécanisme de compensation, ou encore de confusion, entre les paramètres. Dans ce cas, il est possible de réduire le $\mathbb{MSEP}$ (mais pas le $\mathbb{MSE}$) en estimant les paramètres ayant de faibles indices de sensibilité. Inversement, si ces valeurs sont très différentes, alors il est pertinent d'estimer les paramètres ayant l'indice de sensibilité le plus élevé.

Cas 2 : orthogonalité

Rappelons que $\mathbf{A}$ mesure le degré d'orthogonalité entre les variables explicatives associées aux paramètres estimés et celles associées aux paramètres non estimés. S'il y a orthogonalité entre les colonnes de la matrice $\mathcal{X}$ ou, de façon moins contraignante, si $\mathcal{X}_e$ est orthogonale à $\mathcal{X}_f$ alors la matrice $\mathbf{A}$ dans l'équation (3.4.13) est nulle et le vecteur $\mathbf{w}$ vérifie

$\mathbf{w} = -\mathbf{x}_f^*$ (équation (3.4.15). Dans ces conditions, non seulement nous réduisons l'erreur d'estimation de l'équation (3.4.10) mais également nous rendons l'équation (3.4.14) indépendante des indices de sensibilité des paramètres fixés et du biais introduit en fixant β_f. De plus, nous affectons le même poids ($\lambda_j = 1 \quad \forall\, j \in \{q+1, \ldots; d\}$) aux indices des différents paramètres non estimées de l'équation (3.4.18). Ce qui assure la réduction du $\mathbb{MSEP}$ en estimant les paramètres les plus influents.

Cas 3 : $d = 2$ et $q = 1$

Considérons un modèle de régression avec deux variables explicatives ($d = 2$) et supposons que le premier paramètre a l'indice le plus grand. Les quantités $\mathbf{A}$, γ, $\mathbf{w}$ et λ valent :

$$\mathbf{A} = \frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i1}^2} \tag{3.4.19}$$

$$\gamma = \left(\frac{\sum_{i=1}^n x_{i1}x_{i2}}{x_2^{*2}\sum_{i=1}^n x_{i1}^2}\right)^2 \tag{3.4.20}$$

$$\mathbf{w} = x_1^*\frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i1}^2} - x_2^* \tag{3.4.21}$$

$$\lambda = \left(\frac{x_1^*\sum_{i=1}^n x_{i1}x_{i2}}{x_2^*\sum_{i=1}^n x_{i1}^2} - 1\right)^2 \tag{3.4.22}$$

Notons que cet exemple simple illustre l'interprétation de $\mathbf{A}$ et $\mathbf{w}$. En fait, $\mathbf{A}$ mesure à un coefficient près la corrélation linéaire entre les deux variables explicatives. Plus précisément, en posant τ_1 (resp. τ_2) l'écart type de la variable explicative X_1 (resp. X_2) et ρ le coefficient de corrélation entre les deux variables explicatives, $\mathbf{A}$ devient $\mathbf{A} = \rho\frac{\tau_1}{\tau_2}$. En se plaçant dans le cadre de la régression normalisée où toutes les variables explicatives ont la même variance ($\tau_1 = \tau_2$), $\mathbf{A}$ devient le coefficient de corrélation entre les deux variables explicatives ($\mathbf{A} = \rho$). Ainsi, nous avons les égalités suivantes $\gamma = \rho^2 x_2^{*-4}$, $\mathbf{w} = \rho x_1^* - x_2^*$ et $\lambda = \left(\frac{x_1^*}{x_2^*}\rho - 1\right)^2$. La pondération γ est alors le carré du coefficient de corrélation à un coefficient près et la pondération λ mesure l'écart quadratique entre le coefficient de corrélation et sa valeur maximale. La corrélation négative aura tendance à surpondérer l'indice du paramètre fixé contrairement à une corrélation positive si x_1^* et x_2^* sont de mêmes signe et vice versa.

Cas 4 **:** $d = 3$ **et** $q = 1$

En estimant un seul paramètre sur les trois considérés, les quantités $\mathbf{A}$, γ, $\mathbf{w}$ et λ s'écrivent,

$$\mathbf{A} = \left(\frac{\sum_{i=1}^{n} x_{i1}x_{i2}}{\sum_{i=1}^{n} x_{i1}^2}, \frac{\sum_{i=1}^{n} x_{i1}x_{i3}}{\sum_{i=1}^{n} x_{i1}^2} \right)' \tag{3.4.23}$$

$$\gamma = \left[\left(\frac{\sum_{i=1}^{n} x_{i1}x_{i2}}{x_2^* \sum_{i=1}^{n} x_{i1}^2} \right)^2, \left(\frac{\sum_{i=1}^{n} x_{i1}x_{i3}}{x_3^* \sum_{i=1}^{n} x_{i1}^2} \right)^2 \right]' \tag{3.4.24}$$

$$\mathbf{w} = \left(x_1^* \frac{\sum_{i=1}^{n} x_{i1}x_{i2}}{\sum_{i=1}^{n} x_{i1}^2} - x_2^*, \frac{x_1^* \sum_{i=1}^{n} x_{i1}x_{i3}}{\sum_{i=1}^{n} x_{i1}^2} - x_3^* \right)' \tag{3.4.25}$$

$$\lambda = \left[\left(\frac{x_1^* \sum_{i=1}^{n} x_{i1}x_{i2}}{x_2^* \sum_{i=1}^{n} x_{i1}^2} - 1 \right)^2, \left(\frac{x_1^* \sum_{i=1}^{n} x_{i1}x_{i3}}{x_3^* \sum_{i=1}^{n} x_{i1}^2} - 1 \right)^2 \right]' \tag{3.4.26}$$

Cas 5 **:** $d = 4$ **et** $q = 2$

En estimant deux paramètres sur quatre, nous avons :

$$a_{13} = \sum_{i=1}^{n} x_{i2}^2 \sum_{i=1}^{n} x_{i1}x_{i3} - \sum_{i=1}^{n} x_{i1}x_{i2} \sum_{i=1}^{n} x_{i2}x_{i3}$$

$$a_{14} = \sum_{i=1}^{n} x_{i2}^2 \sum_{i=1}^{n} x_{i1}x_{i4} - \sum_{i=1}^{n} x_{i1}x_{i2} \sum_{i=1}^{n} x_{i2}x_{i4}$$

$$a_{23} = \sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}x_{i3} - \sum_{i=1}^{n} x_{i1}x_{i2} \sum_{i=1}^{n} x_{i1}x_{i3}$$

$$a_{24} = \sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}x_{i4} - \sum_{i=1}^{n} x_{i1}x_{i2} \sum_{i=1}^{n} x_{i1}x_{i4}$$

$$\mathbf{A} = \frac{1}{\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2 - (\sum_{i=1}^{n} x_{i1}x_{i2})^2} \begin{pmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{pmatrix} \tag{3.4.27}$$

$$\gamma = \frac{1}{\left[\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2 - (\sum_{i=1}^{n} x_{i1}x_{i2})^2\right]^2} \left(\frac{a_{13}^2 + a_{23}^2}{(x_3^*)^2}, \frac{a_{14}^2 + a_{24}^2}{(x_4^*)^2} \right)' \tag{3.4.28}$$

$$\mathbf{w} = \frac{1}{\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2 - (\sum_{i=1}^{n} x_{i1}x_{i2})^2} \left(x_1^* a_{13} + x_2^* a_{23} - x_3^*, x_1^* a_{14} + x_2^* a_{24} - x_4^*\right)' \tag{3.4.29}$$

$$\lambda = \frac{1}{\left[\sum_{i=1}^{n} x_{i1}^2 \sum_{i=1}^{n} x_{i2}^2 - (\sum_{i=1}^{n} x_{i1}x_{i2})^2\right]} \left[\left(\frac{x_1^* a_{13} + x_2^* a_{23}}{x_3^*} - 1 \right)^2, \left(\frac{x_1^* a_{14} + x_2^* a_{24}}{x_4^*} - 1 \right)^2 \right]'$$

Plus une variable explicative issue du sous groupe de paramètres non estimés est fortement corrélée avec une autre issue du sous groupe de paramètres estimés, plus la

pondération des indices diffèrent. Dans ce cas, il est possible de réduire le $\mathbb{MSEP}$ en estimant les paramètres ayant de faibles indices par un mécanisme de confusion d'effets.

3.5 Simulation

En dehors des cas extrêmes de l'orthogonalité et de la confusion des effets, la relation entre les indices de sensibilité et le $\mathbb{MSEP}$ dépend des pondérations λ associées aux différents paramètres fixés. Ces pondérations dépendent des corrélations entre les variables explicatives, qui changent d'un jeu de données à un autre. En pratique, nous disposons souvent d'un jeu de données corrélé. Du fait qu'une grande valeur de la pondération λ augmente le $\mathbb{MSEP}$ selon l'équation (3.4.18), nous réalisons des simulations pour évaluer les valeurs des pondérations en se plaçant dans des conditions réelles où les variables explicatives sont partiellement corrélées. Une étude comparative des valeurs des pondérations λ et des indices de sensibilité à travers ces simulations permet de vérifier si l'estimation d'un paramètre ayant un faible indice de sensibilité pourrait réduire beaucoup plus le $\mathbb{MSEP}$ que supposé intuitivement. Ces simulations visent aussi à comparer la stratégie décrite par les postulats **P**1-**P**6 à la méthode de sélection et d'estimation des paramètres LASSO. Pour faire cette comparaison nous utilisons l'un des modèles qui a été présenté avec la méthode LASSO dans Tibshirani (1996) [150]. Nous nous plaçons dans les conditions de cet article pour générer les données sauf qu'en accord avec le postulat **P**5, nous allons prendre la taille n considérée petite pour estimer tous les paramètres. Ceci ne met pas en cause la méthode LASSO du fait que cette méthode s'applique pour des tailles très petites par rapport au nombre de paramètres.

3.5.1 Modèle et données simulées

Nous considérons les conditions décrites dans l'article de Tibshirani (1996) [150] pour simuler les observations. Nous considérons un modèle linéaire particulier du postulat **P**6 pour lequel le vecteur de vrais paramètres vaut $\beta = [3, 1.5, 0, 0, 2, 0, 0, 0]'$. Les réalisations des variables explicatives ($\mathcal{X}$) sont générées suivant une loi normale d'espérance nulle et de matrice de variance - covariance définie comme suit :

$$\mathbb{C}\mathrm{ov}(X_{j_1}, X_{j_2}) = 3 \times \rho^{|j_1 - j_2|}, \tag{3.5.30}$$

dans le but de pouvoir introduire de petites corrélations entre les variables explicatives. Les variables X_i et X_j sont les plus corrélées lorsque $j_1 = j_2 + 1$ ou $j_2 = j_1 + 1$ et cette corrélation vaut $\rho = 0.5$.

Les observations ($\mathbf{y}$) sont générées suivant l'équation :

$$\mathbf{y} = \mathcal{X}\beta + \varepsilon, \tag{3.5.31}$$

avec ε un vecteur gaussien de moyenne nulle et de matrice de variance-covariance $\Sigma_\epsilon = 3\times\mathbf{I}$ avec $\mathbf{I}$ la matrice identité.
La taille des observations est fixée à $n = 15$ conformément au postulat $\mathbf{P}5$.

3.5.2 Point de prédiction

Le point $\mathbf{x}^*$ ($\mathbf{P}4$) auquel s'effectuera la prévision est fixé à une réalisation de la loi normale $\mathcal{N}(0, 3)$ et vaut $\mathbf{x}^* = [0.537, -0.266, 1.026, 0.178, 0.970, -1.397, -0.889, 0.981]'$. Nous avons choisi la loi $\mathcal{N}(0, 3)$ de manière à pouvoir tirer un $\mathbf{x}^*$ qui soit dans le même espace que les variables explicatives. Nous nous plaçons dans le cadre d'une interpolation. Pour des raisons de simplicité (de temps en fait) nous avons fixé définitivement $\mathbf{x}^*$ dans toutes les simulations considérées dans cette section sans le faire varier.

3.5.3 Méthodes d'analyse simulées

Le calcul des indices de sensibilité se fait à l'aide de la formule de l'équation (3.2.5). Selon le postulat $\mathbf{P}2$, les modélisateurs doivent fixer la valeur la plus plausible μ_j et le degré d'incertitude σ_j sur chacun des paramètres β_j. Pour les simulations présentées ici, nous partons du principe que les modélisateurs évaluent l'incertitude de façon globalement correcte. Le degré d'incertitude σ_j est donc choisi comme étant la valeur absolue de la différence entre la vraie valeur β_j du paramètre et sa valeur considérée la plus plausible μ_j. Dans notre cas, nous avons choisi $\mu_j = 0$, et donc $\sigma_j = \|\beta_j\|$, puisque le modèle choisi pour le phénomène à prédire est un modèle creux. Remarquons que si un β_j est nul alors son degré d'incertitude (σ_j) est aussi nul. Ce qui suppose que nous connaissons avec certitude la vraie valeur du paramètre. Pour éviter cette situation non réaliste, nous ajoutons un σ_0 à tous les σ_j. σ_0 est alors le plus petit degré d'incertitude que nous avons sur la valeur d'un paramètre inconnu. Dans le cas des simulations nous l'avons fixé à 0.005. Par ailleurs les distributions B_j sont supposées normales :

$$B_j \sim \mathcal{N}(\mu_j, \sigma_j^2) \quad \forall\, j \in \{1, 2, \ldots, d\}.$$

Dans la procédure de sélection des paramètres à estimer par les indices de sensibilité, il est indispensable de se donner un seuil afin de pouvoir déterminer le sous groupe de paramètres à estimer. En effet, les indices de sensibilité fournissent uniquement un classement des paramètres. De même, pour la méthode LARS (LASSO) (Zou *et al.*, 2005

[163]), il est nécessaire de fixer la pénalité pour sélectionner les paramètres. Dans cette étude, nous fixons les différents seuils ou pénalités par la validation croisée (voir Section 1.1.4). Pour chaque valeur du seuil, nous subdivisons l'échantillon en 5 groupes et nous utilisons 4 groupes pour faire l'estimation et le dernier groupe pour faire la prédiction. A la fin de la procédure, nous retenons la valeur du seuil qui minimise le $\mathbb{MSEP}$ estimé par cette procédure. Durant le processus de la validation croisée, nous utilisons les mêmes groupes aussi bien pour la méthode LASSO que pour la méthode basée sur les indices de sensibilité. Nous faisons de même pour la prédiction.

Tous les résultats des simulations sont obtenus en utilisant le logiciel statistique R.

3.5.4 Résultats

Afin de stabiliser les résultats issus des simulations, nous générons 1000 échantillons de taille $n = 15$. En d'autres termes, à chaque jeu de simulation, nous disposons de 15 observations pour estimer 8 paramètres. Sur chacune des simulations, nous appliquons quatre méthodes d'estimation : MCO, LASSO, sélection des paramètres par l'analyse de sensibilité suivie de l'estimation par la méthode MCO (AS+MCO) et enfin la sélection des paramètres suivie de l'estimation par la méthode LASSO (AS+LASSO). Les résultats figurent dans les Tables 3.1 et 3.2.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Indices ($\mathbb{IS}$ en %)	30.1	3.71	0.182	0.005	65.35	0.338	0.13	0.16
Nombre de fois fixé	0	474	857	982	0	728	954	918
λ_{min}	0	0	0	0	0	0	0	0
λ_{mean}	0	1.30	0.78	10.30	0	1.38	1.51	1.15
λ_{max}	0	31.12	6.28	350.82	0	4.77	15.57	14.94
Estimations (AS+MCO)	3.06	1.364	-0.02	0.006	2.015	0.012	0.005	-0.006
Vraies valeurs des paramètres	3	1.5	0	0	2	0	0	0

TABLE 3.1 – Comparaison des pondérations des indices de sensibilité. λ_{min} correspond à la valeur minimale des 1000 pondérations ; λ_{mean} à leur moyenne et λ_{max} à la valeur maximale. Les estimations présentées sont les moyennes des 1000 estimations obtenues par la stratégie AS+MCO.

Méthodes d'estimation	$\mathbb{MSEP}$	Meilleure méthode (%)
MCO	6.55	15.4
LASSO	2.66	28.1
AS+MCO	2.04	32.2
AS+LASSO	2.01	28.5

TABLE 3.2 – Valeurs moyennes du $\mathbb{MSEP}$ pour 4 méthodes d'estimation et proportions des cas où la méthode a la plus petite valeur du $\mathbb{MSEP}$.

Dans la Table 3.1 résumant les statistiques des analyses de sensibilité, remarquons que les paramètres β_1, β_5 ayant les plus grands indices ne sont jamais fixés et que les paramètres β_3, β_4, β_7, β_8 sont fixés dans plus de 80 % des 1000 simulations réalisées. Pour les 4 paramètres fixés, notons une différence significative entre les différentes valeurs de pondérations λ. En moyenne, la pondération λ_4 correspondant au paramètre β_4 est 10 fois plus importante que celles de λ_3, λ_7, λ_8 et pour certaines simulations (valeurs maximales des λ) ce rapport atteint 20 fois plus. Cette différence significative entre les valeurs de λ peut dégrader la réduction du $\mathbb{MSEP}$ lorsque les paramètres les plus influents sont sélectionnés pour l'estimation. Partant de la relation (3.4.18), l'estimation des paramètres ayant des indices faibles β_6 et β_7 ($\mathbb{IS}_{\beta_6} = 0.34\%$ et $\mathbb{IS}_{\beta_7} = 0.13\%$) devrait plus réduire le $\mathbb{MSEP}$ que l'estimation du paramètre β_2 dont l'indice est pourtant 10 (resp. 30) fois plus grand que celui de β_6 (resp. β_7). En fait, la pondération (λ_{max}) de β_2 est 8 (resp. 2) fois plus importante que la pondération associée à β_6 (resp. β_7).

Les valeurs nulles de tous les λ_{min} dans la Table 3.1 signifient qu'au moins une fois parmi toutes les simulations, chacun des 8 paramètres du modèle fut sélectionné et estimé. Les simulations dans lesquelles chacun des paramètres est estimé, la sélection et l'estimation des paramètres importants ne réduisaient pas d'avantage le $\mathbb{MSEP}$ par rapport à l'estimation des 8 paramètres (pourcentage de la méthode MCO dans la Table 3.2) . Ce résultat est probablement dû soit à la confusion des effets auquel cas la sélection des paramètres par les indices n'a aucun effet sur le $\mathbb{MSEP}$ soit à l'explosion des valeurs des pondérations λ.

La moyenne des estimations des paramètres par la stratégie AS+MCO est très proche des vraies valeurs des paramètres (Table 3.2). Les stratégies LASSO, AS+MCO, AS+LASSO sont au même niveau de performances en termes de $\mathbb{MSEP}$. Chacune de ces trois stratégies conduit au $\mathbb{MSEP}$ le plus faible dans environ 30% de toutes les simulations réalisées. La performance de la méthode de LASSO sur le modèle considéré n'est pas étonnante du

fait que la méthode LASSO s'adapte bien aux modèles sparses. Ce qui fut le cas de notre modèle. La performance égale entre les méthodes LASSO et AS+MCO souligne l'intérêt de l'approche AS+MCO dans ce cas de figure.

Par contre la stratégie AS+MCO est 2 fois meilleure que l'estimation brute MCO et le $\mathbb{MSEP}$ de la méthode MCO est 3 fois plus grand que celui de la stratégie AS+MCO. La stratége AS+MCO assure un gain important en terme du $\mathbb{MSEP}$.

3.6 Discussion

Une pratique courante consiste à sélectionner les paramètres clés à estimer en se basant sur les indices de sensibilité dans le cadre des fonctions de réponse sur-paramétrés par rapport aux observations disponibles. En s'appuyant sur cette pratique, nous avons formalisé les différents concepts utilisés par les modélisateurs. A l'aide de la décomposition du $\mathbb{MSE}$ et du $\mathbb{MSEP}$, nous avons établi une relation formelle entre les qualités du modèle et les indices de sensibilité dans le cas particulier d'un modèle linéaire. Toutes choses égales par ailleurs, estimer les paramètres les plus influents contribue à la réduction du $\mathbb{MSE}$ et du $\mathbb{MSEP}$.

Cependant, cette relation est beaucoup plus complexe que cela même pour le modèle linéaire considéré. La sélection des principaux paramètres à estimer par le biais des indices de sensibilité et la fixation du reste de paramètres ne réduisent systématiquement pas le $\mathbb{MSE}$ et le $\mathbb{MSEP}$. Le $\mathbb{MSEP}$ par exemple, est une somme de trois termes d'erreur dont l'une est fonction des indices de sensibilités. Si les paramètres à faibles effets sont fixés à des valeurs totalement écartées de leurs vraies valeurs alors il est possible de ne pas réduire le $\mathbb{MSEP}$. De plus la relation entre le $\mathbb{MSEP}$ et les indices dépend directement des variables explicatives à travers les pondérations des indices. Ces pondérations qui dépendent des données disponibles peuvent compromettre la réduction du $\mathbb{MSEP}$ en sélectionnant les paramètres clés à estimer à l'aide des indices sauf dans le cas particulier où les variables explicatives sont orthogonales.

La comparaison de la stratégie AS+MCO et de la méthode LASSO montre une performance équivalente entre les deux méthodes de sélection de paramètres en terme de qualité prédictive du modèle sur un modèle très bien adapté à la méthode LASSO. Cette égalité de performance n'est évidement possible que si nous disposons à priori des connaissances pertinentes sur le degré d'incertitudes sur les différents paramètres pour conduire l'analyse de sensibilité. Contrairement à la procédure de sélection LASSO, la méthode AS+MCO

peut s'appliquer aussi bien sur des modèles sparses que sur des modèles non sparses qui sont largement rencontrés en modélisation agronomique et environnement.

Le gain en $\mathbb{MSEP}$ des modèles sur-paramétrés est discutable lorsque que nous sélectionnons les paramètres à l'aide des indices. Il serait intéressant de définir de nouveaux indices qui prennent en compte les pondérations qui figurent dans la relation entre le $\mathbb{MSEP}$ et les indices. Ainsi, l'estimation des paramètres jugés influents grâce aux nouveaux indices va contribuer à améliorer la qualité prédictive du modèle.

Les résultats obtenus sur le modèle linéaire peuvent être étendus aux modèles complexes utilisés par les modélisateurs soit en les linéarisant grâce au développement limité de Taylor soit en construisant des meta-modèles linéaires. La complexité de la relation établie dans ce chapitre sur un modèle aussi simple ne peut qu'attirer notre attention sur la qualité du modèle quand on n'estime que les paramètres clés.

Chapitre 4

Analyse de sensibilité multivariée pour les modèles dynamiques non-linéaires

Introduction

Afin de pouvoir étudier les relations qui existent entre les indices de sensibilité et la qualité du modèle dans le cas des modèles dynamiques, complexe, non-linéaires, il est nécessaire d'avoir une méthode d'analyse de sensibilité multivariée qui permet de hiérarchiser les facteurs d'entrée du modèle. Dans ce chapitre, nous proposons une méthode d'analyse de sensibilité à valeur générique basée sur la décomposition de l'inertie qui prend en compte les différentes corrélations existant entre les sorties des modèles dynamiques. Nous distinguons trois approches dans le développement de la méthode i) les facteurs et les sorties du modèle sont supposés être discrets ii) les sorties restent discrètes mais les facteurs deviennent continus pour mieux balayer la gamme d'incertitude des facteurs iii) les sorties et les facteurs sont tous continus du fait que les phénomènes dynamiques sont en réalité des phénomènes continus. Ensuite, nous comparons les deux premières approches sur l'indice de nutrition azotée (INN) du modèle AZODYN pour évaluer les limites de la première approche du fait qu'elle ne prend en compte toute la gamme d'incertitude. L'ensemble de ces travaux est soumis pour publication dans le journal Reliability Engineering & System Safety (RESS).

Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models

Matieyendou Lamboni[1], Hervé Monod[1,*], David Makowski[2]

[1]INRA, Unité MIA (UR341), Domaine de Vilvert, F78352 Jouy-en-Josas Cedex, France
[2]INRA, UMR 211 INRA AgroParisTech, BP 01, F78850, Thiverval-Grignon, France

* Corresponding author : Hervé MONOD.

