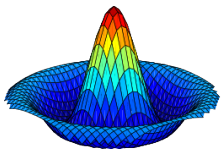
ECOLE-CHERCHEURS ANALYSE DE SENSIBILITE ET EXPLORATION DE MODELES

APPLICATION AUX SCIENCES DE LA NATURE ET DE L'ENVIRONNEMENT

du 7 au 11 juin 2010

Support de cours

par :
Claude Bruchou,
Robert Faivre,
Thierry Faure,
Stanie Mahevas,
David Makowski,
Hervnod



MEXICO

version du 14 mai 2010

Table des matières

Chapitre 1

Méthodes de criblage par discrétisation de l'espace

1.1 Introduction

1.1.1 Les modèles de simulation en questions

De nombreux domaines techniques et scientifiques (industrie aéronautique, nucléaire, économie, climatologie, agronomie,...) mettent en oeuvre des modèles de simulation numérique visant à mimer le comportement d'un système dit complexe. La modélisation du climat remonte à 1950 ([**?**]). La notion de complexité fait notamment référence au grand nombre de processus mis en jeu, à leurs interactions et à l'intégration de différents niveaux de perception ou échelles. Les enjeux sont cognitifs pour tester de nouveaux concepts. Par exemple, le concept d'émergence, rencontré notamment en épidémiologie, correspond à l'apparition de propriétés dans un système complexe. Quand l'analyse détaillée des processus pris isolément est suffisante pour induire les propriétés globales du système, on parle d'émergence faible. Quand on ne peut pas faire le lien (limite temporaire ou indépassable de la connaissance ?) entre les propriétés des processus et celles du système pris dans son ensemble, on parle d'émergence forte. Autre exemple, la résilience est la capacité d'un système (industriel, biologique,..) à reprendre son fonctionnement après un choc majeur. La démarche de modélisation permet d'élaborer et tester virtuellement des hypothèses. Elle permet de concrétiser les résultats d'expériences de pensée sous les hypothèses d'une représentation scientifique. Les enjeux de la modélisation sont aussi technico-scientifiques (système de culture, processus industriel,..). Le modèle génère potentiellement un ensemble de possibles qui ne sont pas forcément tous réellement possibles. Dans la mesure où il mime correctement le système réel, on espère remplacer des expérimentations difficiles ou coûteuses voire impossibles à mettre en oeuvre par des simulations. Faute de validations

conséquentes (l'exhaustivité est d'ailleurs impossible), un doute subsiste un tant soit peu sur le niveau de réalisme des simulations numériques. Il est clair qu'un modèle n'est pas équivalent à la réalité qu'il représente. Pourtant la tentation de le penser existe de façon plus ou moins consciente ([**?**]) du fait de l'empathie du modélisateur pour son modèle. Empathie qui résulte notamment des efforts consentis dans le processus d'élaboration mais aussi d'un certain consensus régnant au sein d'une communauté scientifique. Le consensus a t-il été établi par l'examen approfondi du code informatique et par l'analyse des résultats de simulations selon divers points de vue ? Les publications consacrées à un modèle ne donnent pas forcément le libre accés aux codes à l'ensemble de la communauté scientifique. On comprend certes que, pour certains enjeux, les secrets de fabrication soient à préserver. L'examen théorique du modèle est-il alors suffisant pour valider l'implémentation et la pertinence des sorties ? Dans quelle mesure des arguments d'autorité ne se substituent-ils pas de façon insidieuse à des arguments réfutables ? Certains consensus ou hypothèses introduits à un moment donné dans un modèle ne risquent-ils pas d'acquérir, par le jeu de leur empilement au sein d'une construction informatique imposante, le statut de connaissances acquises pour des utilisateurs moins bien informés des hypothèses sous-jacentes ? La question du suivi et des modifications des contenus du modèle au cours du temps se pose donc. Comment gérer la validation du modèle global suite à la modification (ou l'ajout) d'une de ses parties ? Lorsque le modèle quitte le laboratoire pour répondre à une demande de la société (concevoir des actions, prédire les conséquenses de décisions,...), le consensus au sein d'une communauté d'experts justifie l'utilisation. Face à un public peu averti, l'expert tenu de tenir un discours positif risque de s'exprimer de façon peu diserte sur les incertitudes du modèle si, de plus, il n'a pas utilisé les méthodes adéquates pour en mesurer l'effet.