Soumis pour publication dans Reliability Engineering & System safety le 10 octobre 2009

Abstract

Many dynamic models are used for risk assessment and decision support in ecology and crop science. Global sensitivity analysis of such models is usually applied separately on each time output, but Campbell, McKay and Williams (2006) advocated global sensitivity analyses on the expansion of the dynamics in a well-chosen functional basis. This paper focuses on the particular case when principal components analysis is combined with analysis of variance. In addition to the indices associated with the principal components, generalised sensitivity indices are proposed to synthesize the influence of each parameter on the whole time series output. Index definitions are given when the uncertainty on the input factors is either discrete or continuous and when the dynamic model is either discrete or functional. A general estimation algorithm is proposed, based on classical methods of global sensitivity analysis. The method is applied to a dynamic wheat crop model with thirteen uncertain parameters. Three methods of global sensitivity analysis are compared when the number of model evaluations must be relatively small : the so-called Sobol-Saltelli method, the extended FAST method, and the fractional factorial design of resolution 6.

Keywords : Dynamic model; Factorial design; Latin hypercube sampling; Principal components analysis; RKHS; Sensitivity analysis; Sobol decomposition.

4.1 Introduction

Global sensitivity analysis is frequently applied to models with multivariate or functional output. In such situations, as mentioned by Campbell *et al.* (2006) [31], it may be unsufficiently informative to perform sensitivity analyses on each output separately or on a few context-specific scalar functions of the output. Indeed, it may be more interesting to apply sensitivity analysis to the multivariate output as a whole. Consequently, there is a need to define criteria and to develop methods specifically adapted to the sensitivity analysis of multivariate or functional outputs.

In particular, consider a model with dynamic output $y(1), \ldots, y(T)$. Conducting separate sensitivity analyses on $y(1), \ldots, y(T)$ gives information on how the sensitivity of $y(t)$ evolves over time. This is interesting, but it leads to much redundancy because of the strong relationship between responses from one time step to the next one. It may also miss important features of the $y(t)$ dynamics because many features cannot be efficiently detected through single-time measurements.

To improve relevance, sensitivity analysis can be applied to pre-defined scalar functions $h(y(1), \ldots, y(T))$ that have a useful interpretation. However, many functions of $y(1), \ldots, y(T)$ are potentially interesting to look at. A general and more sophisticated approach consists in modelling the output as a joint function of time and of the input variables and uncertain parameters. Several examples are illustrated in Chapter 7 of Fang *et al.* (2006) [50], based on spatio-temporal, functional or semiparametric modelling tools.

However there is also a need to apply data-driven methods that can identify the most interesting features in the $y(t)$ dynamics and perform sensitivity analyses on these features. Campbell *et al.* (2006) [31] proposed a simple and very useful approach to do so. It consists in *(i)* performing an orthogonal decomposition of the multivariate output, and *(ii)* applying sensitivity analysis to the most informative components individually. There is a large collection of available methods for the first step : it can be based either on a data driven method such as principal component analysis, or on the projections of output on a polynomial, spline, or Fourier basis defined by the user. The second step can also be performed by several different methods of sensitivity analysis, such as factorial design, FAST, or Sobol and its most recent versions developed by Saltelli et al. (see Saltelli *et al.*, 2008 [127]).

The method proposed by Campbell *et al.* (2006) [31] allows to restrict attention to a

few components rather than a whole dynamic. However, there is a need also for a synthetic criterion to summarise the sensitivity over the whole dynamic. In addition, this criterion must be adapted to discrete or continuous uncertainty distributions, whereas the examples in [31] are restricted to the first case. In this paper, we first show that there is a full "factorial by component" decomposition of the output variability or inertia, as illustrated in Lurette *et al.* (2009)[95] and Lamboni *et al.* (2008) [91]. Based on this decomposition, we propose a new synthetic sensitivity criterion for discrete factors first. We then extend this criterion to the cases when the input factors and output are continuous, and estimation methods are proposed and compared through simulations on a crop model.

Section 2 presents the methods. It starts with the case when the input and output are discrete and a complete or fractional factorial design is used (Section 4.2.2). It follows with the case when the input and/or the output are continuous (Sections 4.2.3 and 4.2.4). In that case, a procedure is proposed to sample the input space and estimate the sensitivity indices. In Section 4.3, the methods are illustrated on a crop model with 13 parameters. The main results are discussed in Section 4.4.

4.2 Methodology

4.2.1 Framework

We consider the sensitivity analysis (SA) of a dynamic model

$$y(t) = f_0(\mathbf{x}, t; \theta), \qquad t \in \mathcal{I} \tag{4.2.1}$$

where $y(t)$ is the scalar output at time t ; $\mathbf{x}$ is a vector of input variables and θ is a vector of parameter values. The time domain $\mathcal{I}$ may be discrete or continuous, and both cases will be considered. For simplicity, Model (4.2.1) is assumed to be deterministic.

To perform the sensitivity analysis, some uncertain parameters and input variables are selected for study, while the others are fixed at given nominal values. The selected parameters and input variables yield the d input factors F_1 to F_d of the sensitivity analysis. Let $\mathbf{z} = (z_1, z_2, \ldots, z_d)'$ denote a scenario, that is, a combination of the levels z_j of the input factors F_j, for j in 1,..., d. The model of interest in this paper is

$$y(t) = f(\mathbf{z}, t), \quad t \in \mathcal{I}, \tag{4.2.2}$$

where $f(\mathbf{z}, t) = f_0(\mathbf{x}, t; \theta)$ for some explicit mapping of $\mathbf{z}$ on $(\mathbf{x}, \theta)$. We further consider Model (4.2.2) only. A single model run is determined by its scenario $\mathbf{z}$. Its output is a vector $\mathbf{y} = (y(1), \ldots, y(T))'$ if $\mathcal{I}$ is discrete or more generally a function $\mathbf{y}(t)$, $t \in \mathcal{I}$.

4.2.2 Discrete-time model with discrete input factors

In this subsection, Model (4.2.2) is a discrete-time dynamic model, with $\mathcal{I} = \{1, 2, \ldots, T\}$. Besides, the uncertainty domain is restricted to a discrete set D_j of n_j values of interest for each input factor F_j. Thus the uncertainty domain of $\mathbf{z}$ is the complete factorial design, that is, the full set of scenarios $\Omega = D_1 \times \ldots \times D_d$ of size $N = \prod_j n_j$.

The full set of output dynamics over the complete factorial design Ω, forms the $N \times T$ matrix :

$$\mathcal{Y} = \begin{pmatrix} y_1(1) & \ldots & y_1(t) & \ldots & y_1(T) \\ \vdots & & \vdots & & \vdots \\ y_i(1) & \ldots & y_i(t) & \ldots & y_i(T) \\ \vdots & & \vdots & & \vdots \\ y_N(1) & \ldots & y_N(t) & \ldots & y_N(T) \end{pmatrix}.$$

Each column $\mathbf{y}(t)$ in $\mathcal{Y}$ represents the values of the output variable at a given time t, for the full set of scenarios, while each row $\mathbf{y}_i$ of $\mathcal{Y}$ is an individual dynamic for a given scenario $\mathbf{z}$.

Anova-based decomposition of variance

Consider first global sensitivity analysis for a univariate output. When the input factors are discrete, this is equivalent to analysis of variance (anova), a classical method in statistics (Campolongo and Saltelli, 2000 [31]).

In the full anova decomposition, the output variance is decomposed across 2^d factorial terms. Each factorial term w is associated with a subset of factors, and thus it can be identified to the corresponding subset of $\{1, \ldots, d\}$. For example, the factorial effect denoted by $w = \emptyset$ corresponds to the general mean of the output. The subset $w = \{j\}$ denotes the main effect, or first order effect, of factor F_j. The subset $w = \{j_1, j_2\}$ denotes the interaction between factors F_{j_1} and F_{j_2}, a second order effect.

Let $\mathbf{h}$ in $\mathbb{R}^N$ denote a univariate output vector across all N scenarios. Because of the orthogonality properties of the complete factorial design, there is a unique decomposition of the Sum of Squares $\mathrm{SS}(\mathbf{h}) = \|\mathbf{h} - \bar{\mathbf{h}}\|^2$:

$$\mathrm{SS}(\mathbf{h}) = \sum_{w, w \neq \emptyset} \mathrm{SS}_w, \tag{4.2.3}$$

where SS_w denotes the anova sum of squares associated with the factorial term w. From a technical point of view, the anova decomposition is associated with a decomposition of $\mathbb{R}^N$ into a direct sum of mutually orthogonal subspaces

$$\mathbb{R}^N = \bigoplus_w^{\perp} V_w. \tag{4.2.4}$$

and the sum of squares associated with the factorial term w is defined by

$$\mathrm{SS}_w = \|\mathbf{S}_w \mathbf{h}\|^2,$$

where $\mathbf{S}_w$ denotes the orthogonal projection matrix on V_w. The subspace V_w is defined by the recurrence relation

$$V_w = W_w \cap \left(\bigoplus_{u \subset w}^{\perp} V_u \right)^{\perp},$$

where W_w is the subspace spanned by the indicator vectors of the combinations of levels of the factors in the subset w and $W_\emptyset$ is spanned by the all-one vector.

Principal Components Analysis

Consider now the dynamic output $\mathbf{y}$. Principal Components Analysis (PCA) allows its expansion in a new basis, so that most information is concentrated in the first few components (Jolliffe, 2002 [82] ; Anderson, 2003 [11] ; Saporta, 2006 [131], Besse, 1992 [21]).

Let Σ denotes either the variance-covariance matrix or the correlation matrix of the columns of $\mathcal{Y}$. Thus

$$\Sigma = \frac{1}{N} \mathcal{Y}_c{}' \mathcal{Y}_c,$$

with $\mathcal{Y}_c$ the matrix obtained by centering and possibly normalising each column of $\mathcal{Y}$. The PCA decomposition is based on the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_T$ and on the normalised eigenvectors $\mathbf{v}_k$ of Σ. For simplicity, we assume that $N \geq T$ and that, in case an eigenvalue multiplicity is larger than 1, the corresponding eigenvectors $\mathbf{v}_k$ are chosen to be mutually orthogonal. Then the principal components (PCs) $\mathbf{h}_k$, for $k = 1, 2, \ldots, T$, are the mutually orthogonal linear combinations of the $\mathcal{Y}_c$ columns defined by $\mathbf{h}_k = \mathcal{Y}_c \mathbf{v}_k$ or, in matrix form, the columns of the $N \times T$ matrix $\mathcal{H} = \mathcal{Y}_c \mathcal{V}$, where $\mathcal{V}$ denotes the $T \times T$ matrix with $\mathbf{v}_k$ in column k. The PCs appear in the expansion of the output vectors in the basis defined by the eigenvectors $\mathbf{v}_k$, which reads

$$\mathbf{y}_i(t) = \bar{y}(t) + \sum_{k=1}^{T} (\mathbf{h}_k)_i \mathbf{v}_k(t).$$

The inertia of $\mathcal{Y}$ is defined by $\mathbb{I} = \mathbb{T}\mathrm{r}(\Sigma)$. If Σ is a correlation matrix, then $\mathbb{I} = T$. Otherwise, the inertia measures the total dispersion or variability among the rows of $\mathcal{Y}$. The principal components satisfy $\|\mathbf{h}_k\|^2 = \lambda_k$, and the eigenvalues satisfy $\sum_k \lambda_k = \mathbb{T}\mathrm{r}(\Sigma) = \mathbb{I}$. Thus, by construction, the principal component matrix $\mathcal{H}$ has the same total inertia as $\mathcal{Y}_c$, but it is mostly concentrated in its first columns.

Sensitivity indices on the principal components

Sensitivity analysis (SA) can be applied to each principal component. By combining the anova and PCA decompositions, the following definitions thus generalise the univariate global sensitivity indices defined, e.g., in Saltelli *et al.* (2000) [126].

Définition 4.2.1 *For Model (4.2.2) with discrete input factors and discrete-time output :*

- *the sensitivity index of factorial term w ($w \neq \emptyset$) for the kth principal component is defined by*
$$\mathbb{SI}_{w,k} = \frac{\mathrm{SS}_{w,k}}{\lambda_k}, \text{ where } \mathrm{SS}_{w,k} = \|\mathbf{S}_w \mathbf{h}_k\|^2;$$
- *the first order sensitivity index of F_j for the kth principal component corresponds to the main effect of F_j ($w = \{j\}$), and so is defined by $\mathrm{FSI}_{F_j,k} = \mathbb{SI}_{\{j\},k}$;*
- *the total sensitivity index of F_j for the kth principal component is defined by*
$$\mathrm{TSI}_{F_j,k} = \sum_{w, j \in w} \mathbb{SI}_{w,k},$$
where the sum includes all the factorial terms w that include factor F_j.

Clearly, the sensitivity indices satisfy $0 \leq \mathbb{SI}_{w,k} \leq 1$ and $\sum_w \mathbb{SI}_{w,k} = 1$ for all factorial terms w and all principal components k. In practice, it is possible to limit the decompositions in equation (4.2.3) and Definition 4.2.2 to a subset of factorial terms, typically the main effects and low order interactions. A final residual term w_{Residual} should then be included to account for the remaining variability.

Generalised sensitivity indices

In addition to the sensitivity for each principal component, it is interesting to quantify the contribution of each factorial term w to the total inertia. This can be done through the same decomposition as performed in multivariate analysis of variance (manova), a generalisation of anova to multivariate responses (Anderson, 2003 [11]).

Définition 4.2.2 *For Model (4.2.2) with discrete input factors and discrete-time output,*
- *the* Generalised Sensitivity Index *of factorial term w ($w \neq \emptyset$) is defined by*

$$\mathbb{GSI}_w = \frac{\mathbb{T}r(\mathcal{Y}_c{}'\mathbf{S}_w\mathcal{Y}_c)}{\mathbb{I}};$$

- *the generalised first order sensitivity index of F_j is defined by* $\mathrm{GFSI}_{F_j} = \mathbb{GSI}_{\{j\}}$ *;*
- *the generalised total sensitivity index of F_j is defined by*

$$\mathrm{GTSI}_{F_j} = \sum_{w, j\in w} \mathbb{GSI}_w.$$

Proposition 4.2.1 *For all factorial terms w, the generalised sensitivity indices satisfy :*
- $\mathbb{GSI}_w = \sum_k \frac{\lambda_k}{\mathbb{I}}\, \mathbb{SI}_{w,k}$ *;*
- $0 \leq \mathbb{GSI}_w \leq 1$ *;*
- $\sum_w \mathbb{GSI}_w = 1$.

Preuve 4.2.1 *Since $\sum_{k=1}^{T} \mathbf{v_k}\mathbf{v_k}'$ is the $T \times T$ identity matrix and $\sum_w \mathbf{S}_w$ is the identity matrix of order N according to (4.2.4), it follows that*

$$\begin{aligned}
\mathbb{I} &= \mathbb{T}r(\mathcal{Y}_c{}'\mathcal{Y}_c) \\
&= \sum_w \mathbb{T}r(\mathcal{Y}_c{}'\mathbf{S}_w\mathcal{Y}_c) \\
&= \sum_w \sum_k \mathbb{T}r(\mathcal{Y}_c{}'\mathbf{S}_w\mathcal{Y}_c\mathbf{v_k}\mathbf{v_k}') \\
&= \sum_w \sum_k \|\mathbf{S}_w\mathbf{h}_k\|^2
\end{aligned}$$

□

The following proposition provides a particularly simple way to calculate the generalised sensitivity indices.

Proposition 4.2.2 *The Generalised Sensitivity Index $\mathbb{GSI}_w$ is equal to the sensitivity index $\mathbb{SI}_w = \|\mathbf{S}_w\mathbf{h}\|^2/\|\mathbf{h}\|^2$, where $\mathbf{h} = \sum_{k=1}^{T} \mathbf{h}_k$.*

Preuve 4.2.2 *See 4.4.1.*

Computation issue

In the present subsection, the domain of input uncertainty is a complete factorial design, consisting of the $N = \prod_j n_j$ possible scenarios $\mathbf{z}$ in Ω. If it numerically possible to calculate the output for these N scenarios, then this yields the matrix $\mathcal{Y}$ and all sensitivity indices defined above can be calculated exactly. On the other hand, if the N simulations are too costly, an alternative is to rely on fractional factorial designs, which can reduce the number of model evaluations dramatically. We refer, e.g., to Box and Draper (1987) [25], Kobilinsky (1997) [86] or Campolongo and Cariboni (2000)[129] for more details, and to Lurette *et al.* (2009) [95] for an application.

4.2.3 Discrete-time model with continous random factors

For many dynamic models, uncertainty in the input factors F_j is better described by independent probability distributions defined on continuous intervals, rather than by finite level sets. Up to normalisation, the uncertainty domain can then be identified to the unit hypercube $\Omega = [0,1]^d$, and a scenario $\mathbf{z}$ can be considered as a realisation of the random vector $\mathbf{Z}$, with $\mathbf{Z}$ following the joint probability distribution of the input factors. Through equation (4.2.2), the random scenario $\mathbf{Z}$ generates a random output vector $\mathbf{Y}$ and the probability distribution of $\mathbf{Z}$ induces a probability distribution on $\mathbf{Y}$. We assume that $\mathbf{Y}$ is square integrable, that is, $\mathbb{E}(Y^2(t)) < +\infty$, $\forall\ t \in \mathcal{I}$. In that framework, the global sensitivity indices are defined through conditional probabilities. For example, the first order sensitivity index of factor F_j on $Y(t)$ is equal to $\mathbb{V}\mathrm{ar}(\mathbb{E}(Y(t)|Z_j))/\mathbb{V}\mathrm{ar}(Y(t))$.

In this subsection, we still assume that $\mathcal{I} = \{1, 2, \ldots, T\}$.

Principal Components decomposition

Let $\mu(t) = \mathbb{E}[Y(t)]$ denote the mean of $Y(t)$ over the uncertainty domain Ω, and let Σ denote the $T \times T$ variance-covariance matrix of $\mathbf{Y}$ over time. As Σ is symmetric and positive definite, Σ and $\mathbf{Y}$ can be expanded in the same way as in Section 4.2.2 :

$$\Sigma = \sum_{k=1}^{T} \lambda_k \mathbf{v}_k \mathbf{v}_k',$$

$$Y(t) = \mu(t) + \sum_{k=1}^{T} H_k \mathbf{v}_k(t), \text{ or, in vectorial form, } \mathbf{Y} = \mu + \mathcal{V}\mathbf{H}, \tag{4.2.5}$$

where $\mathbf{v}_k$, for $k \in \mathcal{I}$, are the normalised orthogonal eigenvectors of Σ associated with the eigenvalues λ_k, with $\lambda_1 \geq \ldots \geq \lambda_T \geq 0$; $\mathcal{V}$ is the eigenvector matrix and $\mathbf{H}$ is a random

vector whose elements are given by

$$\begin{aligned} H_k &= [\mathbf{Y} - \mu]' \mathbf{v}_k \\ &= \sum_{t=1}^{T} [f(\mathbf{Z}, t) - \mu(t)] \mathbf{v}_k(t). \end{aligned}$$

The random variate H_k is the principal component $\mathbf{h}_k$ of Section 4.2.2, adapted to the probabilistic framework. Conditional to $\mathbf{Z} = (Z_1, \ldots, Z_d)$, it is fixed. Marginally, it is centered with variance λ_k, that is $\mathbb{V}\mathrm{ar}(H_k) = \mathbb{E}(H_k^2) = \lambda_k$.

The finite expansion of the model output in (4.2.5) shows that all the variability of the model output is contained in the variances of H_k, $k \in \mathcal{I}$ with decreasing importance (variance) when k increases. These considerations motivate the use of the random principal components H_k to analyze the model output uncertainty.

Sensitivity indices

The variance based sensitivity indices have been defined by Sobol (1993) [142], and Saltelli *et al.* (2000 [126], 2002 [124], 2008 [127]) in the univariate context with continuous input factors. An adaptation to the multivariate case is proposed in the generalisation of Definition 4.2.2 given below. For this definition and below, $Z_{[w]}$ denotes the subset $\{F_j, j \in w\}$ of the factors that belong to the factorial effect w.

Définition 4.2.3 *Let $\mathbb{V}_{w,k}$, for $w \subset \{1, \ldots, d\}$, be defined recursively by $\mathbb{V}_{\emptyset,k} = 0$ and $\mathbb{V}_{w,k} = \mathbb{V}ar[\mathbb{E}(H_k \mid Z_{[w]}] - \sum_{u,u\subset w} \mathbb{V}_{u,k}$. For Model (4.2.2) with continuous input factors and discrete-time output,*

- *the sensitivity index of factorial term w ($w \neq \emptyset$) for the kth principal component is defined by*
$$\mathbb{SI}_{w,k} = \frac{\mathbb{V}_{w,k}}{\lambda_k},$$
- *the first order sensitivity index of F_j for the kth principal component is defined by* $\mathrm{FSI}_{F_j,k} = \mathbb{SI}_{\{j\},k}$ *;*
- *the total sensitivity index of F_j for principal component $\mathbf{h}_k$ is defined by* $\mathrm{TSI}_{F_j,k} = \sum_{w,j\in w} \mathbb{SI}_{w,k}$.

In addition to the component-based sensitivity indices, it is useful again to define a synthetic sensitivity index to measure the contribution of each factor to the whole variability of the model output.

Définition 4.2.4 *For Model (4.2.2) with continuous input factors and discrete-time output,*

- *the* Generalised Sensitivity Index *of factorial term w is defined by*

$$\mathbb{GSI}_w = \frac{\mathbb{GV}_w}{\mathbb{I}} \text{ with } \mathbb{GV}_w = \sum_k \mathbb{V}_{w,k};$$

- *the first order generalised sensitivity index of F_j is defined by* $\mathrm{GFSI}_{F_j} = \mathbb{GSI}_{\{j\}}$ *;*
- *the total generalised sensitivity index of F_j is defined by* $\mathrm{GTSI}_{F_j} = \sum_{w,j\in w} \mathbb{GSI}_w$.

Let $H = \sum_{k=1}^{T} H_k$. Thus H denotes the sum of the principal components and $\mathbb{V}\mathrm{ar}(H) = \mathbb{I}$. The proposition below shows that the generalised sensitivity indices are equivalent to the standard sensitivity indices applied to H.

Proposition 4.2.3 *Let $\mathbb{V}_{w,\mathrm{G}}$, for $w \subset \{1,\ldots,d\}$, be defined recursively by $\mathbb{V}_{\emptyset,\mathrm{G}} = 0$ and $\mathbb{V}_{w,\mathrm{G}} = \mathbb{V}ar[\mathbb{E}(H \mid Z_w)] - \sum_{u,u\subset w} \mathbb{V}_{u,\mathrm{G}}$. The Generalised Sensitivity Index $\mathbb{GSI}_w$ is equal to the sensitivity index $\mathbb{SI}_w$ defined on H, so*

$$\mathbb{GSI}_w = \frac{\mathbb{V}_{w,\mathrm{G}}}{\mathbb{V}ar(H)}.$$

Preuve 4.2.3 *See 4.4.2.*

Estimation

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_n$ denote the outputs of a sample of scenarios $\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_n$. The eigenvalues and eigenvectors defined above can be estimated by using the standard estimator of the covariance matrix Σ, given by :

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{y}_i - \bar{\mathbf{y}}][\mathbf{y}_i - \bar{\mathbf{y}}]',$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$. The estimators $\hat{\lambda}_k$ and $\hat{\mathbf{v}}_k$ of λ_k and $\mathbf{v}_k$ are respectively the kth eigenvalue and the kth eigenvector of the empirical covariance matrix $\hat{\Sigma}$. Based on these estimators, predictions of the principal component scores $H_k, k \in \{1, 2, \ldots,\}$ defined in (4.2.5) can be calculated for each sample, by

$$\hat{H}_{i,k} = \sum_{t=1}^{T}[\mathbf{y}_i - \bar{\mathbf{y}}]\hat{\mathbf{v}}_k(t).$$

Based on these considerations, we propose the quite natural algorithm below to estimate the PCs and the sensitivity indices :

Algorithme 4.2.1 *Computation of the Principal Components (PC) and Sensitivity Indices (SI) when factors are continuous and output is discrete*

Step 1 :	*choose one among the different variance based methods allowing to estimate global sensitivity indices (Sobol or FAST for example)*
Step 2 :	*sample the input space* Ω *according to the method chosen at Step 1, to get the factor values in* n *scenarios* $\mathbf{z}_i$
Step 3 :	*calculate the* n *model outputs* $\mathbf{y}_i$
Step 4 :	*perform Principal Components Analysis on these outputs, to get the estimates* $\hat{\lambda}_k$, $\hat{\mathbf{v}}_k$, $\hat{H}_{i,k}$, *for* $k = 1, \ldots, K$, *where* $K \leq T$ *is chosen according to the percentage of inertia that the user wants to keep*
Step 5 :	*compute the SIs on the PCs* $\hat{H}_{i,k}$ *estimated at step 4, for* $k = 1, \ldots, K$, *by the method chosen at step 1*
Step 6 :	*compute the generalised sensitivity indices by applying the method chosen at step 1 to the sum of the principal components, as follows from Proposition 4.2.3.*

If the scenarios $\mathbf{z}_i$ are sampled independently according to the probability distribution of $\mathbf{Z}$, the eigenvectors ($\hat{\mathbf{v}}_k$) and eigenvalues ($\hat{\lambda}_k$) satisfy convergence in probability towards their true values (λ_k, $\mathbf{v}_k$), together with more detailed properties that can be found in Hall *et al.* (2006a, 2006b) [64], [63], Bosq(2000) [24], Dauxois *et al.* (1982) [44] and Anderson (1963) [9]. It follows that $\hat{H}_k$ also converges in probability towards H_k. In addition, the precision on the estimated quantities associated with the principal components analysis can be assessed by bootstrap, as proposed by Hall *et al.* (2006b), and this can be added to the Step 4 of Algorithm 4.2.1.

These properties apply directly if the scenarios are generated by Monte Carlo sampling. However, this is not the case with the standard sensitivity analysis methods, and so there is a need to better assess the convergence properties of the principal components estimates and of the sensitivity indices, when Algorithm 4.2.1 is applied with different SA methods. In this paper, this issue is addressed through the simulations on a case study presented in Section 4.3.

4.2.4 Functional output

Many phenomena are continuous by nature and need to be modelled as a function or a curve instead of discrete points. For example, in crop science, phenomena such as crop biomass or gaseous emission are frequently observed once a day or at another time interval limited by technical constraints, but should in fact be considered as continuous over time. Hall *et al.* (2006b [63]) give conditions under which the recorded data must be treated as functional data or not ; they suggest that data recorded with a high frequency should often be treated as functional data.

In this subsection we treat any dynamic phenomenon as a continuous function over the time interval $\mathcal{I} = [0, T]$. The input factors are also supposed to be continuous. As in Section 4.2.3, their levels are supposed to vary in the unit hypercube $\Omega = [0, 1]^d$.

Consider the dynamic model defined by

$$Y(t) = f(\mathbf{Z}, t), \quad t \in \mathcal{I} = [0, T], \tag{4.2.6}$$

where $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_d)$ is treated as a random variable, as in Section 4.2.3. Hence, the model output is a random or stochastic function and we assume that it is square integrable, that is, $\int_0^T \mathbb{E}(\mathbf{Y}^2(t))dt < \infty$.

Functional PCA

As a random function on $\mathcal{I}$, the model output $\mathbf{Y}$ is intrinsically infinite dimensional. Sensitivity analyses can be applied to only a subset of output variables or to predefined functions $h(\mathbf{Y})$ with a reasonable interpretation. For example, h may be the mean $h(\mathbf{Y}) = \int_T \mathbf{Y}(t)dt$ when one aims to have a synthetic view on the global model output. Alternatively, functional PCA allows finite dimension analysis of the infinite dimensional problem by concentrating the information contained in the data in a few uncorrelated components (Ramsay *et al.*, 1997 [118] and 2002 [119]).

Consider the random function $\mathbf{Y}$ defined in (5.2.3) and let $\mu(t) = \mathbb{E}(Y(t))$ and $K(t_1, t_2) = \mathbb{C}\text{ov}[Y(t_1), Y(t_2)]$ denote respectively its mean and covariance operators. It is common to use the covariance operator K for characterising a random function. Since the covariance function is positive definite on $\mathcal{I}\times\mathcal{I}$, the Moore-Aronszajn theorem (Aronszajn, 1950 [14] ; Berlinet *et al.*,2004 [20]) shows that there exists a unique Reproducing Kernel Hilbert Space (RKHS) $\mathbb{F}$ associated to the kernel $K(.,.)$. The kernel $K(.,.)$ is assumed to be square integrable, and the model output is now considered as a random

variable on the Hilbert space $\mathbb{F}$.

Mercer's theorem (Indritz, 1963 [76] ; Berlinet *et al.*, 2004 [20] : Hall *et al.*, 2006a [64], 2006b [63]) implies a spectral decomposition of the covariance function

$$K(t_1, t_2) = \sum_{k=1}^{+\infty} \theta_k \psi_k(t_1) \psi_k(t_2), \tag{4.2.7}$$

where $\theta_1 \geq \theta_2 \geq \ldots \geq 0$ are ordered eigenvalues of K and the corresponding normalised orthogonal eigenfunctions are ψ_k, for $k = 1, \ldots, \infty$. The Karhunen-Loeve expansion or functional principal component expansion of $\mathbf{Y}$ is given by

$$Y(t) = \mu(t) + \sum_{k=1}^{+\infty} \xi_k \psi_k(t), \tag{4.2.8}$$

where

$$\xi_k = \int_{\mathcal{I}} [Y(t) - \mu(t)] \psi_k(t) dt \tag{4.2.9}$$

is the kth functional principal component score. By the Karhumen-Loeve expansion (4.2.8), the random character of the output $\mathbf{Y}$ is transfered into the functional principal component scores $\xi_k, k \geq 1$. In particular, the random effects $\xi_k, k \geq 1$ are centered and uncorrelated with variance $\theta_k = \mathbb{E}(\xi_k^2)$, and

$$\mathbb{V}\text{ar}\{Y(t)\} = \sum_{k=1}^{+\infty} \theta_k \psi_k^2(t). \tag{4.2.10}$$

The functional inertia is naturally defined by the series

$$\mathbb{I} = \sum_{k=1}^{+\infty} \theta_k = \int_{\mathcal{I}} \mid K(t,t) \mid^2 dt. \tag{4.2.11}$$

The strength with which the curve ψ_k contributes to the random function is proportional to the standard deviation θ_k of ξ_k. As in classical principal component analysis, the high proportion of inertia explained by the first functional components motivates their use as a support for the sensitivity analysis of functional data.

All the series defined in (4.2.7), in (4.2.8) and (4.2.11) are absolutely and uniformly almost sure convergent (Hall, 2006b [63]).

Sensitivity analysis

The variance based sensitivity analysis or any global sensitivity analysis (Saltelli *et al.*, 2008 [127] ; Saltelli, 2002 [124] ; Sobol, 1993 [142] ; Saltelli *et al.*, 2000 [126]) allow to identify the important factors on the kth functional principal components.

To define the sensitivity indices precisely, Definitions 4.2.3 and 4.2.4 can be used directly, provided H_k is replaced by ξ_k. Proposition 4.2.3 can also be adapted to the functional principal component decomposition, using the random variable $\xi = \sum_{k=1}^{+\infty} \xi_k$, whose variance is equal to the inertia.

Proposition 4.2.4 *Let $\mathbb{V}_{w,\mathrm{G}}$, for $w \subset \{1, \ldots, d\}$, be defined recursively by $\mathbb{V}_{\emptyset,\mathrm{G}} = 0$ and $\mathbb{V}_{w,\mathrm{G}} = \mathbb{V}ar[\mathbb{E}(\xi \mid Z_w)] - \sum_{u,u\subset w} \mathbb{V}_{u,\mathrm{G}}$. The Generalised Sensitivity Index $\mathbb{GSI}_w$ is equal to the sensitivity index $\mathbb{SI}_w$ defined on ξ, so*

$$\mathbb{GSI}_w = \frac{\mathbb{V}_{w,\mathrm{G}}}{\mathbb{I}}.$$

The proposition shows that the generalised sensitivity indices can be defined directly on ξ, which contains, in some sense, all the variability of the model output. In particular, the generalised main and total indices are given by

$$\begin{aligned} \mathrm{GFSI}_{F_j} &= \frac{\mathbb{V}\mathrm{ar}[\mathbb{E}(\xi \mid \mathbf{z}_j)]}{\mathbb{I}} \\ \mathrm{GTSI}_{F_j} &= \frac{\mathbb{V}\mathrm{ar}(\xi) - \mathbb{V}\mathrm{ar}[\mathbb{E}(\xi \mid -\mathbf{z}_j)]}{\mathbb{I}} \end{aligned}$$

The number of functional principal components is infinite, but in practice, only the first functional principal components carry useful information for statistical analysis. Hence, approximate sensitivity indices based on the first P functional principal components scores can be computed.

Estimation

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_n$ denote a sample of n independent realisations of $\mathbf{Y}$ in the RKHS $\mathbb{F}$. The standard estimator of the kernel $K(t,s)$ is given by :

$$\hat{K}(t,s) = \frac{1}{n-1} \sum_{i=1}^{n} [\mathbf{y}_i(t) - \bar{\mathbf{y}}(t)][\mathbf{y}_i(s) - \bar{\mathbf{y}}(s)],$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$. This empirical kernel is used for eigenvalue and eigenfunction estimations. More precisely, the estimators $\hat{\theta}_k$ and $\hat{\psi}_k$ of θ_k and ψ_k are respectively the kth eigenvalue and the kth eigenfunction of the empirical operator $\hat{K}$. The convergence

in probability of these estimators and other properties are available in Hall *et al.* (2006a, 2006b) [64],[63]. Now, predictions of the functional principal component scores $\xi_k, k \geq 1$ defined in (4.2.9) can be calculated for each sample, by

$$\hat{\xi}_{i,k} = \int_{\mathcal{I}} [\mathbf{y}_i - \bar{\mathbf{y}}]\hat{\psi}_k(t)dt.$$

In practice, the functional output must be discretised according to the domain of application or to technical constraints. Then the algorithm 4.2.1 can be used to compute the sensitivity indices.

4.3 Case study

4.3.1 Description of the model Azodyn

The model AZODYN (Jeuffroy and Recous, 1999 [81]) simulates the wheat crop development, in order to guide farmers in their crop fertilisation strategies. The main goal of the model AZODYN is to develop interesting strategies for fertilisation that meet performance objectives such as grain protein content or preservation of the environment.

AZODYN simulates crop and soil components among which yield, biomass, protein content of grains, residual mineral nitrogen in the soil at harvest and the nitrogen nutrition index (INN). In this paper, attention is focused on the INN output, because plant development and fertilisation management are based on INN. AZODYN simulates INN at a daily time step. It includes input variables such as temperature and soil type, together with a total of 69 parameters, among which 13 are genetic parameters, that is, parameters which depend on the wheat cultivar. Following Makowski *et al.* (2006) [97], sensitivity analysis will be restricted to these 13 genetic parameters. Their uncertainty intervals are listed in Table 4.1 and Figure 1 shows a sample of simulated INN dynamics when drawing the 13 uncertain genetic parameters in their uncertainty intervals and fixing the other parameters to given nominal values.

4.3.2 Simulation experiments

Reference simulation design

The model AZODYN is too complex for one to calculate the principal components and sensitivity indices analytically, but it runs very fast. It was thus possible to define a large simulation design to provide reliable and precise estimates of the sensitivity indices,

Parameter	Interpretation	Nominal value	Uncertainty interval
$E_{b\max}$	radiation use efficiency	3	2.7-3.3
$E_{i\max}$	ratio of intercepted to incident radiation	0.94	0.9-0.99
K	extinction coefficient	0.7	0.6-0.8
Tep.flo	duration between earing and flowering	150	100-200
D	ratio of leaf area index to critical nitrogen	0.035	0.020-0.045
R	ratio of total to above ground nitrogen	1.25	1-1.5
Lambda	parameter for nitrogen use efficiency	35	25-45
Mu	parameter for nitrogen use efficiency	0.75	0.6-0.9
DJPF	temperature threshold	200	150-250
NGM2MAXVAR	maximal grain number	128	107.95-146.05
P1GMAXVAR	maximal weight of one grain	56	47-65
RDTMAXVAR	maximal yield	118	100-137
REM2	fraction of remobilised nitrogen	0.7	0.5-0.9

TABLE 4.1 – Uncertainty intervals for AZODYN model genetic parameters

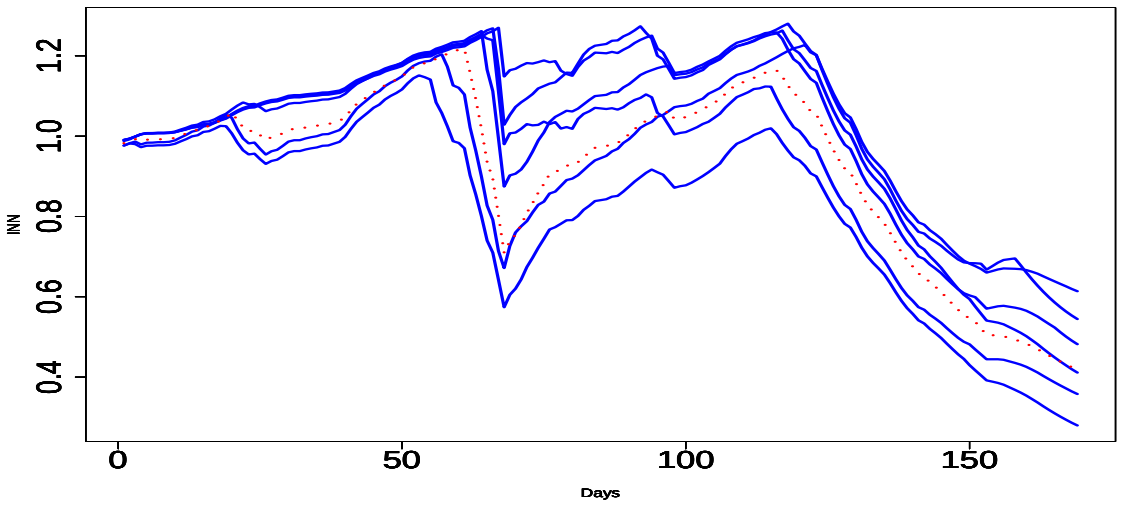

FIGURE 4.1 – Uncertainty on AZODYN-INN due to the variability of the input parameters : illustration by six INN dynamics resulting from six randomly sampled input combinations. The dotted curve corresponds to the nominal values of the input factors.

to be used as reference values.

The method proposed by Sobol (1993)[142] and adapted by Saltelli (2002)[124] was implemented. In the following, it will be called the Sobol-Saltelli method. It requires two initial sets of n_0 Monte Carlo simulations and $n_0(d+2)$ simulations overall, since an additional sample of size n_0 is required for each input factor. To get a good covering of the input space, the size was set at $n_0 = 10000$ and Latin Hypercubes were used rather than Monte Carlo samples, for both initial sets (McKay *et al.*, 1979 [101]). Confidence bands on the principal components were calculated by bootstrap with 100 replications (Hall *et al.*, 2006a)[64]. In addition, the whole process was replicated five times and confidence intervals were calculated by the quantile method (Hyndman and Fan, 1996)[75] on the estimated principal components and sensitivity indices.

Smaller simulation designs to compare methods

The reference design defined above required $150,000$ model evaluations, which would be too expensive for many models. In Saltelli *et al.* (2008) [127] (p.164), the order of magnitude given for n_0 when using the Sobol-Saltelli method is between a few hundreds to a few thousands. With such a number of model evaluations, alternative methods include :

- discretising the continous factors and using fractional factorial designs ;
- applying the Sobol-Saltelli method ;
- using alternative methods for continuous factors such as extended FAST (Saltelli *et al.*, 1999 [130]).

In order to compare these methods on AZODYN when the sample size is relatively small, the sensitivity indices were computed with approximatively 6500 model evaluations for each method.

A fractional factorial design was constructed with all 13 factors at three levels. The levels were the mean and the bounds of the uncertainty intervals. A complete 3^{13} factorial design would have required $1,594,323$ simulations. Instead, a regular fraction of size $3^8 = 6561$ and resolution 6 was used. Resolution 6 means that the main effects and two-factor interactions can be estimated, provided interactions of order larger than 3 are assumed to be zero (see Box and Draper, 1987 [25] ; Kobilinsky, 1997 [86]). Consequently the selected fraction allowed to compute the main and second-order sensitivity indices, with possible bias due to higher order interactions. It also allowed to estimate total indices by summations over the main effect and second-order interactions of each input factor, with the same possible sources of bias.

For the Sobol-Saltelli method, two simulation designs were run with $n = 6,555$ ($= 437 \times 15$) model evaluations, one by using Monte Carlo sampling and the other by using Latin hypercube sampling.

An alternative approach to compute the SI is the Extended FAST method. The FAST method was proposed by Cukier *et al.* (1973)[40] and its extended version (eFAST) was introduced by Saltelli *et al.* (1999)[130], to compute the first order and total sensitivity indices efficiently ($n_0 \times d$). The eFAST method calculates the sensitivity indices of each factor by an approximated one-dimensional integral, following the sequence of scenarios defined by

$$z_{i,j} = G_j(\sin \omega_j \, s_i), \quad s_i = \frac{2(i-1) - (n_0 - 1)}{n_0 - 1} \, \pi,$$

where $z_{i,j}$ is the level of factor F_j in the scenario $\mathbf{z}_i$, for $i = 1, \ldots, n_0$, and ω_j are frequency parameters. The size was set at $n_0 = 504$ for each input factor, resulting in $n = 504 \times 13 = 6552$ simulations.

Confidence bands on the principal components were calculated by the bootstrap method with 100 replications, for the Sobol-Saltelli and eFAST methods.

Software

The fractional design was constructed by the algebra methods implemented in the FACTEX procedure of the SAS© 8.0/QC module (SAS Institute Inc., 2008 [132]). All the other computations were performed by using the R statistical software (Venables and Ripley, 2003 ; R Development Core Team, 2007 [116]), including the R sensitivity package (Pujol, 2008 [115]) and the lhs package (Carnell, 2009 [33]).

4.3.3 Results

Principal Components

In the PCA applied to the reference design simulations, 68.1%, 16.6% and 7.8% of inertia were associated with the first three components. The correlations between the principal components and the output at time t, for $t =, 1 \ldots, T = 170$, are shown in the top row of Figure 4.2. The estimations of PCs were very precise so that the confidence bounds and the estimations are almost confounded. The first component was positively correlated with the output at all times t, with relatively little variation across time. The first principal component thus corresponded to the global INN amount. The second component discriminated the INN evolution between the beginning of plant growth and the

end. In fact, at the beginning, plants need more nitrogen for their growth than at the end. The third principal component accounted for a smaller proportion of inertia. It was mainly associated with the difference between nitrogen need at the middle of plant growth and the very last grain filling days.

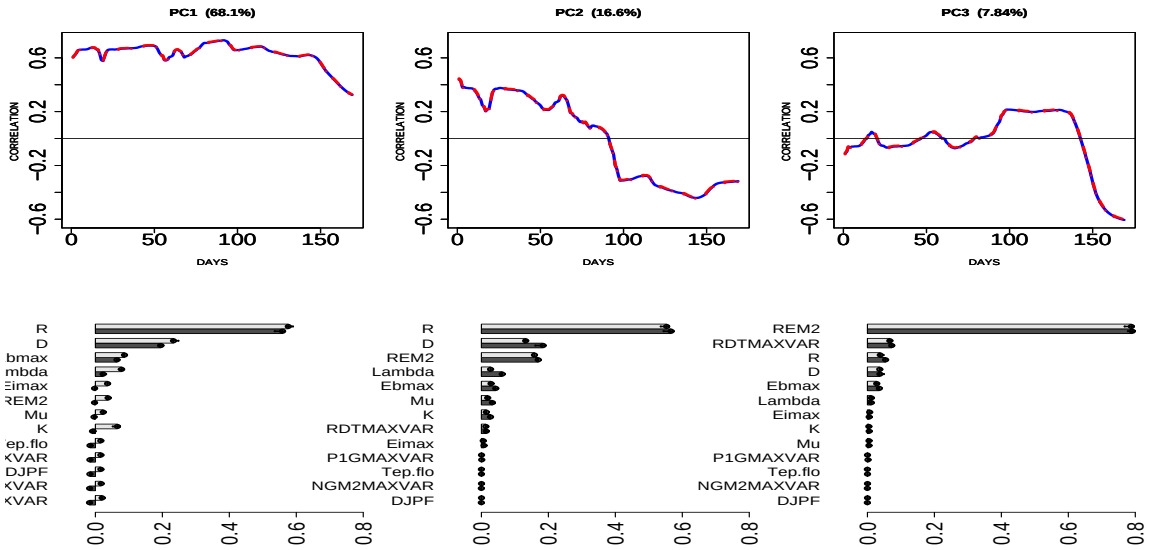

FIGURE 4.2 – Multivariate Sensitivity Analysis on the results of the reference design (Latin hypercube sampling and Sobol-Saltelli method, 150,000 simulations in total replicated five times). Top row : first three Principal Components with (almost confounded) bootstrap confidence limits. Bottom row : sensitivity indices of the first three principal components. The pale bars (respectively the dark bars) correspond to the average main effects (respectively total effects) over the five replications. The small black bars show the extreme values over the five replications.

Over all tested methods and sample sizes the first PC inertia varied in the interval $[67, 75]$, the second in $[14, 17]$ and the third one in $[6, 8]$. The empirical principal components are displayed in Figure 4.3 for the factorial design and in Figure 4.4 for the eFAST method. The results for the Sobol-Saltelli with $6,555$ simulations method are not shown but the computations of PCs with Sobol-Saltelli and the fractional design were almost the same and were very close to those obtained with the reference design. For eFAST, differences with the reference were larger, in particular for the third principal component that included two picks in the first half of the dynamics, absent from the reference. The bootstrap confidence bands were very close to the mean and so did not cater at all for

the differences between the eFAST and reference principal components.

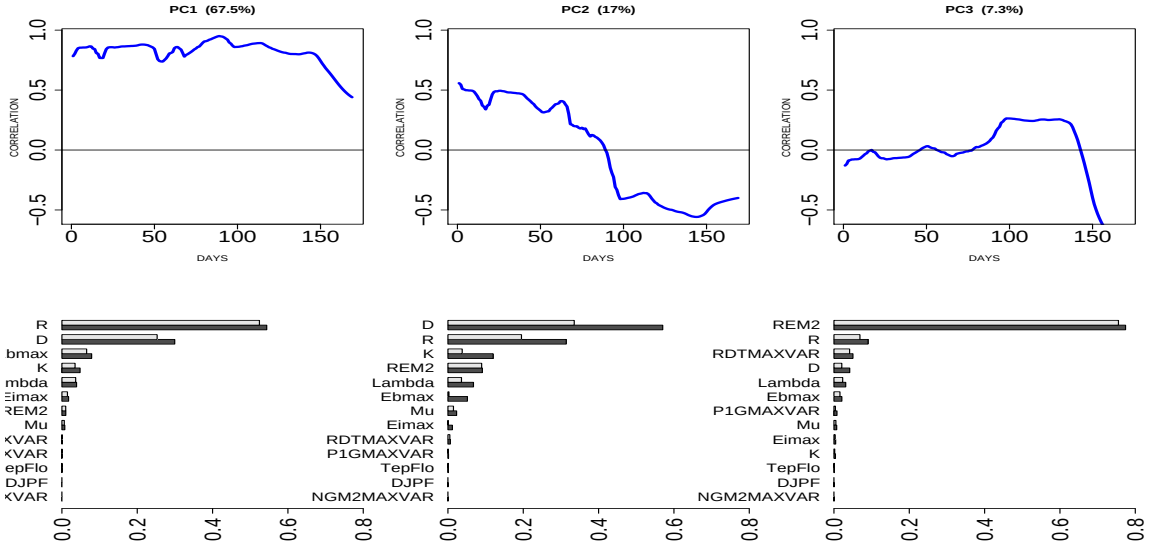

FIGURE 4.3 – Multivariate Sensitivity Analysis on the results of the fractional factorial design (6,561 simulations). Top row : first three Principal Components. Bottom row : sensitivity indices of the first three principal components. The pale bars (respectively the dark bars) correspond to the average main effects (respectively total effects) over the five replications.

Sensitivity indices on the principal components

For the reference design and for the first three principal components, the main-effect and total sensitivity indices are displayed in the bottom row of Figure 4.2. The graphics include 95% confidence intervals calculated by the quantile method using the five design replications. They are small enough to give confidence on the ranking of factors importance. To obtain such a precision, however, it proved necessary to use Latin hypercube rather than Monte Carlo sampling.

The bottom rows of Figures 4.3 and 4.4 show the ANOVA and eFAST first order and total sensitivity indices. The results for the Sobol-Saltelli method with 6,555 simulations are not shown because they were hardly interpretable, with several negative indices or with first order indices smaller than total indices, even for factors which appeared to be

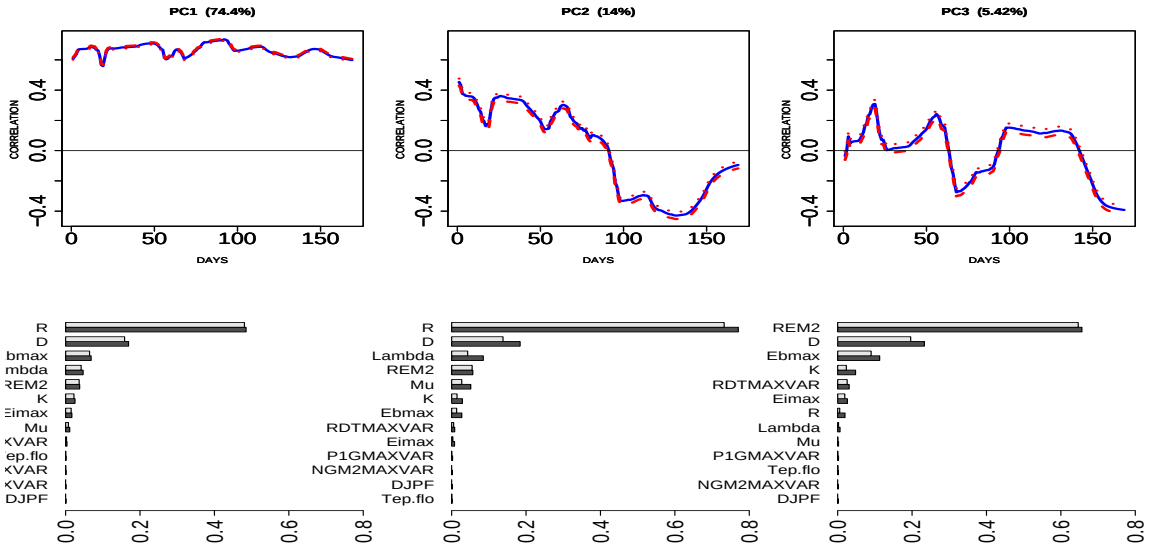

FIGURE 4.4 – Multivariate Sensitivity Analysis on the results of the eFAST method (6,552 simulations). Top row : first three Principal Components, with bootstrap confidence limits. Bottom row : sensitivity indices of the first three principal components. The pale bars (respectively the dark bars) correspond to the main effects (respectively total effects).

influent according to the reference.

The sensitivity indices on the principal components show that at least 6 parameters were not influent, according to all methods used in this paper. In all cases, the global amount of INN (PC_1) was mainly sensitive to parameter R, but also to D and $E_{b\max}$. The classification of the remaining parameters differed between the methods.

The difference between the INN at the beginning and at the end of the dynamics (PC_2) was mainly due to the parameters R, D, REM2 for the reference and for the eFAST method, and to D,R, REM2, K for the factorial design. In the factorial design case, D was the most influent parameter instead of R and moreover parameter K which seemed to be less important in the reference and eFAST method was more important than REM2. Indeed, the reference and the eFAST method did capture little interaction among parameters so that parameter K did not appear important. A major difference between the two types of methods concerned the first order and total sensitivity indices for R and D. There again the differences were partly due to more interactions being detected by the fractional design.

Finally, the third principal component was mainly sensitive to parameter REM2. This result confirmed the role of REM2 in the last stages of wheat growth, during grain filling. The sensitivity on REM2 was stable and largely dominant, but more for the reference and the fractional design than for the eFAST method.

Generalised sensitivity indices

Figure 4.5 shows the average main and total generalised sensitivity indices (GSI) and their 95% confidence intervals for the reference and for the eFAST method and fractional design. The confidence interval was computed by using the five replications of the reference.

The generalised sensitivity indices shown in Figure 4.5 provide a unique ranking of model parameters according to their influence on AZODYN-INN outputs. The generalised sensitivity indices approximated by the first five principal components gave almost the same parameter ranking for the three methods. Note that this ranking differed from those obtained on each principal component and only 6 parameters (R, D, REM2, $E_{b\max}$, Lambda, K) of 13 had a non negligible influence on the simulated INN values, all dates mixed together. The ranking of the less important parameters differed between the methods, which could be explained by their small sensitivity indices. As mentioned for the

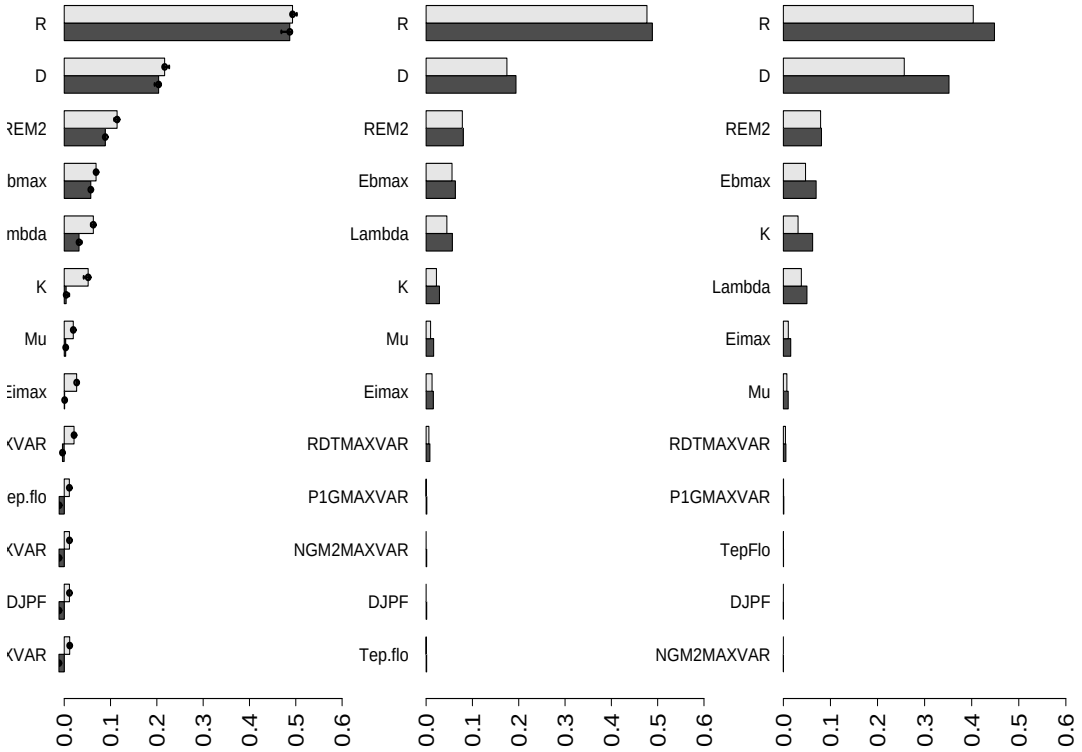

FIGURE 4.5 – Estimated Generalised Sensitivity Indices (GSI). Left : Sobol-Saltelli method ($n = 150,000$). Center : eFAST method ($n = 6552$). Right : fractional design and anova ($n = 6561$). The pale bars (respectively the dark bars) correspond to the main effects (respectively total effects). The small black bars show the extreme values over five replications.

principal component indices, the fractional design detected more interaction than the Sobol and eFAST methods.

4.4 Discussion

The combination of Principal Components Analysis and global Sensitivity Analysis leads to indices that provide rich information on dynamic model behaviour. In this paper, we show that the whole output variability can thus be decomposed into meaningful sensitivity indices. In the AZODYN-INN example, the rankings of the factors were quite different between the principal components, suggesting that different structural properties of the model were associated with the different components.

The generalised sensitivity indices synthetise the information that is spread between the time outputs or between the principal components. The factor ranking provided by these indices can be convenient for selecting a subset of parameters to be estimated from data. Thus, if the available data is scarce and allows calibration of very few parameters, the generalised sensitivity indices can be used to select these parameters. This approach was tested in Lamboni *et al.* (2009) [90] on a model of NO_2 gas emission in agricultural fields. In the AZODYN-INN example, the generalised sensitivity indices helped to identify the most interesting cultivar parameters to assess in field trials.

Although our paper focused on PCA, many alternative decompositions can be worth applying to a multivariate model output. Examples are given by Campbell *et al.* [31]. In this paper, only PCA decomposition was considered, but several sensitivity analysis methods were compared.

When the uncertainty intervals are continuous, the Sobol-Saltelli method looks most appropriate, since its Monte Carlo basis ensures convergence of the principal components estimates and of the sensitivity indices, together with adequate bootstrap assessment of its own precision. The results of our simulations on AZODYN-INN, however, showed that this method may require a very large number of simulations to give coherent and reliable results. Using Latin hypercube rather than Monte Carlo sampling appeared as a major way of improvement when mixing PCA and global SA by using the Sobol-Saltelli method.

In our examples, the eFAST method gave more coherent sensitivity indices than the Sobol-Saltelli one, when the number of simulations was relatively small (around 6500 for 13 parameters). Because of its deterministic sampling path, it may occasionally generate

bias in the principal components, that gets undetected because the conditions to apply bootstrap are not satisfied.

The fractional factorial design appeared as an interesting alternative to the methods based on intensive sampling. In the AZODYN-INN example, it provided reliable information on the principal components and on the sensitivity indices. However, the components and indices that are estimated with a factorial design are inherently biased because they are based on a discretisation of the input space. Using a fraction rather than a complete factorial design may be another source of bias. So a factorial design should be used only if the model can be assumed to be well approximated by a low-degree polynomial function, or if there is specific interest in studying a few values per input factor.

Appendix

4.4.1 Proof of Proposition 4.2.2

By combining the orthogonality properties of $\mathbf{h}_k$, for $k = 1, 2, \ldots, T$, and the anova decomposition we have,

$$\begin{aligned} ||\mathbf{h}||^2 &= \sum_{k=1}^{T} ||\mathbf{h}_k||^2 \\ &= \sum_{k=1}^{T} \sum_{w} ||\mathbf{S}_w \mathbf{h}_k||^2 \\ &= \sum_{w} \sum_{k=1}^{T} ||\mathbf{S}_w \mathbf{h}_k||^2 , \end{aligned}$$

and on the other hand,

$$||\mathbf{h}||^2 = \sum_{w} ||\mathbf{S}_w \mathbf{h}||^2 ,$$

It follows from the unique anova decomposition of $\mathbf{h}$ that $||\mathbf{S}_w \mathbf{h}||^2 = \sum_{k=1}^{T} ||\mathbf{S}_w \mathbf{h}_k||^2$. □

4.4.2 Proof of Proposition 4.2.3

We first show that H_{k_1} and H_{k_2} are uncorrelated for $k_1 \neq k_2$. Without loss of generality, let us assume that $\mathbf{Y}$ is centered. Then

$$\begin{aligned} \mathbb{E}(H_{k_1} H_{k_2}) &= \mathbb{E}\left[\mathbb{T}\mathrm{r}(\mathbf{Y}\mathbf{Y}' v_{k_1} v'_{k_2})\right] \\ &= \mathbb{T}\mathrm{r}\left[\mathbb{E}(\mathbf{Y}\mathbf{Y}') v_{k_1} v'_{k_2}\right] \\ &= \mathbb{T}\mathrm{r}\left(\Sigma\, v_{k_1} v'_{k_2}\right) \\ &= \lambda_{k_1} \mathbb{T}\mathrm{r}\left(v_{k_1} v'_{k_2}\right) \\ &= 0. \end{aligned}$$

The unique variance decomposition of H_k reads $\mathbb{V}\mathrm{ar}(H_k) = \sum_{\{j\} \in \{1,2,\ldots,d\}} \mathbb{V}_{\{j\},k} + \sum_{\{j_1,j_2\}} \mathbb{V}_{\{j_1,j_2\},k} + \ldots$ Since H_{k_1} and H_{k_2} are uncorrelated for $k_1 \neq k_2$, it follows that

$$\begin{aligned} \mathbb{V}\mathrm{ar}(H) &= \sum_{k=1}^{T} \mathbb{V}\mathrm{ar}(H_k) \\ &= \sum_{k=1}^{T} \sum_{w} \mathbb{V}_{w,k} \\ &= \sum_{w} \mathbb{G}\mathbb{V}_w, \end{aligned}$$

and on the other hand,

$$\mathbb{V}\mathrm{ar}(H) = \sum_{w} \mathbb{V}_{w,\mathrm{G}}.$$

The unique decomposition of $\mathbb{V}\mathrm{ar}(H)$ implies that $\mathbb{V}_{w,\mathrm{G}} = \mathbb{G}\mathbb{V}_w$, and so $\mathbb{GSI}_w = \frac{\mathbb{V}_{w,\mathrm{G}}}{\mathbb{V}\mathrm{ar}(H)}$.

□

Chapitre 5

Lien entre les indices de sensibilité et le $\mathbb{MSEP}$ pour un modèle non-linéaire dynamique : cas du modèle CERES-EGC

Introduction

Après avoir établis une relation formelle entre analyse de sensibilité et un critère de validation de modèle tel que le $\mathbb{MSEP}$ (Chapitre 3) sur un modèle linéaire et grâce au développement de l'analyse de sensibilité multivariée du Chapitre 4 qui permet d'avoir un indice unique pour chaque facteur, ce chapitre contribue à la mise en œuvre des méthodes développées sur des modèles dynamiques complexes. Dans un premier temps, nous conduisons i) l'analyse de sensibilité globale sur chaque sortie des modèles (WWDM et CERES-EGC) pour savoir l'évolution de l'importance des différents paramètres tout au long de la dynamique ii) l'AS sur les 3 premières composantes principales qui résument au maximum l'information contenue dans toute la dynamique iii) l'AS multivariée pour avoir une seule hiérarchisation des différents facteurs du modèle en termes d'importance sur la variabilité globale de la dynamique. Ces différentes méthodes furent comparées et sont complémentaires. La seconde partie de ce chapitre cherche à valider de manière empirique et sur un modèle non-linéaire, complexe et dynamique (CERES-EGC) la relation entre les indices de sensibilité et le $\mathbb{MSEP}$ établie dans le Chapitre 3 de ce mémoire. L'ensemble de ce travail est valorisé par un article qui est paru dans *F*ield Crops Research.

Multivariate global sensitivity analysis for dynamic crop models

Matieyendou Lamboni[1,*], David Makowski[2], Simon Lehuger[3], Benoit Gabrielle[3] and Hervé Monod[1]

[1]INRA, Unitïé MIA (UR341), Domaine de Vilvert, F78352 Jouy-en-Josas Cedex, France
[2]INRA, UMR 211 INRA AgroParisTech, BP 01, F78850, Thiverval-Grignon, France
[3]INRA, AgroParisTech UMR1091 EGC, F78850, Thiverval-Grignon, France

* Corresponding author : Matieyendou Lamboni.

Abstract

Dynamic crop models are frequently used in ecology, agronomy and environmental sciences for simulating crop and environmental variables at a discrete time step. They often include a large number of parameters whose values are uncertain, and it is often impossible to estimate all these parameters accurately. A common practice consists in selecting a subset of parameters by global sensitivity analysis, estimating the selected parameters from data, and setting the others to some nominal values. For a discrete-time model, global sensitivity analyses can be applied sequentially at each simulation date. In the case of dynamic crop models, simulations are usually computed at a daily time step and the sequential implementation of global sensitivity analysis at each simulation date can result in several hundreds of sensitivity indices, with one index per parameter per simulation date. It is not easy to identify the most important parameters based on such a large number of values. In this paper, an alternative method called multivariate global sensitivity analysis was investigated. More precisely, the purposes of this paper are i) to compare the sensitivity indices and associated parameter rankings computed by the sequential and the multivariate global sensitivity analyses, ii) to assess the value of multivariate sensitivity analysis for selecting the model parameters to estimate from data. Sequential and multivariate sensitivity analyses were compared by using two dynamic models : a model simulating wheat biomass named WWDM and a model simulating N_2O gaseous emission in crop fields named CERES-EGC. N_2O measurements collected in several experimental plots were used to evaluate how parameter selection based on multivariate sensitivity analysis can improve the CERES-EGC predictions. The results showed that sequential and multivariate sensitivity analyses provide modellers with different types of information for models which exhibit a high variability of sensitivity index

values over time. Conversely, when the parameter influence is quite constant over time, the two methods give more similar results. The results also showed that the estimation of the parameters with the highest sensitivity indices led to a strong reduction of the prediction errors of the model CERES-EGC.

Keywords : crop model, mean squared error of prediction, N_2O emission, sensitivity analysis, parameter estimation.

5.1 Introduction

Dynamic crop models are frequently used in ecology, agronomy and environmental sciences for simulating crop and environmental variables of interest at a discrete time step. These models are useful for the management of endangered species (e.g Santangelo *et al.*, 2007), for understanding intraspecific and interspecific competition (e.g Yakubu *et al.*, 2002 ; Wu *et al.*, 2007), for pest management (e.g Matsuoka and Seno, 2008), for predicting plant growth (e.g Bechini *et al.*, 2006 ; Boote *et al.*, 1996 ; Passioura, 1996) or for greenhouse gas management (Gabrielle *et al.*, 2006b). For instance, CERES-EGC is a discrete-time model that simulates the emission of nitrous oxide (N_2O), a potent greenhouse gas, into the atmosphere on a daily time step (Gabrielle *et al.*, 2006a). Discrete-time models can include many parameters whose values are uncertain. The uncertainty on the parameters is a major source of uncertainty on the model predictions. Consequently, the estimation of the uncertain parameters from experimental data is an important step and model performances depend for a large part on the accuracy of the parameter estimates (Butterbach-Bach *et al.*, 2004 ; Gabrielle *et al.*, 2006a ; Lehuger *et al.*, 2009 ; Makowski *et al.*, 2006a ; Wallach *et al.*, 2001). Model predictions based on inaccurate parameter values are unreliable and hardly meaningful.

In general, it is impossible to estimate all parameters of complex models simultaneously (Bechini *et al.*, 2005). A common strategy consists in selecting a subset of parameters to be calibrated using sensitivity analysis, and fixing the others to some nominal values (Makowski *et al.*, 2006a ; Makowski *et al.*, 2006b ; Monod *et al.*, 2006 ; Wallach *et al.*, 2001). Several local and global sensitivity analysis methods have been developed and applied to identify the parameters with a large influence on model outputs (Cariboni *et al.*, 2007 ; Homma and Saltelli, 1996 ; Saltelli *et al.*, 2000b ; Saltelli *et al.*, 2004 ; Saltelli *et al.*, 2006). Methods of global sensitivity analysis are useful and are easy to interpret. They allow modelers to perform factor prioritization, i.e to determine which subset of parameters accounts for most of the output uncertainty. Those factors with a small contribution can be set to some nominal value or to any value within their uncertainty range. The use of sensitivity analysis to select the parameters to estimate relies mainly on the intuitive idea that predictions are more accurate when the parameters with the greatest influence are estimated accurately. There have been few attempts to formalize this kind of relationship or even to check it empirically in realistic situations (Brun *et al.*, 2001 ; Tremblay and Wallach, 2004).

For a discrete-time model, global sensitivity analysis methods can be applied sequen-

tially at each simulation date. In the case of dynamic crop models, simulations are usually computed at a daily time step and the sequential implementation of global sensitivity analysis at each simulation date can result in several hundreds of sensitivity indices, with one index per simulation date. It is not easy to identify the most important parameters based on such a large number of values (Campolongo *et al.*, 2007). In addition, there is a high level of redundancy between neighbouring dates and interesting features of the dynamics may be missed out.

Campbell *et al.* (2006) proposed to use a different approach with dynamic models, called multivariate global sensitivity analysis. Their idea was to decompose the time series of model outputs upon a complete orthogonal basis and to compute sensitivity indices on each component of the decomposition. The orthogonal basis can be determined by principal component analysis from a set of model outputs computed using various combinations of parameter values. When the variability of the model outputs can be explained by a small number of principal components, this approach allow modelers to analyze the sensitivity of a large number of dynamic model outputs by computing a small number of sensitivity indices ; one index value per parameter and per principal component instead of one index value per parameter and per simulation date. Although this approach looks promising, it has never been applied to dynamic crop models

The purpose of this paper was to study the usefulness of multivariate global sensitivity for dynamic crop models. More specifically, our objectives are i) to compare the sensitivity indices and associated parameter rankings computed by the sequential and by the multivariate global sensitivity analyses, ii) to assess the value of multivariate sensitivity analysis for selecting the model parameters to estimate from data. Sequential and multivariate global sensitivity analyses were compared by using two dynamic models : a model simulating wheat biomass named WWDM (Makowski *et al.*, 2004 ; Monod *et al.*, 2006) and a model simulating N_2O gaseous emission in crop fields named CERES-EGC (Gabrielle *et al.*, 2006a, 2006b). N_2O measurements collected in several experimental plots were used to evaluate how parameter selection based on multivariate sensitivity analysis can improve the CERES-EGC predictions.

5.2 Material and methods

5.2.1 Models

The Winter Wheat Dry Matter model

The Winter Wheat Dry Matter model (WWDM) is a dynamic crop model running at a daily time step (Makowski *et al.*, 2004). It has two state variables, the above-ground winter wheat dry matter $U(t)$, in g/m^2 and the leaf area index LAI(t) with t the day number from sowing ($t = 1$) to harvest ($t = 223$). The state variable $U(t)$ is calculated on a daily basis in function of the cumulative degree-days $T(t)$ (over a basis of O°C) and of the daily photosynthetically active radiation PAR(t). The model equations are defined by

$$U(t+1) = U(t) + E_b E_{i\max} \left[1 - e^{K.\,\mathrm{LAI}(t)}\right] \mathrm{PAR}(t) + \varepsilon(t) \tag{5.2.1}$$

and

$$\mathrm{LAI}(t) = L_{\max} \left(\frac{1}{1 + e^{-A(T(t)-T_1)}} - e^{B(T(t)-T_2)} \right) \tag{5.2.2}$$

where $\varepsilon(t)$ is a random term with zero expectation representing the model error. Only the deterministic part of the model was considered for this paper and so the error term was set to zero. The dry matter at sowing ($t = 1$) was also set to zero : $U(1) = 0$. In addition, the constraint $T_2 = \frac{1}{B}\log[1 + \exp(A \times T_1)]$ was applied, so that LAI$(1) = 0$

Seven uncertain parameters were considered for the sensitivity analysis. Uncertainty intervals in Table 5.1 were given by agronomists (Monod *et al.*, 2006). Usually the climate should form one or several input factors for the sensitivity analysis. Here, preliminary investigations on 14 annual climate series showed little differences between years. For simplicity, results with a single series are presented. The model output considered in the following text is the dynamic evolution of the dry matter from sowing ($t = 1$) until harvest ($t = 223$). It is represented in Figure 5.1 for the nominal parameter values and for a sample of other possible parameter values drawn within the uncertainty ranges.

The CERES-EGC model

CERES-EGC was adapted from the CERES suite of soil-crop models, with a focus on the simulation of environmental outputs such as nitrate leaching or the emission of nitrogen oxides (Gabrielle *et al.*, 2006b). The CERES models are available for a large number of crop species that share the same soil components (Jones and Kiniry, 1986). CERES-EGC runs at a daily time step, and requires daily rain, mean air temperature

Parameter	Interpretation	Nominal value	Uncertainty interval
E_b	radiation use efficiency	1.85	0.9-2.8
$E_{i\max}$	maximal ratio of intercepted to incident radiation	0.94	0.9-0.99
K	coefficient of extinction	0.7	0.6-0.8
$L_{\max}$	maximal value of LAI	7.5	3-12
T_1	temperature threshold	900	700-1100
A	-	0.0065	0.0035-0.01
B	-	0.00205	0.0011-0.0025

TABLE 5.1 – Uncertainty intervals for the parameters of the winter wheat dry matter model.

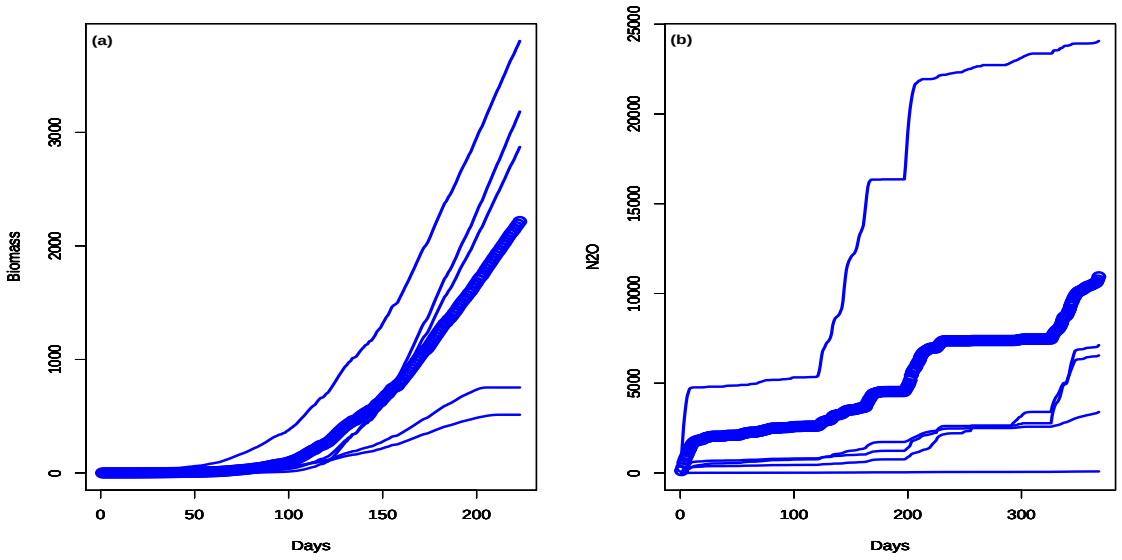

FIGURE 5.1 – Daily simulated values of the winter wheat dry matter model in g/m^2 (a) and of the CERES-EGC N_2O emissions in gN ha-1day-1 (b), for the nominal values of the parameters (thick lines) and for a sample of other possible values drawn within the uncertainty ranges.

and Penman potential evapo-transpiration as forcing variables.

The nitrous oxide emission module simulates the production of N_2O (kg N ha^{-1} day^{-1}) in soils through both the nitrification (Ni) and the denitrification (De) pathways (Hénault *et al.*, 2005 ; Hénault and Germon, 2000). Nitrous oxide emissions resulting from the two processes are soil-specific proportions of total denitrification and nitrification pathways, and are calculated according to the equations presented in the Appendix. For details on model equation see Lehuger *et al.* (2009). The N_2O sub-model of CERES-EGC involves a total set of 15 main parameters including the potential denitrification rate (PDR, kg N ha^{-1} day^{-1}), the maximum nitrification rate (MNR, kgN ha^{-1} day^{-1}), the fractions of nitrified (c) and denitrified (r) N. Parameter uncertainty intervals are listed in Table 5.2.

5.2.2 Methods of sensitivity analysis for time series output

Structure of the simulated data

The output of a dynamic crop model with discrete time step can be written

$$y(t) = f(\mathbf{z}, t; \theta), \tag{5.2.3}$$

where $y(t)$ is the scalar output on day t for $t = 1, 2, \ldots, T$, $\mathbf{z}$ is a vector of input variables and θ is a vector of parameters. Both input variables and parameters may be used as input factors for the sensitivity analysis. However, in the applications presented in this paper, $\mathbf{z}$ was fixed and so only the parameters in θ were used as input factors.

Simulations were performed according to complete or fractional factorial designs (Box and Draper, 1987 ; Ginot *et al.*, 2006). Each parameter was studied at three levels, the mean and the bounds of its uncertainty interval, making it possible to assess linear and quadratic effects. For the WWDM model, a complete 3^7 factorial design (seven parameters at three levels) was constructed, with $N = 3^7 = 2187$ simulations. For the CERES-EGC model ; a fractional factorial design 3^{15-7} (fifteen parameters at three levels and $N = 3^8 = 6561$) was constructed with the FACTEX procedure of the SASï©8.0/QC module (SAS Institute Inc., 2008) . It was of resolution 5, which means that all main effects and two-factor interactions could be estimated (Box and Draper, 1987 ; Kobilinsky, 1997).

Suppose that N simulation runs are performed according to a factorial design on the input factors. Then the output can be stored in a $N \times T$ matrix :

Parameter	Interpretation	Unit	Nominal value	Uncertainty interval
$Seuil\_wfps$	Water field pore space response threshold	-	0.62	0.4-0.8
K_m	Half saturation constant (denitrification)	mg N-NO_3 kg^{-1} soil	22	5-120
$Seuil\_t$	Temperature threshold	-	11	10-15
q_dix_un	Q10 factor for low temperature	-	89	60-120
q_dix_deux	Q10 factor for high temperature	-	2.1	1-4.8
$Puissance$	Exponent of power function	-	1.74	0-2
Opt_wfps	Optimum WFPS for nitrification	-	0.6	0.35-0.75
Min_wfps	Minimum WFPS for nitrification	-	0.1	0.05-0.15
Max_wfps	Maximum WFPS for nitrification	-	0.8	0.8-1
K_m_amm	Half saturation constant (nitrification)	mg N-NO_3 kg^{-1} soil	10	1-50
Q_dix_nit	Q10 factor for nitrification	-	2.1	1.9-13
PDR	Potential denitrification rate	kg N ha^{-1} day^{-1}	7	0.1-20
MNR	Maximun nitrification rate	kg N ha^{-1} day^{-1}	1 9	4-13
r	Fraction of denitrified N	-	0.25	0.09-0.9
c	Fraction of nitrified N	-	0.018	0.0002-0.1

TABLE 5.2 – Uncertainty intervals for the parameters of the CERES-EGC model.

$$\mathbf{Y} = \begin{pmatrix} y_1(1) & \dots & y_1(t) & \dots & y_1(T) \\ \vdots & & \vdots & & \vdots \\ y_i(1) & \dots & y_i(t) & \dots & y_i(T) \\ \vdots & & \vdots & & \vdots \\ y_N(1) & \dots & y_N(t) & \dots & y_N(T) \end{pmatrix}.$$

Each column $\mathbf{y}(t)$ in $\mathbf{Y}$ represents the simulated values of the output variable at a given time t, while each row of $\mathbf{Y}$ is an individual dynamic for a given set of input values. The rows of $\mathbf{Y}$ constitute a sample of output dynamics in $\mathbf{R}^T$ over the uncertainty domain of the input factors. In the following text, we assume that $N \geq T$.

Method 1 : sequential global sensitivity analyses

Sensitivity analysis of a time series output can first be performed separately on each output variable $y(t)$. Because orthogonal factorial designs were used for the simulations, classical analyses of variance (ANOVA) were performed. The complete variance decomposition is

$$\begin{aligned} \mathrm{SS}(\mathbf{y}(t)) &= \mathrm{SS}_1(t) + \ldots + \mathrm{SS}_i(t) + \ldots + \mathrm{SS}_K(t) + \mathrm{SS}_{1.2}(t) + \ldots && (5.2.4) \\ &\quad + \mathrm{SS}_{i.j}(t) + \ldots + \mathrm{SS}_{K-1.K}(t) + \ldots \mathrm{SS}_{1\ldots.K}(t), && (5.2.5) \end{aligned}$$

where SS_i is the main effect of parameter i, $\mathrm{SS}i.j$, say, is the interaction between parameters i and j, and K is the total number of parameters. With a complete factorial design, the decomposition can be calculated with all factorial terms. With a fractional design of resolution 5, the decomposition must be limited to two-factor interactions because of confounding.

Sensitivity indices defined by

$$\mathbb{SI}_W = \frac{\mathrm{SS}_W}{\mathrm{SS}(y(t))}, \tag{5.2.6}$$

were derived from the ANOVA sums of squares at each time t and for each factorial term W in the model (see Monod *et al.*, 2006). The dynamic evolution of sensitivity indices was represented graphically as proposed by Saltelli *et al.*, 2000 (Figure 5.2). Computations were performed using the R statistical software (Venables and Ripley, 2003 ; R Development Core Team, 2007).

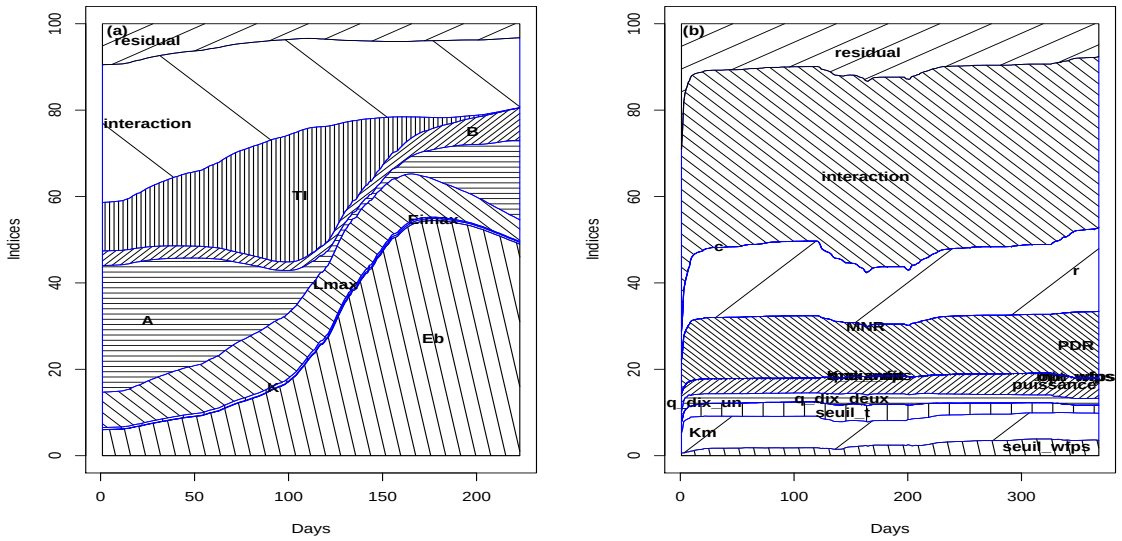

FIGURE 5.2 – Time-dependent pie charts of sensitivity indices for the WWDM model (a) and for the CERES-EGC model (b). Residual indicates the interactions between three or more parameters.

Method 2 : PCA-based multivariate global sensitivity analysis

Sensitivity analyses can also be applied to a pre-defined function h of the model outputs $h(y(1), \ldots y(T))$ with a biological interpretation. For example, $h = y(t_1) - y(t_2)$ may represent the difference in biomass between two key stages of plant growth. As many other features in the $y(t)$ dynamics are potentially interesting to look at, it is useful to identify automatically the linear combinations that contain most variability between the output dynamics. This identification step was performed here by Principal Components Analysis (PCA) of matrix $\mathbf{Y}$ (Krzanowski and Marriott, 1990 ; Anderson, 2003).

By definition, the total inertia is the sum of the $y(t)$ variances and the first principal component is the linear combination $\mathbf{h}_1 = l_{1,1}\mathbf{y}(t_1) + \ldots + l_{1,T}\mathbf{y}(t_T)$ of the columns of $\mathbf{Y}$ with the maximum proportion of inertia (or variance) subject to the constraint $\sum_{t=1}^{T} l_{1,t}^2 = 1$. Overall, there are T principal components in a decreasing order of importance as measured by the percentage of inertia. When an orthogonal factorial design is used, ANOVA-based sensitivity analysis can be applied to each principal component $\mathbf{h}_k$ according to the variance decomposition :

$$\begin{aligned} \mathrm{SS}(\mathbf{h}_k) &= \mathrm{SS}_{1,k} + \ldots + \mathrm{SS}_{i,k} + \ldots + \mathrm{SS}_{K,k} + \mathrm{SS}_{1.2,k} + \ldots + \\ &\quad \mathrm{SS}_{i.j,k} + \ldots + \mathrm{SS}_{K-1.K,k} + \ldots \mathrm{SS}_{1.\ldots.K,k}, \end{aligned} \tag{5.2.7}$$

Sensitivity indices $\mathbb{SI}_{w,k}$ were derived from ANOVA sums of squares for each main effect or interaction W and each principal component $k = 1, \ldots, T$ (see Table 5.3). The principal components were represented graphically by plotting the coefficients $l_{k,t}$ as a function of t. The sensitivity indices were also represented graphically by drawing a Pareto plot for each principal component k of interest (Figures 5.3, 5.4).

A synthetic multivariate sensitivity index

In addition to the PCA-based sensitivity indices, a synthetic sensitivity index $\mathbb{GSI}_w$ can be calculated to measure the contribution of each factorial term W to the total inertia between output dynamics. We propose to call it the generalized sensitivity index of factorial term w. $\mathbb{GSI}_w$ is equal to the weighted sum of the $\mathbb{SI}_{w,k}$ indices over the principal components k, with weights proportional to the inertia associated with the components k (Lamboni *et al.*, 2008). Generalized sensitivity indices can be represented by Pareto plots (Figure 5.5).

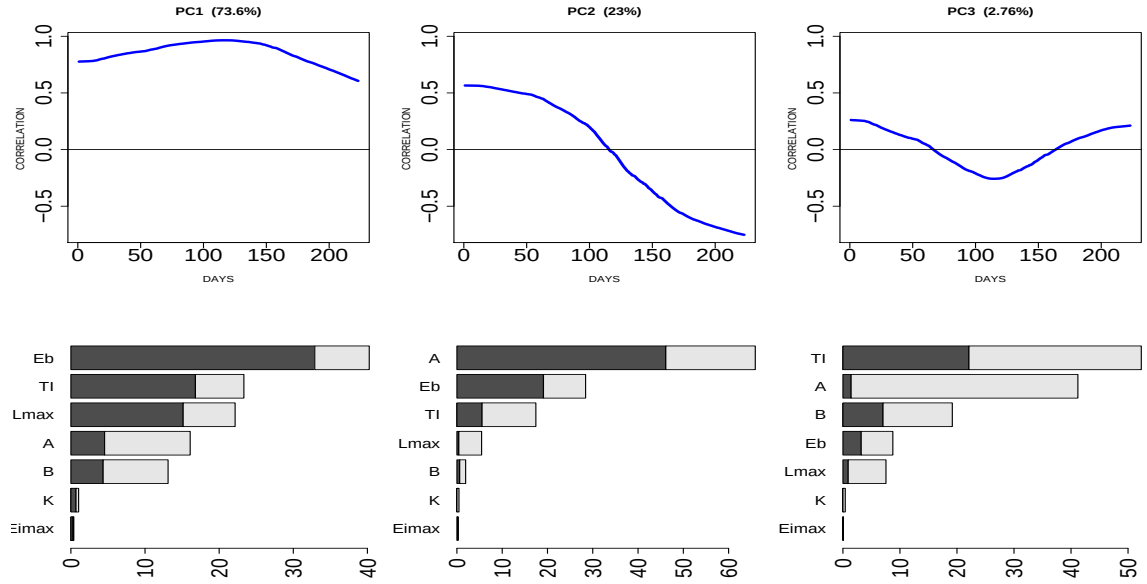

FIGURE 5.3 – PCA-based sensitivity analysis of the WWDM model. Columns : principal components 1 to 3. Top row : correlation coefficients (y-axis) between the principal component and y(t) (with t on the x-axis). Bottom row : first order sensitivity indices (dark bars) and total sensitivity indices (dark + pale bars).

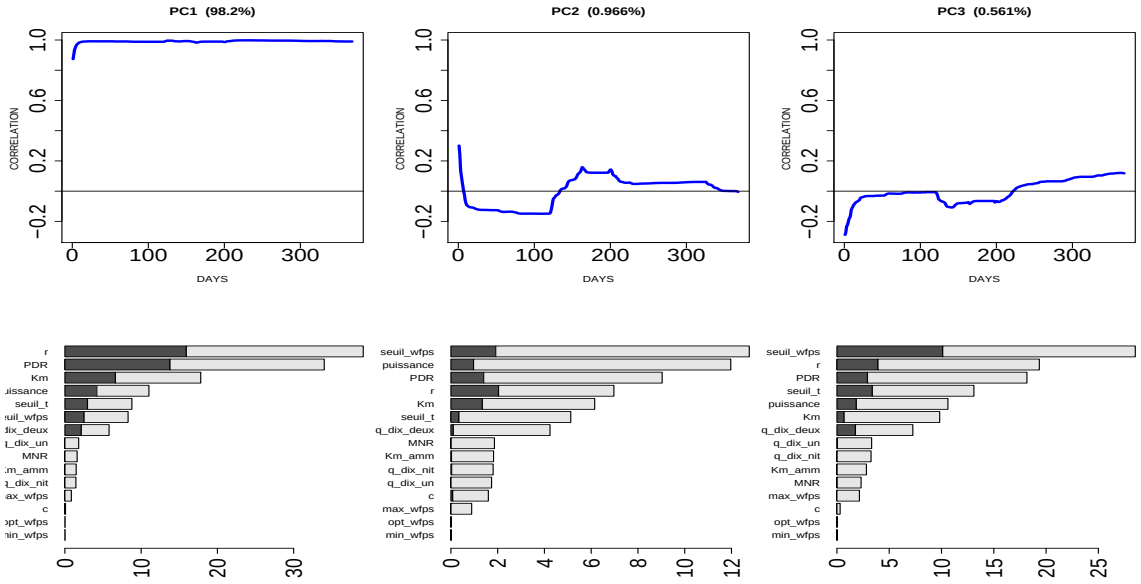

FIGURE 5.4 – PCA-based sensitivity analysis of the CERES-EGC model. Columns : principal components 1 to 3. Top row : correlation coefficients (y-axis) between the principal component and y(t) (with t on the x-axis). Bottom row : first order sensitivity indices (dark bars) and total sensitivity indices (dark + pale bars).

Factorial term	Principal Component PC_1	PC_2	PC_3	...	PC_T	Inertia
A	$SS_{A,1}$	$SS_{A,2}$	$SS_{A,3}$	...	$SS_{A,T}$	$SS_{A,\text{total}}$
B	$SS_{B,1}$	$SS_{B,2}$	$SS_{B,3}$	...	$SS_{B,T}$	$SS_{B,\text{total}}$
⋮	⋮	⋮	⋮		⋮	⋮
AB	$SS_{AB,1}$	$SS_{AB,2}$	$SS_{AB,3}$	...	$SS_{AB,T}$	$SS_{AB,\text{total}}$
⋮	⋮	⋮	⋮		⋮	⋮
W	$SS_{W,1}$	$SS_{W,2}$	$SS_{W,3}$	...	$SS_{W,T}$	$SS_{W,\text{total}}$
⋮	⋮	⋮	⋮		⋮	⋮
Inertia	λ_1	λ_2	λ_3	...	λ_T	$\mathbb{I}$

TABLE 5.3 – Sum of squares decomposition of the total inertia based on principal component analysis and MANOVA. A and B denote the first two parameters, AB their interaction, and W a generic factorial term.

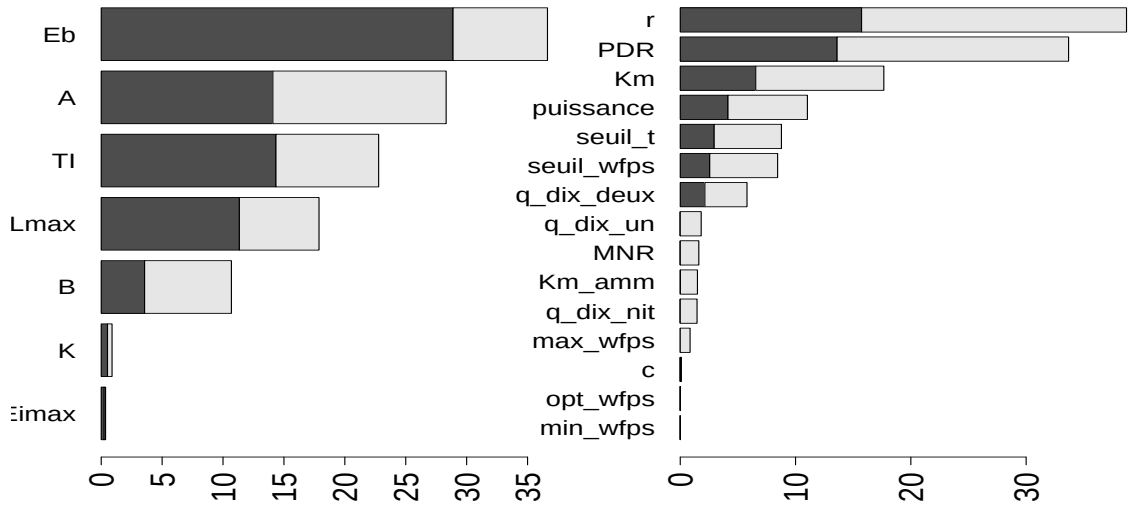

FIGURE 5.5 – Generalized Sensitivity Indices for the WWDM model (left) and for the CERES-EGC model (right). The main sensitivity indices are in dark bars and interaction ones are in pale bars. The total length of any bar represents the total sensitivity index.

In practice, only the first principal components carry useful information on the model output and higher-order interactions can be neglected. An approximation of $\mathbb{GSI}_w$ based on a subset of the principal components is called an approximate generalized sensitivity index and denoted by $\hat{\mathbb{GSI}}_w$ in the following text. To quantify the approximation made when restricting the interpretation to a subset of principal components and a subset of factorial terms, the total proportion of inertia associated with these subsets is called the approximation global quality and denoted by GQ. If GQ is close to 1 then the approximation accounts for most of the inertia of $\mathbf{Y}$, whereas low GQ suggests that $\hat{\mathbb{GSI}}_w$ indices must be interpreted with much caution. In addition to the global quality criterion, dynamic coefficients of determination R_t^2 can be used to measure the quality of any approximation when coming back to the original time series output $y(t)$ (Lamboni *et al.*, 2008).

5.2.3 Mean Squared Error of Prediction ($\mathbb{MSEP}$) of the CERES-EGC model

$\mathbb{MSEP}$ values were computed in order to see if model predictions were improved when sets of parameters with high sensitivity indices were estimated from experimental data. Only the CERES-EGC model was considered at this step because no experimental data was available for estimating the parameters of WWDM.

Experimental data

$\mathbb{MSEP}$ were computed from data collected in an experimental trial carried out in Villamblaim (Central France), during the 1998-1999 growing season. A winter wheat crop was grown with conventional management on a haplic Calcisol soil. The emissions of N_2O emissions were monitored at 18 different dates throughout the growing season using static chambers with eight replicates. At each sampling date, the chambers were closed with an airtight lid, and the head space was sampled 4 times over a period of 2 hours. The gas samples were stored in 3-mL Vacutainer tubes (Terumo Europe N.V., Leuven, Belgium), and analyzed in the laboratory by gas chromatography. See Gabrielle *et al.* (2006a) for a detailed description of the experimental methods. All the input variables required by CERES-EGC were measured in the experimental plot.

Statistical analysis

All possible sets of one, two or three parameters were defined from the 15 model parameters. A pooled value of the generalized sensitivity index defined above was calculated

for each set by summing all the main effect and interaction terms of the parameters included in the set. Each set of one, two or three parameters was then estimated by using half of the 18 N_2O measurements. The other nine measurements were used to assess the errors of prediction of the fitted model. Data were permuted in order to estimate the $\mathbb{MSEP}$ by cross-validation for each set of one, two or three parameters. Finally, the relationship between the generalized sensitivity index and the $\mathbb{MSEP}$ was investigated graphically. In the estimation procedure, the non-selected parameters were fixed at their nominal values while the selected ones were estimated by least squares, subject to the constraint that all parameter values should remain in their uncertainty interval. We did not estimate more than three parameters because of the limited number of data.

5.3 Results

5.3.1 Sequential global sensitivity analyses

The results obtained by sequential sensitivity analysis of the WWDM model (Figure 5.2a) showed that the values of the sensitivity indices strongly vary over time. Before $t = 50$, the most important parameter is A followed by parameters T_1, $L_{\max}$, and B. The strong influence of A at the beginning of the growing season is due to its influence on the increase of the leaf area index which occurs at this stage. The influence of A decreases and becomes less important than T_1, E_b, and $L_{\max}$ after $t = 110$. In the second half of the growth cycle, biomass growth becomes very sensitive to the parameter E_b due to the effect of this parameter on the conversion of the intercepted radiation into biomass. The influence of parameter A increases again after $t = 150$ and until harvest due to leaf senescence. Interactions between parameters are important during the whole growing period.

Unlike the WWDM model, the sensitivity indices computed for CERES-EGC are quite similar over time. The simulated values of N_2O emission are driven by the same three parameters during the whole period of simulation ; r, PDR, K_m (Figure 5.2b). For this model too, the interactions are important ; the sensitivity index associated to interaction represents more than 40 % of the total variability. This strong interaction is due to the fact that the parameter effects are not additive.

5.3.2 Multivariate sensitivity analysis

Results of the principal components and sensitivity principal indices are presented in Figure 5.3 for WWDM. For this model, the first three components explained 99% of the

total inertia of the simulated dry matter dynamics. The inertia percentage associated with the first three components were equal to $\lambda_1 = 0.73$, $\lambda_2 = 0.23$ and $\lambda 83 = 0.03$ respectively. The first component was positively correlated with all outputs $y(t)$. The largest correlations were obtained in the middle of the simulation period but the correlation values were quite similar over time. According to these correlation values, the first principal component corresponds to the global amount of dry matter produced during the growing season. The sensitivity indices computed for this component (Figure 5.3, bottom row) show that the global amount of dry matter was mainly sensitive to parameter E_b, but also to T_1 and $L_{\max}$. The second principal component was positively correlated with dry matter during the first part of the growing season and negatively correlated with dry matter during the second half of the growing season. Thus, this principal component corresponds to the difference between early and late dry matter productions. It was mainly sensitive to parameter A. Finally, the third principal component accounted for a much smaller part of inertia, associated with the difference between the dry matter produced the middle of the growing season and the dry matter produced both very early and late. It was sensitive to T_1.

For CERES-EGC, the first three principal components explained more than 99% of the model output inertia. The first component corresponds to the mean of the simulated N_2O emission values. This component is sensitive to the parameters r, PDR, and K_m (Figure 5.4). The second and third principal components correspond to the difference between the N_2O values simulated during the first and second halves of the simulation period, but the second component is positively correlated to the values simulated at a very early stage. The second component is strongly influenced by $Seuil\_wfps$ and $puissance$ whereas the third component is strongly influenced by $Seuil\_wfps$ and r.

5.3.3 Generalized sensitivity indices

The Generalized sensitivity indices (GSI) are shown in Figure 5.5. These indices provide a unique ranking of the parameters. For WWDM, Figure 5.5a shows that E_b, and then A and T_1 had the strongest influence on the simulated dry matter values, all dates mixed together. Such ranking is quite convenient for selecting a set of parameters to be estimated from data. Thus, if one has to choose two parameters for calibration, it should be E_b and A according to Figure 5.5a.

For CERES-EGC, Figure 5.5b shows that the parameters with the strongest effects were r, PDR and then K_m over the entire simulation period. Parameters min_wfps,

opt_wfps and c had negligible effects. Thus, if one had to estimate two parameters of the CERES-EGC model, it should be r and PDR according to the generalized sensitivity indices.

5.3.4 Parameter selection and estimation (CERES-EGC model)

Figure 5.6 shows the empirical relationship between MSEP and the generalized sensitivity index (GSI) values for the CERES-EGC model when one, two or three parameters were estimated from data. $\mathbb{MSEP}$ are presented in function of the pooled sensitivity index of each set of parameters. Overall, $\mathbb{MSEP}$ strongly decreases when $\mathbb{GSI}$ increases. This result shows that the $\mathbb{MSEP}$ was decreased, and so the prediction accuracy improved, when the parameters with the highest sensitivity indices were estimated. This is an argument in favor of $\mathbb{GSI}$ for selecting the parameters to estimate from data. According to Figure 5.6, the estimation of the sets of parameters with the highest sensitivity indices is a guarantee for a low $\mathbb{MSEP}$. Figure 5.6 also shows that, in some cases, low $\mathbb{MSEP}$ were reached by estimating sets of parameters with low sensitivity indices but this was not systematic. The estimation of sets of parameters with low sensitivity indices is thus very risky.

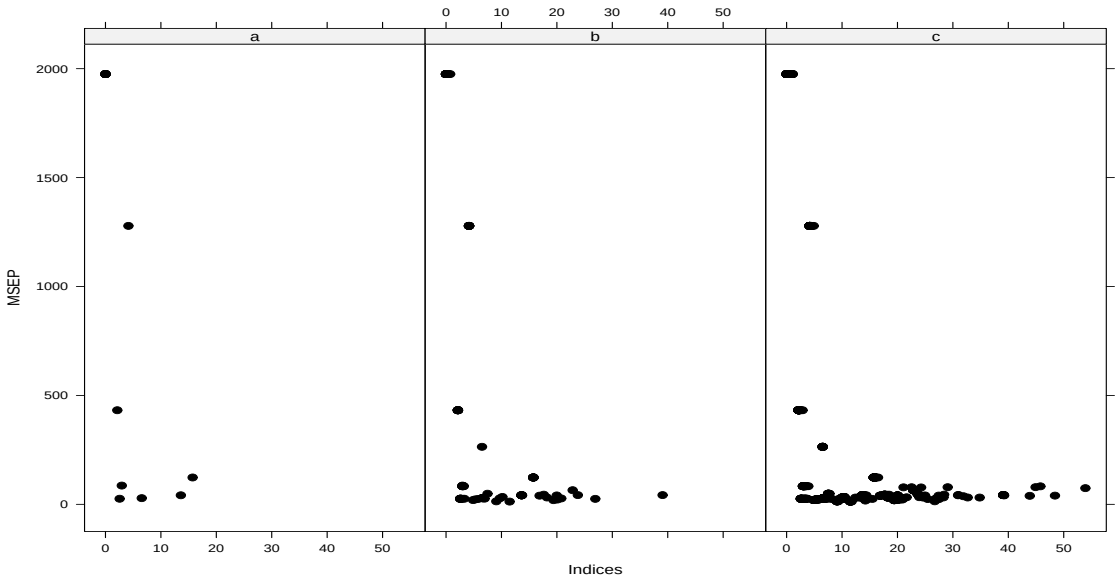

FIGURE 5.6 – Empirical relation between $\mathbb{MSEP}$ and generalized sensitivity index (GSI) when estimating one parameter (a) ; two parameters (b) and three parameters (c) of CERES-EGC model.

5.4 Discussion and conclusion

The results presented above first show that, when performing sensitivity analysis on a dynamic model, it is most useful to consider the output over the whole time series. The practical comparison between the sequential and multivariate sensitivity analysis showed that these methods are complementary.

In the sequential sensitivity analysis (Saltelli *et al.*, 2000a ; Pacala *et al.*, 1996), the sensitivity index is a function of time that measures when any given factor is more influent. Conducting separate sensitivity analyses on $y(1), \ldots, y(T)$ gives information on how the sensitivity of $y(t)$ evolves over time. However, it usually leads to a high level of redundancy because of the strong relationship between responses from one date to the next one. It may also miss important features of the $y(t)$ dynamics because many features cannot be efficiently detected through single-time measurements.

The second type of multivariate sensitivity analysis, proposed by Campbell *et al.* (2006), decomposes the crude outputs into the non-correlated principal components and computes sensitivity indices on each PC. After interpreting the PCs, this analysis allows to understand more precisely the role of some parameters in the model. Sensitivity indices on each PC can give different ranking of model parameters and we showed that the overall effect of each parameter can be summarized by a global single sensitivity index.

These two methods are useful because they provide modellers with different types of information. The application of both methods is more interesting for models which exhibit much variability of sensitivity indices over time. Conversely, when the parameter influence is quite constant over time, the two methods give more similar results.

In addition to yielding information on model behaviour, sensitivity indices can be useful to select parameters before calibration. This is an intuitive and reasonable statement. However, there is no automatic relationship between sensitivity indices, which are based on simulated data purely, and prediction quality, which depends on experimental data. Discrepancies may arise because of modelling approximations, bad choice of the uncertainty intervals and nominal values, mesurement errors, or correlations between parameter estimates resulting from partial confounding in the data. To our knowledge, the relationship has rarely been verified using real data and a rigorous cross-validation approach (Brun *et al.*, 2001 ; Tremblay and Wallach, 2004). In this paper we proposed and applied such an approach in the particular case when predictions have to be made to complement

observations that are too much dispersed in time. We found a relationship between $\mathbb{MSEP}$ and sensitivity indices ; our results showed that the estimation of the parameters with the highest sensitivity indices led to a strong reduction of the prediction errors of the model CERES-EGC. However, the estimation of parameters with the highest sensitivity indices did not lead systematically to the very smallest $\mathbb{MSEP}$ and, conversely, small sensitivity indices did not lead systematically to a high $\mathbb{MSEP}$. In our opinion, these results give weight to the use of sensible sensitivity analyses for selecting the parameters to estimate, especially since such a data-free approach to selection avoids selection bias.

The methods presented in this paper can be applied to any dynamic model predicting one or several output variables at a discrete time step. In the future, it will be interesting to apply and evaluate them on other modelling and prediction situations. Besides, they are quite flexible and extensions can be researched in several directions. For example, principal components could be made more flexible by considering functional principal components (Ramsay and Silverman, 1997) or Legendre polynomials (Campbell *et al.*, 2006), while well-designed Monte Carlo simulations could be a useful alternative to factorial designs in various situations.

Acknowledgements

We are grateful to colleagues of the Mexico ("Méthodes pour l'EXploration Informatique des modèles COmplexes") network for useful discussion.

Appendix : Equations of the N_2O emission sub-model : CERES-EGC

$N_2O(t)$ is the nitrous oxide emissions on day t, and is calculated by :

$$N_2O(t) = r \times D_e(t) + c \times N_i(t),$$

where the denitrification process D_e equation is :

$$D_e(t) = \begin{cases} PDR \frac{[NO_3^-(t)]}{K_m + [NO_3^-(t)]} \left[\frac{WFPS(t) - Seuil\_wfps}{1 - Seuil\_wfps}\right]^{puissance} F_T(t) & \text{if } WFPS(t) \geq Seuil\_wfps \\ 0 & \text{else} \end{cases}$$

with,

$$F_T(t) = \begin{cases} \exp\left[\frac{[T(t) - Seuil\_t]\ln(Q\_dix\_un) - 9\ln(Q\_dix\_deux)}{10}\right] & \text{if } T(t) < Seuil\_t \\ \exp\left[\frac{[T(t) - 20]\ln(Q\_dix\_deux)}{10}\right] & \text{else} \end{cases}$$

and the nitrification process N_i is described by :

$$N_i(t) = MNR \times \exp\left[\left[\frac{[T(t) - 20]\ln(Q\_dix\_nit)}{10}\right] \times \frac{[NH_4^+(t)]}{K_{m\_amm} + [NH_4^+(t)]} \times F_w(t)\right.$$

with,

$$F_w(t) = \begin{cases} \frac{WFPS(t) - Min\_wfps}{Opt\_wfps - Min\_wfps} & \text{if } Min\_wfps \leq WFPS(t) \leq Opt\_wfps \\ \frac{Max\_wfps - WFPS(t)}{Max\_wfps - Opt\_wfps} & \text{if } Opt\_wfps \leq WFPS(t) \leq Max\_wfps \end{cases}$$

In these equations, $WFPS(t)$ is the soil Water Filled Pore Space input variable at days t ; $NO_3^-(t)$ is the soil nitrate content (mg N kg^{-1} soil) at days t ; $T(t)$ is the temperature input variable at days t and $NH_4^+(t)$ is the soil ammonium content (mg N kg^{-1} soil). The CERES-EGC model parameters are listed in Table 5.2.

References

[1] Anderson, T.W., 2003.*An Introduction to Multivariate Statistical Analysis* (3rd ed., 721 pp.). Wiley, New York.
[2] Bechini, L., Bocchi, S., Maggiore, T., Confalonieri, R., 2006. Parameterization of a crop growth and development simulation model at sub model component level. An example for winter wheat (Triticum aestivum L.). *Environmental Modelling & Software* 21, 1042-1054.
[3] Boote, K.J., Jones, J.W., Pickering, N.B., 1996. Potential uses and limitations of crop models. *Agronomy Journal* 88, 704-716.
[4] Box, G.E.P., Draper, N.R. (1987). *Empirical Model Building and Response Surfaces*. Wiley, New York.
[5] Brun, R., Reichert, P., Kunsch, H.R. 2001. Practical identifiability of large environmental simulation models. *Water Resources Research* 37, 1015-1030.
[6] Butterbach-Bahl, K., Kesik, M., Miehle, P., Papen, H., Li C., 2004. Quantifying the regional source strength of N-trace gases across agricultural and forest ecosystems with process based models. *Plant and Soil260, 311-329*.
[7] Campbell, K., McKay, M.D., Williams, B.J., 2006. Sensitivity analysis when model outputs are functions. *Reliability Engineering and System Safety* 91, 1468-1472.
[8] Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environmental Modelling & Software 22, 1509-1518.
[9] Cariboni, J., Galtelli, D., Liska, R., Saltelli, A., 2007. The role of sensitivity analysis in ecological modelling. *Ecological Modelling* 203,167-182.
[10] Gabrielle, B., Laville, P., Hénault, C., Nicoullaud, B., Germon, J. C., 2006a. Simulation of nitrous oxide emissions from wheat-cropped soils using CERES. *Nutrient Cycling in Agroecosysyems* 74, 133 ?146.
[11] Gabrielle, B., Laville, P., Duval, O., Nicoullaud, B., Germon, J. C., Hénault, C., 2006b. Process-based modeling of nitrous oxyde emissions from wheat-cropped soils at the sub-regional scale. *Global Biogeochemical Cycles* 20, GB4018 X.1-x.13.
[12] Ginot, V., Gaba, S., Beaudouin, R., Aries, F., Monod H., 2006. Combined use of local and ANOVA-based global sensitivity analyses for investigation of a stochastic dynamic model : application to the case study of an individual-based model of a fish population. *Ecological Modelling* 193, 479-491.
[13] Hénault, C., Germon, J. C., 2000. NEMIS, a predictive model of denitrification on the field scale. *European Journal of Soil Science* 51, 257 ?270.
[14] Hénault, C., Bizouard, F., Laville, P., Gabrielle, B., Nicoullaud, B., Germon, J. C., Cellier, P., 2005. Predicting in situ soil N_2O emission using NOE algorithm and soil database. *Global Change Biology* 11, 115 ?127.

[15] Homma, T., Saltelli, A., 1996. Importance measure in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety* 52, 1-17.
[16] Jones, C. A., Kiniry, J. R., 1986. CERES-N Maize : A Simulation Model of Maize Growth and Development. *Texas A & M University Press*, College Statio, Temple, TX.
[17] Kobilinsky, A., 1997. Les plans factoriels. In : Droesbeke J.-J., Fine J., Saporta G. (Eds.), *Plans d'expériences : application à l'entreprise*. Editions Technip, Paris, Chapter 3, pp. 69-209.
[18] Krzanowski, W.J., Marriott, F.H.C., 1990. *Multivariate Analysis. Part1.* Distribution, Ordination, Inference. E. Arnold, London.
[19] Lamboni, M., Makowski, D., Monod, H., 2008. Multivariate global sensitivity analysis for discrete-time models. *Technical report* 2008-3, 17 pp, Unité MIA, INRA Jouy-en-Josas.
[20] Lehuger, S., Gabrielle, B., VanOijen, M., Makowski, D., Germon, J. C., Morvan, T., Henault, C., 2009. Bayesian calibration of the nitrous oxide emission module of an agro-ecosystem model. Agric. Ecosyst. Environ., in press.
[21] Makowski, D., Jeuffroy, M.-H., Gue´ rif, M., 2004. In : Van Boekel, et al. (Eds.),Bayesian Methods for Updating Crop Model Predictions, Applications for Predicting Biomass and Grain Protein Content. *Bayesian Statistics and Quality Modelling in the Agro-food Production Chain*. Kluwer, Dordrecht, pp. 57-68.
[22] Makowski, D., Hillier, J., Wallach, D., Andrieu, B., Jeuffroy, M.-H., 2006a. Parameter estimation for crop models. In :Wallach,D.,Makowski, D., Jones, J. (Eds.), *Working with Dynamic Crop Models*. Elsevier, Amsterdam, pp. 101-150.
[23] Makowski, D., Naud, C., Jeuffroy, M.-H., Barbottin, A., Monod, H., 2006b. Global sensitivity analysis for calculating the contribution of genetic parameters to the variance of crop model predictions. *Reliability Engineering and System Safety* 91, 1142-1147.
[24] Matsuoka, T., Seno, H., 2008. Ecological balance in the native population dynamics may cause the paradox of pest control with harvesting.J*ournal of Theoretical Biology* 252, 87-97.
[25] Monod, H., Naud, C., Makowski, D., 2006. Uncertainty and sensitivity analysis for crop models. In : Wallach, D., Makowski, D., Jones, J. (Eds.), *Working with Dynamic Crop Models*. Elsevier, Amsterdam, pp. 55-100.
[26] Pacala, S.W., Canham, C.D., Saponara, J., Silande rJr., J.A., Kobe, R.K., Ribbens, E., 1996. Forest models defined by field measurements : estimation, error analysis, and dynamics. *Ecological Monographs* 66, 143.
[27] Passioura, J.R., 1996. Simulation models : science, snake oil, education or engineering. *Agronomy Journal* 88, 690-694.
[28] R Development Core Team, 2007. R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[29] Ramsay, I., Silverman, B., 1997. *Functional Data Analysis*. Springer, New York, p. 311.

[30] Saltelli, A., Chan, K., Scott, E.M. (Eds.), [90_TD$DIFF]2000a. Sensitivity Analysis. Wiley, New York, p. 475.

[31] Saltelli, A., Ratto, M., Tarantola, A., Campolongo, F., 2006. Sensitivity analysis practices : strategies for model based inference. *Reliability Engineering and System Safety* 91, 1109-1125.

[32] Saltelli, A., Tarantola, A., Campolongo, F., 2000b. Sensitivity analysis as an ingredient of modelling. *Statistical Science* 15, 377-395.

[33] Saltelli, A., Tarantola, A., Campolongo, F., Ratto, M., 2004. *Sensitivity Analysis in Practice*. Wiley, New York.

[34] Santangelo, G., Bramanti, L., Iannelli, M., 2007. Population dynamics and conservation biology of over-exploited Mediterranean red coral. *Journal of Theoretical Biology* 244, 416-423.

[35] SAS Institute Inc., 2008. SAS/QC Users Guide. SAS Institute Inc., Cary, NC http : v8doc.sas.comsashtmlqcchap14index.htm.

[36] Tremblay, M., Wallach, D., 2004. Comparison of parameter estimation methods for crop models. *Agronomie* 24, 351-365.

[37] Venables, W.N., Ripley, B.D., 2003. *Modern Applied Statistics with S*, 4th ed. Springer, Berlin.

[38] Wallach, D., Goffinet, B., Bergez, J.E., Debaeke, P., Leenhardt, D., Aubertot, J.N., 2001. Parameter estimation for crop models : a new approach and application to a corn model. *Agronomy Journal* 93, 757-766.

[39] Wu, H.-H., Lee, H.-J., Horng, S.-B., Berec, L., 2007. Modeling population dynamics of two cockroach species : effects of the circadian clock, interspecific competition and pest control. *Journal of Theoretical Biology* 249, 473-486.

[40] Yakubu, A.-A., Castillo-Chavez, C., 2002. Interplay between local dynamics and dispersal in discrete-time metapopulation models. *Journal of Theoretical Biology* 218, 273-288.

Conclusion et perspectives

Les modèles dynamiques sont des outils importants et pratiques pour l'aide à la décision. Ils permettent en particulier de prédire les observations et de simuler les impacts de certaines pratiques à moindre coût. En plus de l'incertitude sur la structure du modèle, les modèles dynamiques utilisés en agronomie et en environnement contiennent souvent de nombreux paramètres incertains alors que les décisions prises à l'aide de ces modèles dépendent des valeurs de ces paramètres. L'importance de l'estimation des paramètres est alors essentielle dans le processus de la modélisation.

L'analyse de sensibilité est souvent un critère de sélection des paramètres à estimer dans la modélisation des phénomènes agronomiques et environnementaux et il est intéressant de bien appréhender les limites de cette procédure pour en faire un usage à bon escient. Nous distinguons deux grandes parties dans ce mémoire : i) le lien entre l'analyse de sensibilité et la qualité de l'estimation et de la prédiction ii) les méthodes d'analyse de sensibilité multivariée utilisées pour l'étude du lien entre les indices et l'estimation, mais aussi intéressantes en soi.

Nous confrontons d'abord la procédure de sélection des paramètres basée sur les observations virtuelles aux critères de qualité du modèle tels que le $\mathbb{MSE}$ et le $\mathbb{MSEP}$. Nous formalisons dans un premier temps les différents concepts utilisés par les modélisateurs dans la pratique pour un modèle dynamique et nous montrons dans un second temps sur un modèle linéaire la relation qui existe entre les indices de sensibilité et les deux critères $\mathbb{MSE}$, $\mathbb{MSEP}$. La sélection des paramètres les plus influents contribue à améliorer la qualité du modèle toutes choses égales par ailleurs. Les résultats du Chapitre 3 suggèrent néanmoins que des critères associant les indices de sensibilité et les caractéristiques du plan d'expérience pourraient être plus pertinents pour améliorer la qualité du modèle. Il serait intéressant de poursuivre dans cette voie pour proposer de nouveaux indices. Cependant, cette relation est complexe même pour le modèle linéaire considéré dans l'étude. D'une manière générale, l'estimation des paramètres les plus influents ne réduit pas systématiquement le $\mathbb{MSEP}$ ni le $\mathbb{MSE}$ et la fixation du reste des paramètres introduit

un biais.

Pour le modèle dynamique non linéaire (CERES-EGC) utilisé comme un cas d'étude, nous retrouvons des résultats cohérents avec l'étude du modèle linéaire. En appliquant cette procédure de sélection de paramètres pour le modèle CERES-EGC afin de prédire les émissions entre des dates d'observations dispersées, nous avons établi une relation empirique entre les indices de sensibilité et le $\mathbb{MSEP}$. Les résultats montrent que l'estimation de paramètres les plus influents réduit l'erreur de prédiction du modèle CERES-EGC. Toutefois, la sélection des paramètres moins influents ne conduit pas systématiquement à augmenter l'erreur de prédiction. Ce résultat d'une certaine manière renforce la sélection de paramètres par les indices de sensibilité du fait que cette procédure de sélection évite le problème de biais de sélection rencontrée lorsque les mêmes données sont utilisées pour sélectionner et estimer les paramètres.

La complexité de la relation établie dans le Chapitre 3 sur un modèle aussi simple attire notre attention sur la qualité du modèle quand on n'estime que les paramètres les plus influents. Les résultats obtenus sur le modèle linéaire peuvent être étendus aux modèles complexes utilisés par les modélisateurs soit en les linéarisant grâce au développement limité de Taylor soit en construisant des méta-modèles qui sont souvent linéaires. Mais, il serait potentiellement intéressant de chercher à relier la qualité du modèle aux indices de sensibilité en considérant directement les définitions théoriques de la qualité du modèle et des indices de sensibilité sans passer par un modèle linéaire.

L'étude de la relation empirique entre le $\mathbb{MSEP}$ et les indices de sensibilité établie dans le cas particulier du modèle dynamique non-linéaire nécessitait le développement d'un indice unique par facteur qui prenne en compte les corrélations entre les différentes sorties du modèle. Dans ce mémoire, nous proposons une méthode générique d'analyse de sensibilité multivariée basée sur la décomposition de l'inertie. Cette méthode prend en compte non seulement toute la dynamique du modèle mais aussi les corrélations induites par les différentes sorties du modèle. La variabilité des dynamiques des modèles est mesurée par l'inertie qui est une métrique dans l'espace des variables et est l'équivalent de la variance dans le cas d'un modèle à une seule sortie. Nous montrons comment décomposer l'inertie en des parts d'inerties expliquées par les différents facteurs et leurs interactions à l'image de la décomposition de la variance. Dans un premier temps, cette décomposition a été effectuée pour des facteurs supposés discrets. Ceci rend cette méthode d'analyse de sensibilité multivariée bien adaptée aux modèles dynamiques qui incluent de nombreux paramètres et nécessitent une méthode moins coûteuse en temps de calcul pour identifier

les facteurs les plus influents. Cette méthode à été exposée plusieurs fois et appliquée à des modèles variés, et toujours appréciée par le recul qu'elle donne aux modélisateurs pour mieux comprendre le comportement de leur modèle. Néanmoins, cette méthode ne permet pas de balayer toute la gamme d'incertitude et les indices obtenus fournissent une information sur la pertinence des facteurs que l'on peut considérer comme incomplète, voire biaisée. Dans un second temps, nous proposons donc une méthode qui prend en compte toute la gamme d'incertitude des facteurs et traitons aussi le cas d'une sortie fonctionnelle.

L'ACP considérée dans le développement de notre méthode d'analyse de sensibilité multivariée ne permet pas de prendre en compte les corrélations non-linéaires existant entres les différentes sorties du modèle. En présence d'un phénomène chaotique par exemple, une façon d'améliorer la procédure proposée serait de substituer la base adaptée fournie par l'ACP par une autre base qui corresponde bien au phénomène. Dans ce cas de figure, il serait par ailleurs intéressant de mettre en œuvre une méthode qui sélectionne les éléments de la base qui contribuent le plus à la variabilité des sorties dynamiques du modèle. Une telle stratégie contribuera à la réduction de la dimension qui est essentielle lorsqu'on manipule des bases choisies à l'avance et permettra de gagner en temps de calcul. On pourra éventuellement faire appel aux différentes méthodes de pénalité notamment les nouvelles approches de sélection telles que LARS (LASSO) pour faire ce choix.

Une étude comparative entre la méthode proposée, l'analyse de sensibilité dynamique (Saltelli *et al.*, 2000 [129]) et l'analyse de sensibilité sur les composantes principales a été effectuée sur le modèle agri-environnemental CERES-EGC et sur le modèle de culture WWDM. Elle montre la complémentarité et la cohérence de ces trois méthodes. Un facteur influent sur toute la dynamique ou sur les premières composantes principales est jugé influent par notre méthode. Par contre, il est plus difficile avec les indices dynamiques et les indices calculés sur les composantes principales de fournir un classement unique des facteurs dès que l'influence des facteurs change dans le temps. Une étude comparative entre les deux méthodes d'analyse de sensibilité multivariée que nous proposons dans ce mémoire (facteurs discrets et facteurs continus) à l'aide du modèle AZODYN montre que la classification des paramètres par l'approche discrète est moins coûteuse en temps de calcul et proche de l'approche intensive considérée comme référence. (729 évaluations du modèle contre 150000).

Bien que, notre méthode d'analyse de sensibilité multivariée ait été appliquée sur plusieurs modèles avec succès, il est souhaitable, de préciser les propriétés de convergence en lien avec la procédure d'échantillonnage. Il serait également intéressant de faire tourner

nos méthodes sur un modèle théorique complexe pour lequel les indices sont calculables analytiquement. Cette confrontation permettra de vérifier son efficacité et sa précision.

Annexe

Annexe 1 : Critère d'information pour les modèles linéaires

Un estimateur sans biais de $\mathbb{MSE}$ (voir Leeb, 2008 [92] pour la dérivation) est MC_n et s'écrit

$$MC_n = SCR + 2d\hat{\sigma} - n\hat{\sigma} \tag{5.4.1}$$

Remarquons qu'en divisant cette équation par $\hat{\sigma}$, nous retrouvons le critère de selection de modèle de Mallows (C_p) introduit en 1964 (Mallows, 1973 [98] ; Mallows, 1995 [99]). Le critère C_p de Mallows selon cette description vise à sélectionner le modèle qui minimise le $\mathbb{MSE}$.

Si l'objectif du chercheur est la prévision, le critère de sélection sera évidement le $\mathbb{MSEP}$. Un estimateur sans biais $nFPE$ du $\mathbb{MSEP}$ présenté dans Leeb (2008) [92] vaut :

$$nFPE = nSCR \times \frac{1 + \frac{d}{n}}{n - d} \tag{5.4.2}$$

où FPE désigne le critère Final Predictor Error (Akaike 1969 [1], 1970 [2]).
Notons que les critères FPE et C_p sont alors des dérivations respectives du $\mathbb{MSEP}$ et $\mathbb{MSE}$ dans le cas d'un modèle linéaire multiple.

Le critère d'information d'Akaike (AIC) (Akaike, 1973 [3] ; Akaike , (1974) [4]) largement utilisé cherche à minimiser l'espérance de l'écart entre la distribution du modèle candidat et celle du "vrai modèle" en utilisant la divergence de Kullback-Leiller comme fonction perte. Akaike (1973) [3] propose un estimateur approximativement non biaisé (AIC) (Leeb 2008 [92]) de cette mesure de risque.

$$AIC = log(\frac{SCR}{n}) + 2\frac{d}{n} \tag{5.4.3}$$

La dérivation formelle de ce critère se trouve dans Amemiya (1980) [7], Trembaly (2004) [151] Ce critère ne faisant intervenir que les distributions du modèle candidat et du « vrai modèle » s'applique aussi bien aux modèles linéaires que non linéaires. Les propriétés asymptotiques et non asymptotiques (convegence en probabilité) du critère AIC sont étudiées dans Shibata (1981) [139], Nishii (1984)[108], Mcquarrie et Tsai (1998) [103]. Hurvich et Tsai (1989) [74] et Bedrick et Tsai (1994) [19] proposent une version corrigée du critère AICc pour prendre en compte les petites tailles d'échantillon par rapport à la dimension du modèle.

$$AICc = log(\frac{SCR}{n}) + 2\frac{d+1}{n-k-2} \tag{5.4.4}$$

L'équivalent de ce critère dans la statistique bayésienne est le critère d'information de Schwartz BIC (Schwartz, 1978 [134] ; Cavanaugh et Neith, 1997[36] ; Cavanaugh et Neith, 1999 [35] ; Haughton, 1991 [65] ; Pauler, 1998 [111] ; Mcquarrie 1999 [102]). Ce critère cherche à maximiser la vraisemblance à posteriori du modèle candidat et la dérivation de cet estimateur se trouve Tremblay (2004) [151].

$$BIC = log(\frac{SCR}{n}) + d\frac{log(n)}{n} \tag{5.4.5}$$

Diverses critères furent considérés dans Leeb (2008) [92] et les références mentionnées dans cet article notamment TIC, GIC, et dans Robert (2006) [121] pour DIC, D'autres techniques de sélection de modèles tels que le stepwise (« backward step », « forward step ») sont présentées dans Muller (2002) [107], Draper et Smith (1981) [46], Hocking (1976) [67].

Les relations et les équivalences entre les différents critères de sélection sont discutées dans Shao (1997) [138], Yang (2007) [161], Shibata (1989) [140], Stone (1977) [145], Leeb (2008) [92], Amemiya (1980) [7]. Une étude comparative par simulation et sur les données réelles des critères $\mathbb{CV}$,AIC, AIC_c,BIC, BIC_c est faite dans Tremblay (2004) [151]. Les critères AIC, AIC_c, FPE, C_p et $\mathbb{CV}$ sont asymptotiquement équivalents.

Annexe 2 : procédure de sélection LARS (LASSO)

Cette procédure de sélection est spécifique aux modèles linéaires (Seber, 1977 [135]), d'équation :

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{5.4.6}$$

où $\mathbf{y}$ est le vecteur d'observations et $\mathbf{X}$ est la matrice des variables explicatives, β est le vecteur de paramètres à estimer et ε représente le terme d'erreur qui modélise l'inadéquation entre les observations et le modèle.

Considérons la famille des régressions pénalisées définie par la contrainte $||\beta||_p < C$ avec C une constante, $p \geq 1$ et $||\bullet||_p$, désigne la norme L_p. Notons qu'à chaque norme L_p ($||\bullet||_p$, $\quad p > 0$) est associée une pénalité. Cette famille de régression connue sous le nom de "bridge regression" (Frank and Friedman, 1993 [53]) consiste à estimer le vecteur de paramètres β par :

$$\hat{\beta}_{bridge} = argmin_\beta \, ||\mathbf{y} - \mathbf{X}\beta||_2 + \lambda \, ||\beta||_p \tag{5.4.7}$$

Les pénalités les plus fréquentes sont la pénalité L_1 ou Lasso ($p = 1$) (Tibshirani, 1996) [150], la pénalité L_2 ou la "ridge regression" ($p = 2$) (Hoerl and Kennard, 1970 [69]). Des études comparatives de ces différentes procédures de régression sont présentées dans Tibshirani, 1996) [150], Fu (1998) [55] et montrent les performances de LASSO. Ces différentes méthodes rétrécissent les coefficients de régression mais seule LASSO (Least Absolute Shrinkage and Selection Operator) a les propriétés de parcimonie dans le sens où il rétrécit certains coefficients et fixent d'autres à 0. Ainsi, il contribue à la fois à la sélection de variables et à la "ridge regression" des variables sélectionnées et fournit un modèle sparse. Il y a plusieurs algorithmes d'estimation des paramètres par la méthode de LASSO (Zou *et al.*, 2005 [163] ; Efron *et al.*, 2004 [47] ; Tibshirani, 1996 [150].

L'algorithme d'estimation des paramètres β proposé par Efron *et al.* [47] pour le LASSO est une adaptation de l'algorithme LARS (Least Angle Regression Stepwise). L'idée principale de l'algorithme repose sur le principe suivant : on cherche à maximiser à chaque étape de l'algorithme, la corrélation entre les variables explicatives et le vecteur de résidus du modèle tout en se déplaçant dans une direction $\mathbf{w}$ qui assure l'équi-corrélation entre les variables explicatives sélectionnées jusqu'à présent (variables actives notées A). L'algorithme est le suivant :

Algorithme 5.4.1 (Efron 2004) *[47] Adaptation du LARS au LASSO*

Etape 1 : *initialisation :*
$A \leftarrow \emptyset, \quad \hat{\beta}_0(A) \leftarrow 0_{\mathbb{R}^d}, \quad \mathbf{X}_A \leftarrow (\ldots s_j \mathbf{x}_j \ldots)_{j \in A},$
$\hat{\mathbf{y}} \leftarrow 0, \quad A \cup A^c = \{1, 2, \ldots, d\},$

Etape 2 : *recherche de la variable à ajouter qui maximise la corrélation*
$\hat{j} = argmax_{j \in A^c} \mid \mathbf{x}_j'(\mathbf{y} - \hat{\mathbf{y}}) \mid, \quad \hat{c}_j = \mathbf{x}_{\hat{j}}'(\mathbf{y} - \hat{\mathbf{y}})$
$s_j \leftarrow sign(\hat{c}_j)$

Etape 3 : *mise à jour :*
$A \leftarrow A \cup \{j\}, \quad \mathbf{X}_A \leftarrow \mathbf{X}_A \cup s_j \mathbf{x}_{\hat{j}}$

Etape 4 : *recherche de la direction* $\mathbf{w}_A$ *d'équi-corrélation*
$\mathbf{w}_A = [1\!\!1_A (\mathbf{X}_A' \mathbf{X}_A)^{-1} 1\!\!1_A]^{-1/2} [\mathbf{X}_A' \mathbf{X}_A]^{-1} 1\!\!1_A$
le vecteur d'équi-angle est $\mathbf{u}_A = \mathbf{X}_A' \mathbf{w}_A$ *et on a*
$\mathbf{X}_A' \mathbf{u}_A = a 1\!\!1_A, \quad \mathbf{X}_A'(\mathbf{y} - \hat{\mathbf{y}}) = c 1\!\!1_A$

Etape 5 : *pas de descente : pas optimale pour qu'une variable intègre A*
$\hat{\gamma} = \min^+_{j \in A^c} \left(\frac{c - c_j}{1 - a_j}, \frac{c + a_j}{1 + a_j} \right)$

Etape 6 : *mise a jour de l'estimation :*
si $\forall j \in A$, $\beta_k[j] \beta_{k+1}[j] < 0$ *alors :*
posons $A \leftarrow A \setminus \{j\}, \quad \gamma_j = -\frac{\beta_k[j]}{s_j \mathbf{w}_A[j]}$
$\tilde{\gamma} \leftarrow \min_{\gamma_j > 0} \{\gamma_j\}$
finsi
si $\tilde{\gamma} < \hat{\gamma}$ *alors*
$\hat{\gamma} \leftarrow \tilde{\gamma}$
finsi
$\beta_{k+1}[j] \leftarrow \beta_k[j] + s_j \hat{\gamma} \mathbf{w}_A[j]$

Etape 7 : *poser* $\hat{\mathbf{y}}_{k+1} \leftarrow \mathbf{X}' \hat{\beta}_{k+1}$ *; reprendre les étapes 2 à 6 et utiliser le critère* C_p *de Mallows comme critère d'arrêt.*

La démonstration du calcul de la direction $\mathbf{w}$, du pas optimal de descente $\hat{\gamma}$, de la convergence de l'algorithme et des propriétés théoriques se trouvent dans Efron (2004) [47]. L'avantage de cet algorithme est qu'il est efficace dans le sens où il intègre une variable à chaque itération et il fournit toutes les solutions de LASSO correspondantes aux différentes contraintes (C) sur le vecteur des paramètres β. Il est également possible de faire les inférences sur β. Tibshirani (1996) [150], Efron *et al.* (2004) [47] montrent que les estimateurs $\hat{\beta}$ sont des fonctions linéaires par morceau en fonction du paramètre γ.

Les limitations de la méthode LASSO furent soulignées dans Zou *et al.* (2005) [163].

En effet, Zou *et al.* (2005) [163] montrent que LASSO est limité dans le cas où le nombre d'observation n est très peu petit par rapport à la dimension d du modèle (microarray $p \gg n$) à l'aide des données génétiques. Dans cette configuration, lasso ne peut sélectionner qu'au plus n variables. Zou *et al.* (2005) [163] proposent de combiner la pénalité L_1 et L_2 pour prendre en compte les insuffisances de LASSO. Cette nouvelle méthode de régression connue sous le nom de "Elastic Net" permet la sélection de variables et gère mieux les corrélations importantes entre les variables. La régression Elastic Net naïf s'écrit :

$$\hat{\beta}_{enn} = argmin_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2 , \tag{5.4.8}$$

ou de manière équivalente en posant $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$

$$\hat{\beta}_{enn} = argmin_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2 + \alpha ||\beta||_2 + (1-\alpha) ||\beta||_1 , \tag{5.4.9}$$

où λ_1,, λ_2 sont des réels positifs. Elastic Net naïf minimise SCR sous contrainte de $J(\beta) = \alpha ||\beta||_2 + (1-\alpha) ||\beta||_1$. Remarquons que si $\alpha = 1$ Elastic Net naif devient la "ridge regression" et que si $\alpha = 0$, on retombe sur LASSO. La pénalité l_1 (LASSO) génère un modèle sparse et la pénalité L_2 remédie la limitation du nombre de variable à sélection (ridge régression) et stabilise la régularisation. Le paramètre α joue un rôle de compromis entre les deux pénalités.

L'estimateur Elastic Net $\hat{\beta}_{en}$ de β est la version normalisée de celui de Elastic Net naïf $\hat{\beta}_{enn}$ (Zou *et al.*, 2005 [163]) et on a :

$$\hat{\beta}_{en} = (1 + \lambda_2)\hat{\beta}_{enn} \tag{5.4.10}$$

Cette normalisation permet d'éviter à Elastic Net naïf de se rapprocher le plus à la "ridge regression" ou à LASSO et le théorème suivant (Zou *et al.*, 2005 [163]) donne les estimateurs directes de β par la méthode Elastic Net $\hat{\beta}_{en}$ et LASSO $\hat{\beta}_{lasso}$.

Théorème 5.4.1 (Zou 05) *[163]*
Etant données les observations $(\mathbf{y}, \mathbf{X})$ *alors :*

$$\hat{\beta}_{en} = argmin_{\beta}\beta' \left(\frac{\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}}{1 + \lambda_2} \right) \beta - 2\mathbf{y}'\mathbf{X}\beta + \lambda_1 ||\beta||_1 . \tag{5.4.11}$$

$$\hat{\beta}_{lasso} = argmin_{\beta}\beta' (\mathbf{X}'\mathbf{X}) \beta - 2\mathbf{y}'\mathbf{X}\beta + \lambda_1 ||\beta||_1 . \tag{5.4.12}$$

La preuve de ce théorème se trouve dans Zou *et al.* (2005) [163].
□

Bibliographie

[1] Akaike, H. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics 21* (1969), 243–247.

[2] Akaike, H. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics 22* (1970), 203–217.

[3] Akaike, H. Information theory and an extension of the maximun likelihood principle. *International Symposium on Information Theory 2nd* (1973), 267–281.

[4] Akaike, H. A new look at the statistical model identification. *IEEE Transaction On Automatic Control 19* (1974), 716–723.

[5] Allen, D. Mean square error of prediction as a criterion for selecting variables. *Technometrics 13 (3)* (1971), 469–475.

[6] Allen, D. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics 16* (1974), 125–127.

[7] Amemiya, T. Selection of regressors. *International Economic Review 21* (1980), 331–354.

[8] Anderson, T. W. The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics 33* (1962), 255–265.

[9] Anderson, T. W. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics 34* (1963), 122–148.

[10] Anderson, T. W. Determination of the order of dependence in normally distributed time series. In *Time Series Analysis*, M. Rosenblatt, Ed. Wiley & Sons, New York., 1963, pp. 425–446.

[11] Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, New York, 2003.

[12] Antoniadis, A. Analysis of variance on function spaces. *Journal of Theoretical and Applied Statistics 15* (1984), 59–71.

[13] Archer, G., Saltelli, A., and Sobol, I. Sensitivity measures, anova-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation 58* (1997), 99–120.

[14] Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society 68* (1950), 337–404.

[15] Azaïs, J. M., and M., B. J. *Le Modèle Linéaire par L'exemple : Régression, Analyse de la Variance et Plans d'Expériences Illustrés par R, SAS et Splus*. Dunod, Paris, 2005.

[16] B., B. M. Water quality modeling : A review of the analysis of uncertainty. *Water Ressource Research 23* (1987), 1393–1442.

[17] Baret, D., and Guyot, D. Potentials and limits of vegetation indices of lai and apar assessment. *Remote sensing of Environment 33* (1991), 161–173.

[18] Bechini, L., Bocchi, S., Maggiore, T., and Confalonieri, R. Parameterization of a crop growth and development simulation model at sub model component level. an example for winter wheat (triticum aestivum l.). *Environmental Modelling & Software 21* (2006), 1042–1054.

[19] Bedrick, E. J., and Tsai, C. Model selection for multivariate regression in small samples. *Biometrics 50* (1994), 226–231.

[20] Berlinet, A., and Thomas-Agnan, C. *Reproducing Kernel Hilbert Space in Probability and Statistics*. Kluwer Academic, 2004.

[21] Besse, P. Pca stability and choice of dimensionality. *Statistics & Probability Letters 13* (1992), 405–410.

[22] Beven, K. Future of distributed modeling. *Hydrology Processes 6* (1992), 253–254.

[23] Boote, K., Jones, J., and Pickering, N. Potential uses and limitations of crop models. *Agronomy Journal 88* (1996), 704–716.

[24] Bosq, D. *Linear Process in Function Space : Theory and applications*, vol. 49 of *Lecture Notes in Statistic*. Springer, 2000.

[25] Box, G., and Draper, N. *Empirical Model Building and Response Surfaces*. Wiley and Sons, , New York, 1987.

[26] Brisson, N., Mary, B., Ripoche, D., Jeuffroy, D., Ruget, F., Nicoullaud, B., Gate, P., Devienne, B., Antonioletto, R., Durr, c., Richard, G., Beaudoin, N., Recous, S., Tayot, X., Plenet, D., Celier, P., Machet, J., Meynard, J., and Delecolle, R. Stics : a generic model for the simulation of crops and their water and niitrogen balances. theory and parameterization applied to wheat and corn. *Agronomie 18* (1998), 311–346.

[27] Brun, R., Kuhni, M., Siegrist, H., Gujer, W., and Reichert, P. Practical identifiability of asm2d parameters-systematic selection and tuning of parameter subsets. *Water Research 36* (2002), 4113–4127.

[28] Brun, R., Reichert, P., and Kunsch H., R. Practical identifiability of large environmental simulation models. *Water Resources Research 37* (2001), 1015–1030.

[29] Butterbach-Bahl, K., Kesik, M., Miehle, P., Papen, H., and Li, C. Quantifying the regional source strength of n-trace gases across agricultural and forest ecosystems with process based models. *Plant and Soil 260* (2004), 311–329.

[30] Cai, Z., Fan, J., and Yao, Q. Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95* (2000).

[31] Campbell, K., McKay, D., and Williams, P. Sensitivity analysis when model outputs are functions. *Reliability Engineering and System Safety 91* (2006), 1468–1472.

[32] Campolongo, F., Cariboni, J., and Saltelli, A. An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software 22* (2007), 1509–1518.

[33] Carnell, R. *Latin Hypercube Samples. Package "lhs", Version 0.5.* CRAN Repository, 2008.

[34] Castillo, E., Sánchez-Maroño, N., Alonso-Betanzos, A., and Castillo, C. Functional network topology learning and sensitivity analysis based on anova decomposition. *Neural Computation 19* (2007), 231–257.

[35] Cavanaugh, J., and Neath, A. Generalizing the derivation of the schwarz information criterion. *Communication in Statistics-Theory and Methods 28* (1997), 49–66.

[36] Cavanaugh, J., and Neath, A. Regression and time series model selection using variants of the schwarz information criterion. *Communication in Statistics-Theory and Methods 26* (1997), 559–580.

[37] Chatterjee, A. An introduction to the proper orthogonal decomposition. *Computational science 78* (2000).

[38] Colbach, N., Clermont, D. C., and Meynard, J. M. Genesys : a model of the influenece of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. temporal evolution of a population of rapeseeed volunteers in a field. *Agriculture, Ecosystem & Environment 83* (2001), 235–253.

[39] Constales, D., and Kacur, J. Computation and sensitivity analysis of the pricing of american call options. *Applied Mathematics and Computation* (2006), 303–307.

[40] Cukier, R., Fortuin, C., Shuler, K., Petschek, A., and Schaibly, J. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients-theory. *Journal Chemical Physics 59* (1973), 3837–3878.

[41] Cukier, R., Levine, R., and Shuler, K. Nonlinear sensitivity analysis of multiparameter model systems. *Journal Computational Physics 26* (1978), 1–42.

[42] Cukier, R., Shuler, K., and Schaibly, J. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients-analysis of the approximations. *Journal Chemical Physics 63* (1975), 1140–1149.

[43] Da-Veiga, S. *Analyse d'incertitudes et de sensibilité : application aux modèles de cinétique chimique*. PhD thesis, Université de Toulouse III, 2007.

[44] Dauxois, J., Pousse, A., and Romain, Y. Asymptotic theory for the pca of a vector random function : some applications to statistical inference. *Journal of Multivariate Analysis 12* (1982), 136–154.

[45] Dong, D., and McAvoy, T. Nonlinear principal component analysis-based on principal curves and neural networks. *American Control Conference 2* (1994), 1284–1288.

[46] Draper, N. R., and Smith, H. *Applied Regression Analysis*, 2nd ed. John Wiley, New York, 1981.

[47] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Annals of Statistics 32* (2004), 407–451.

[48] Efron, B., and Stein, C. The jackknife estimate of variance. *The Annals of Statistics 9* (1981), 586–596.

[49] Escoufier, Y. La dépendance de deux aléas vectoriels : critères et visualisation. *Revue de la Statistique Appliquée 21* (1973), 5–16.

[50] Fang, K. T., Li, R., and Sudjianto, A. *Design and Modelling for computer Experiments*. Computer science and Data Analysis. Chapman and Hall,Taylor and Francis Group, 2006.

[51] Fisher, R., and Yates, F. The 6×6 latin squares. *Mathematical Proceedings of the Cambridge Philosophical Society 30* (1934).

[52] Fisher, R. A. *Statistical Methods for Research Workers*, 13th ed.

[53] Frank, I., and Friedman, J. H. A statistical view of some chemometrics regression tools (with discussion). *Technometrics 35* (1993), 109–148.

[54] Frontier, S. Etude de la décroissance des valeurs propres dans une analyse en composante principale : comparaison avec le modèle de bâton brisé. *Journal of Experimental Marine Biology and Ecology 25* (1976), 67–75.

[55] Fu, W. Penalized regressions : the bridge versus the lasso. *Journal of Computational and Graphical Statistics 7* (1998), 397–416.

[56] Furbinger, J. M. Sensitivity analysis for modellers. *Air Infiltration Review 17* (1996).

[57] Gabrielle, B., Laville, P., Duval, O., Nicoullaud, B., Germon, J. C., and Hénault, C. Process-based modeling of nitrous oxyde emissions from wheat-cropped soils at the sub-regional scale. *Global Biogeochemical Cycles 20* (2006), X.1–x.13.

[58] Gabrielle, B., Laville, P., Hénault, C., Nicoullaud, B., and Germon, J. Simulation of nitrous oxide emissions from wheat-cropped soils using ceres. *Nutrient Cycling Agroecosystem* (2006), 133–146.

[59] Gallant, A. *Nonlinear Statistical Models*. Wiley. John Wiley, 1987.

[60] George, E., and Foster, D. Calibration and empirical bayes variable selection. *Biometrika 77* (2000), 731–747.

[61] Ginot, V., Gaba, S., Beaudouin, R., Aries, F., and Monod, H. Combined use of local and anova-based global sensitivity analyses for investigation of a stochastic dynamic model : application to the case study of an individual-based model of a fish population. *Ecological Modelling 193* (2006), 479–491.

[62] Goldstein, M., and Rougier, J. Bayes linear calibrated prediction for complex systems. *Journal of American Statistical Association 101* (2006), 1132–1143.

[63] Hall, P., Muller, H. G., and Wang, J. L. Properties of functional principal component methods for functional and longitudinal data analysis. *The Annals of Statistics 34* (2006), 1493–1517.

[64] Hall, P., and Nasab, M. On properties of functional principal components analysis. *Journal of Royal Statistical Society B 68* (2006), 109–126.

[65] Haughton, D. Consistency of a class of information criteria for model selection in nonlinear regression. *Communication in Statistics-Theory and Methods 20* (1991), 1619–1629.

[66] Hénault, C., and Germon, J. Nemis, a predictive model of denitrification on the field scale. *European Journal of Soil Science 51* (2000), 257–270.

[67] Hocking, R. The analysis and selection of variables in linear regression. *Biometrics 32* (1-49).

[68] Hoeffding, W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics 19* (1998), 293–325.

[69] Hoerl, A., and Kennard, R. Ridge regression based estimation for nonorthogonal problems. *Technometrics 12* (1970), 55–67.

[70] Homma, T., and Saltelli, A. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety 52* (1996), 1–17.

[71] Hooker, G. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics 16* (2007), 709–732.

[72] Hotelling, H. Relations between two sets of variates. *Biometrica 28* (1936), 321–377.

[73] Huang, J. Projection estimation in multiple regression with application to functional anova model. *The Annals of Statistics 26* (1998), 242–272.

[74] Hurvich, C. M., and Tsai, C. Regression and times series model selection in small samples. *Biometrika 76* (1989), 297–307.

[75] Hyndman, R., and Fan, Y. Sample quantiles in statistical packages. *American Statistician 50* (1996), 361–365.

[76] Indritz, J. *Methods in Analysis*. Macmillan, New York, 1963.

[77] Jackerman, A. J., and Hornberger, G. M. How many complexity is warranted in rainfail-rainoff model. *Water Ressource Research 29* (1993), 2637–2649.

[78] Jackson, D. Stopping rules in principal components analysis : A comparison of heuristical and statistical approaches. *Ecology 74* (1993), 2204–2214.

[79] Jackson, J. E. *A User's Guide to Principal Components*. John Wiley and Sons, New York, USA, 1991.

[80] Jacques, J. *Contribution à l'analyse de sensibilité et à l'analyse discriminate généralisée*. PhD thesis, Université Joseph Fourier, 2005.

[81] Jeuffroy, M., and Recous, S. Azodyn : a simple model for simulating the date of nitrogen defiency for decision support in nitrogen fertilization. *European Journal of Agronomy 10* (1999), 129–144.

[82] Jolliffe, I. *Principal Component Analysis*, 2nd ed. Springer, New York, USA, 2002.

[83] Jolliffe, I. T. *Principal Component Analysis*. Springer-Verlag, New York, USA, 1986.

[84] Jones, C., and Kiniry, J. Ceres-n maize : A simulation model of maize growth and development. *Texas A&M University Press, College Statio, Temple, TX* (1986).

[85] Kanso, A., Chebbo, G., and Tasssin, B. Application of mcmc-gsa model calibration method to urban runoff quality modeling. *Reliability Engineering & System Safety 91* (2006), 1398–1405.

[86] Kobilinsky, A. Les plans factoriels. In *Plans d'Expériences : Application à l'entreprise*, J. Droesbeke, J. Fine, and G. Saporta, Eds. Editions Technip, Paris, 1997, ch. 3, pp. 69–209.

[87] Kontoravdi, C., Asprey, S. P., Pistikopoulos, E. N., and Mantalaris, A. Application of global sensitivity analysis to determine goals for design of experiments : An example study on antibody-producing cell cultures. *Biotechnology Prog.* (2005).

[88] Kosambi, D. Statistics in function space. *Journal of Indian Mathematical Society 7* (1943), 76–88.

[89] Kunisch, K., and Volkwein, S. Galerkin proper orthogonal decomposition for parabolic problems. *Numerische mathematik 90* (2001), 117–148.

[90] Lamboni, M., Makowski, D., Lehuger, S., Gabrielle, B., and Monod, H. Multivariate global sensitivity analysis for dynamic crop models. *Journal Of Fields Crop Reasearch 113* (2009), 312–320.

[91] Lamboni, M., Makowski, D., and Monod, H. Multivariate global sensitivity analysis for discrete-time models. Rapport technique 2008-3, INRA, UR341 Mathématiques et Informatique Appliquées, Jouy-en-Josas, France, 2008.

[92] Leeb, H., and Potscher, B. *Model selection.* Handbook of Financial Time Series, 2008.

[93] Lehuger, S., Gabrielle, B., VanOijen, M., Makowski, D., Germon, J., Morvan, T., and Hénault, C. Bayesian calibration of the nitrous oxide emission module of an agro-ecosystem model. *Agriculture, Ecosystems & Environment 133* (2009), 208–222.

[94] LeMieux, C., and Owen, A. B. Quasi-regression and the relative importance of the anova components of a function. In *Monte Carlo and Quasi-Monte Carlo Methods*, K. T. Fang, F. J. Hickernell, and H. Niederreiter, Eds. 2000.

[95] Lurette, A., Touzeau, S., Lamboni, M., and Monod, H. Sensitivity analysis to identify key parameters influencing salmonella infection dynamics in a pig batch. *Journal of Theoretical Biology 258* (2009), 43–52.

[96] Makowski, D., Jeuffroy, M.-H., and Guérif, M. Bayesian methods for updating crop model predictions. applications for predicting biomass and grain protein content. In *Bayesian Statistics and Quality Modelling in the Agro-food Production*

Chain, M. Van Boekel, A. Stein, and A. H. C. van Bruggen, Eds. Kluwer, Dordrecht, 2004, pp. 57–68.

[97] Makowski, D., Naud, C., Jeuffroy, M., Barbottin, A., and Monod, H. Global sensitivity analysis for calculating the contribution of genetic parameters to the variance of crop model prediction. *Reliability Engineering and System Safety 91* (2006), 1142–1147.

[98] Mallows, C. Some comments on c_p. *Technometrics 15* (1973), 661–675.

[99] Mallows, C. More comments on c_p. *Technometrics 37* (1995), 362–372.

[100] Marrel, A. *Mise en œuvre et utilisation du metamodèle processus gaussien pour l'analyse de sensibilité*. PhD thesis, Institut National des Sciences Appliquées de Toulouse, 2008.

[101] McKay, M., Beckman, R., and Conover, W. A comparison of three methods of selecting values of input variables in the analysis of output from a computer code. *Technometrics 21* (1979), 239–245.

[102] McQuarrie, A. A small sample correction for the schwarz bic model selection criterion. *Statistics and Probability Letters 44* (1999), 79–86.

[103] McQuarrie, A., and Tsai, C.-L. Regression and time series model selection. *World Scientific Publication Co* (1998).

[104] Meynard, J., Cerf, M., Guichaud, L., Jeuffroy, M. H., and Makowski, D. Which decision support tools for the environment management of nitrogen ? *Agronomie 22* (2002), 817–829.

[105] Monod, H., Naud, C., and Makowski, D. Uncertainty and sensitivity analysis for crop models. In *Working with Dynamic Crop Models*, D. Wallach, D. Makowski, and J. Jones, Eds. Elsevier, Amsterdam, 2006, pp. 55–100.

[106] Morris, M. Factorial sampling plans for preliminary computational experiments. *Technometrics 33* (1991), 161–174.

[107] Muller, A. *Subset Selection in Regression*, 2nd ed. Chapman et Hall, 2002.

[108] Nishii, R. Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics 12* (1984), 758–765.

[109] Oreskes, N., Shrader, F., and Belitz, K. Verification, validation and confirmation of numerical models in the earth science. *Science 236* (1994), 641–646.

[110] Passo, A., Esposito, A., Porcu, E., Reverberi, A. P., and F., V. Statiscal sensitivity analysis of packed column reactors for contaminated wastewater. *Environmetrics 14* (2003), 743–759.

[111] Pauler, D. K. The schwartz criterion and related methods for normal linear models. *Biometrika 85* (1998), 13–27.

[112] Pearson, K. Lines and planes of closest fil to systems of points in space. *Philosophical Magazine* (1901), 559–572.

[113] Peres-Neto, P., Jackson, D., and Somers, K. How many principal components ? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis 49* (2005), 974–997.

[114] Perrin, C., Michel, C., and Andreassian, V. Does a large number of parameters enhance model performance ? comparative assessment of common catchment model structure on 429 catchments. *Journal of Hydrology 242* (2001), 275–301.

[115] Pujol, G. *Sensitivity Analysis. Package "sensitivity", Version 1.4-0.* CRAN Repository, 2008.

[116] R Development Core Team. *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008.

[117] Rabitz, H. System analysis at molecular scale. *Science 246* (1989), 221–226.

[118] Ramsay, J., and Silverman, B. *Functional Data Analysis.* Statistics. Springer-Verlag, 1997.

[119] Ramsay, J., and Silverman, B. *Applied Functional Data Analysis : Methods and Case Studies.* Springer Series in Statistics. Springer-Verlag, 2002.

[120] Ratto, M., Tarantola, S., and Saltelli, A. Sensitivity analysis in model calibration : Gse-glue approach. *Computer Physics Communications 136* (2001), 212–224.

[121] Robert, C. P. *Le choix bayésien, Principes et pratique.* Springer. Statistique et probabilités appliquées., 2006.

[122] Ruget, F., Brisson, N., Delécolle, R., and Faivre, R. Sensitivity analysis of the crop simulation model stics in order to choose the main parameters to be estimated. *Agronomie 22* (2002), 133–158.

[123] Ruskai, M., Beylkin, G., Coifman, R., Daubechies, I., Mallat, S., Meyer, Y., and Raphael, L. *Wavelets and Their Applications.* Collection Jones and Bartlett books in mathematics. ones & Bartlett, 1992.

[124] Saltelli, A. Making best use of model evaluations to compute senstivity indices. *Computer Physics Communication 145* (2002), 280–297.

[125] Saltelli, A. Sensitivity analysis for importance assessment. *Risk Analysis 22* (2002), 579–590.

[126] Saltelli, A., Chan, K., and Scott, E. *Variance-Based Methods*. Probability and Statistics. John Wiley and Sons, 2000. CHAPTER 8.

[127] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. *Global Sensitivity Analysis : The Primer*. Wiley, 2008.

[128] Saltelli, A., Tarantola, A., Campolongo, F., and Ratto, M. *Sensitivity Analysis in Practice*. Wiley, New York, 2004.

[129] Saltelli, A., Tarantola, S., and Campolongo, F. Sensitivity analysis as an ingredient of modelling. *Statistical Science 15* (2000), 377–395.

[130] Saltelli, A., Tarantola, S., and Chan, K.-S. A quantitative model-independent methods for global sensitivity analysis of model output. *Technometrics 41* (1999), 39–56.

[131] Saporta, G. *Probabilité, Analyse des Données et Statistique*, 2nd ed. Technip, 2006.

[132] SAS Institute Inc. *SAS/QC User's Guide*. SAS Institute Inc., Cary, NC., 2008.

[133] Schumaker, L. *Spline Functions Basic Theory*. Collection Pure and applied mathematics. John Wiley, 1981.

[134] Schwarz, G. Estimating the dimension of model. *The Annals of statistics 6* (1978), 461–464.

[135] Seber, G. *Linear Regression Analysis*. Wiley. John Wiley N. Y., 1977.

[136] Seber, G., and Wild, C. *Nonlinear Regression*. John Wiley, N. Y., 1989.

[137] Shao, J. Linear model selection by cross validation. *Journal of American Statistical Association 88* (1993), 486–494.

[138] Shao, J. An asymptotic theory for linear model selection (with discussion). *Statistical Sinica 7* (1997), 221–264.

[139] Shibata, R. An optimal selection of regression variables. *Biometrika 68* (1981), 45–51.

[140] Shibata, R. Statisticals aspect of model selection. In *From Data to Model*, J. Willems, Ed., Springer. Springer, New York, 1989, pp. 215–240.

[141] Sievanen, R., and T.E., B. Adjusting a process-based growth model for varying site conditions through parameter estimation. *Canadian Journal Forrest Research 23* (1993), 1837–1851.

[142] Sobol, I. Sensitivity analysis for non-linear mathematical model. *Mathematical Modelling and Computational Experiments 1* (1993), 407–414.

[143] Stone, C. J. The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics 22* (1994), 179–184.

[144] STONE, M. Cross-validatory choice and assessment of statistical predictions (with discussion). *Royal Statistical Society 36* (1974), 111–147.

[145] STONE, M. An asympototic equivalence of choice of model by cross validation and akaike's criterion. *Royal Statistical Society 39* (1977), 44–47.

[146] TAO, J. *Data Driven Shrinkage Strategies for Quasi Regrssion*. PhD thesis, Department of statistics, Stanford University, 2003.

[147] TAO, J., and Owen, A. B. Quasi-regression with shrinkage. *Mathematics and Computers in Simulation 62* (2003), 231–241.

[148] Tarantola, S., Gatelli, D., and Mara, T. Random balance design for the estimation of first order global sensitivity indices. *Reliability Engineering and System Safety 91* (2006), 717–727.

[149] Theil, H. *Economic Forecasts and Policy*, 2nd ed. North Holland, Amsterdam, 1961.

[150] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society 58* (1996), 267–288.

[151] Tremblay, M. *Estimation des paramètres des modèles de culture*. PhD thesis, Université Paul Sabatier, 2004.

[152] Volkwein, S. Model reduction using proper orthogonal decomposition. Technical report, 2008.

[153] Wall, M., Rechsteiner, A., and Rocha, L. Singular value decomposition and principal component analysis : in a pratical approach to microarray data analysis, 2003.

[154] Wallach, D., and Génard, M. Effect of uncertainty in input and parameter values on model predictor error. *Ecological Modeling 105* (1998), 337–345.

[155] Wallach, D., and Goffinet, B. Mean square error of prediction in models for studying ecological and agronomic systems. *Biometrics 43* (1987), 561–573.

[156] Wallach, D., Goffinet, B., Bergez, J., Debaeke, P., Leenhardt, D., and Aubertot, J. Parameter estimation for crop models : a new approach and application to a corn model. *Agronomy Journal 93* (2001), 757–766.

[157] Wallach, D., Goffinet, B., Bergez, J., Debaeke, P., Leenhardt, D., and Aubertot, J. The effect of parameter uncertainty on a model with adjusted parameters. *Agronomie 22* (2002), 159–170.

[158] Wallach, D., Makowski, D., and Jones, J. *Working with Dynamic Crop Models : Evaluation, Analysis, Parameterization and Application.* Elsevier, 2006.

[159] Wang, X., and Fang, K. The effective dimension and quasi-monte carlo integration. *Journal of Complexity 19* (2003), 101–124.

[160] Whittaker, A. *The performance of cross validation indices used to select among competing covariance structure models.* PhD thesis, The university of Texas at Austin, 2003.

[161] Yang, Y. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics 35* (2007), 2450–2473.

[162] Zebra, K. E., and Collins, J. P. Spatial heterogeneity and individual variation in diet of aquatic predator. *Ecology 73* (1992), 268–279.

[163] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B 67* (2005), 301–320.

[164] Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15* (2006), 265–286.

Liste des productions scientifiques

Publications

- **[1.] Sensitivity analysis to identify key parameters influencing Salmonella infection dynamics in a pig batch, 2009 (Journal of Theoretical Biology, volume 258, pp. 43-52)**
  Amandine Lurette, Suzanne Touzeau, Matieyendou Lamboni, Hervé Monod

- **[2.] Multivariate global sensitivity analysis for dynamic crop models, 2009 (Field Crops Research, volume 113, pp. 312-320)**
  Matieyendou Lamboni, David Makowski, Simon Lehuger, Benoit Gabrielle and Hervé Monod

- **[3.] Predicting and mitigating the global warming potential of agro-ecosystems (soumis en septembre 2009)**
  S. Lehuger, B. Gabrielle, P. Laville , M. Lamboni, P. Cellier , B. Loubet

- **[4.] Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models (soumis en octobre 2009 dans le journal Reliability Engineering & System Safety)**
  Matieyendou Lamboni, Hervé Monod, David Makowski

- **[5.] Multivariate global Sensitivity Analysis for discrete-time models (Rapport Technique 2008-3, INRA, Unité MIA Jouy-en-josas)**
  Matieyendou Lamboni, David Makowski, Hervé Monod

Conférences et séminaires

- International conference on Sensitivity Analysis of Model Outputs (SAMO), Budapest, Hongrie, 18-22 juin 2007 (poster)

- 39ième journées de la Société Française Des Statistiques (SFDS) Anger, France, 11-15 juin 2007 (Exposé)

- Groupe de Recherche GDR Mascot Num, CEA Cadarache, France, 12-14 mars 2008 (poster)

- 2nd International Biometrics Society (IBS) Channel Network Conference Ghent, Belgium, 6-8 April 2009 (Exposé)

Enseignements

- **2007-2008 :** TD mathématique en L1 MPI à l'IUT de Créteil (Université Marne la vallée Paris XII) : algèbre linéaire (espace vectoriel, systèmes d'équation, calcul matricielle), fonction paramètrique et trigonométrie, suites numériques

- **2008-2009** : TD mathématique en L1 MPI à l'IUT de Créteil (Université Marne la vallée Paris XII) : logique, algèbre linéaire, fonction d'une variable, continuité et dérivabilité, équation différentielle, calcul d'intégrale, développement limité.

Résumé

Des modèles dynamiques sont souvent utilisés pour simuler l'impact des pratiques agricoles et parfois pour tester des règles de décision. Ces modèles incluent de nombreux paramètres incertains et il est parfois difficile voire impossible de tous les estimer. Une pratique courante dans la littérature consiste à sélectionner les paramètres clés à l'aide d'indices de sensibilité calculés par simulation et de n'estimer que les paramètres les plus influents. Bien que cette démarche soit intuitive, son intérêt réel et ses conséquences sur la qualité prédictive des modèles ne sont pas connus. Nos travaux de recherches ont pour ambition d'évaluer cette pratique des modélisateurs en établissant une relation entre les indices de sensibilité des paramètres d'un modèle et des critères d'évaluation de modèles tels que le $\mathbb{MSEP}$ (Mean Square Error of Prediction) et le $\mathbb{MSE}$ (Mean Square Error), souvent utilisés en agronomie. L'établissement d'une telle relation nécessite le développement d'une méthode d'AS qui fournit un unique indice par facteur qui prend en compte les corrélations entre les différentes sorties du modèle obtenues à différentes dates. Nous proposons un nouvel indice de sensibilité global qui permet de synthétiser les effets des facteurs incertains sur l'ensemble des dynamiques simulées à l'aide de modèle. Plusieurs méthodes sont présentées dans ce mémoire pour calculer ces nouveaux indices. Les performances de ces méthodes sont évaluées pour deux modèles agronomiques dynamiques : Azodyn et WWDM. Nous établissons également dans ce mémoire, une relation formelle entre le $\mathbb{MSE}$, le $\mathbb{MSEP}$ et les indices de sensibilité dans le cas d'un modèle linéaire et une relation empirique entre le $\mathbb{MSEP}$ et les indices dans le cas du modèle dynamique non linéaire CERES-EGC. Ces relations montrent que la sélection de paramètres à l'aide d'indices de sensibilité n'améliore les performances des modèles que sous certaines conditions.

Mots clés : Analyse de Sensibilité (AS) ; ACP ; Décomposition de Karhunen Loève ; Décomposition de l'inertie ; Estimation des paramètres ; Inertie ; $\mathbb{MSEP}$; Modèle de culture ; Modèle agro-environnemental ; Modèle dynamique ; Sélection de paramètres.

Abstract

Dynamic models are often used to simulate the impact of agricultural practices and sometimes to test some decision rules. These models include many uncertain parameters and it is sometimes difficult or impossible to estimate all the paramters. A common practice in literature is to select key parameters by using sensitivity index and then to estimate the most influent parameters. Although this approach is intuitive, his real interest and its consequences on the models predictive quality are not well known. Our research work aims to evaluate the practice of modellers by establishing a relationship between the sensitivity indices of model parameters and some model quality measures such as the *msep* (Mean Square Error of Prediction) and the $\mathbb{MSE}$ (Mean Square Error) often used in agronomy. Establishing such a relationship requires the development of a Sensitivity Analysis (SA) method that provides a unique index per factor and takes into account correlations between different model outputs. We propose a new sensitivity index that synthetizes the effects of uncertain factors on all the dynamic outputs obtained from dynamic models. Several methods are presented in this paper to calculate the new indices. The performance of these methods are evaluated on two agricultural dynamics models : Azodyn and WWDM. We also establish, in this paper, a formal relationship between $\mathbb{MSE}$, the $\mathbb{MSEP}$ and sensitivity indices in the case of a linear model and an empirical relationship between the $\mathbb{MSEP}$ and the new synthetic index in the case of a nonlinear dynamic model : CERES-EGC. These relations show that parameter selection by using sensitivity index improves models performance under some conditions.

Keywords : Agri-environment models ; Dynamic models ; Inertia ; Inertia expansion ; Karhunen Loeve expansion ; Parameter estimation ; Parameters selection ; PCA ; Sensitivity Analysis (SA) ; $\mathbb{MSE}$; $\mathbb{MSEP}$.