Du fait de l'usage généralisé d'ordinateurs puissants on voit apparaître des situations où les modèles sont producteurs de réalités qui interagissent avec celles produites par d'autres modèles. L'usage de la modélisation en finance pourrait être emblématique de cette situation qui peut sembler déroutante au scientifique habitué à confronter une représentation rationnelle au réel supposé indépendant. Dans les sciences de l'environnement, le réel résiste d'une façon ou d'une autre à une conceptualisation liée à des objectifs (trop) humains. Dans certains contextes socio-économiques, la recherche d'une vérité objective indépendante des acteurs n'est pas forcément primordiale. Le modèle informatique est un moyen d'aboutir à un consensus entre acteurs. La modélisation est vue comme un "langage permettant aux acteurs d'exprimer leurs intérêts et de représenter leur vision des problèmes" ([**?**]). Les rationalités ou points de vue des différents acteurs se confrontent pour aboutir à des décisions acceptées par tous dans la mesure du possible. En résumé, la modélisation mixant un virtuel "hyper-réel", du fait du mimétisme, et des informations provenant du réel à un niveau de complexité ou de complication élevé grâce à la puissance de l'ordinateur, un nouveau rapport au monde serait à l'oeuvre. De plus, la recherche d'un consensus, et par là même l'acceptation d'une certaine subjectivité, remplacerait, mais c'est moins nouveau, la recherche d'une vérité qui transcende les acteurs.

Pour transposer dans ce contexte la célèbre gravure de Goya, les rêves (ou le sommeil) de la raison engendreront-ils des monstres (pas forcément malfaisants) aux comportements émergents surprenants ?

Nous avons tenté à partir de quelques interrogations tirées de la littérature et des pratiques de l'auteur de convaincre le lecteur, modélisateur ou utilisateur (mais qui peut aussi porter les deux casquettes), de la nécessité d'introduire une distance critique avec le modèle. Dans cet objectif, quel est le niveau d'analyse nécessaire pour élaborer cette critique et dans quels buts ? Avec quelles méthodes rationnellement discutées et admises faire cette analyse ? Pour le concepteur, l'analyse sera détaillée dans un objectif de recherche (validation du code, confrontation à des données issues d'observation,...). Pour l'utilisateur, l'analyse sera effectuée en fonction d'un objectif particulier (déterminer les facteurs à prendre en compte dans un scénario). La proposition de l'analyse de sensibilité globale est de construire un schéma ou synthèse interprétable du modèle. Elle peut devenir l'outil partagé par le modélisateur et l'utilisateur du modèle et permettre une véritable intégration du discours d'expertise basé sur la modélisation, dans la pratique démocratique. Un but de ce texte est de proposer quelques clés utiles pour la compréhension du modèle ou, plus modestement, une description statistique de son comportement basée sur l'étude des résultats de simulations.

1.1.2 Introduction à l'analyse de sensibilité

La première étape du travail de modélisation consiste à établir une représentation ou un schéma conceptuel qui agrège les connaissances établies ou hypothétiques sur des processus reliés entre eux. Le schéma conceptuel peut prendre la forme d'un diagramme représentant les processus par des boîtes avec leur entrées, leurs sorties et les échanges (flux de matière, informations,...) inter-processus. Nous supposerons que cette représentation théorique se formalise à l'aide d'une fonction mathématique $\mathcal{G}$ (pour usine à gaz) à valeurs dans $\mathbb{R}$. Une sortie d'intérêt du modèle théorique est représentée par le scalaire $\mathcal{G}(\mathbf{x})$ image d'un vecteur $\mathbf{x} = (x^{(1)}, .., x^{(K)})$, constitué de K facteurs d'intérêt du modèle. Cette représentation simplifiée a été choisie car elle est adaptée non seulement aux méthodes présentées dans ce chapitre mais aussi aux problématiques effectivement rencontrées en pratique. Nous n'aborderons pas ici la question de sorties qualitatives ou vectorielles. Nous supposerons que chaque facteur prend ses valeurs dans un intervalle borné préalablement défini. L'espace de définition Ω des facteurs sera assimilé à un hypercube de $\mathbb{R}^K$. En pratique, la définition des bornes de Ω n'est pas sans conséquence sur les résultats des analyses. Elle est en général argumentée par la littérature sur le sujet et résulte d'un consensus d'experts. On peut être amené à reconsidérer les hypothèses sur les bornes de variation des facteurs.

Il est nécessaire de réaliser une implémentation de $\mathcal{G}$ sous forme d'un code informatique pour le calcul effectif du modèle. Un nouvel objet mathématico-informatique en

résulte. Il sera représenté par une fonction G appelée fonction code (FC) en référence à la double nature de fonction mathématique et de code informatique. L'espace de définition de G est l'espace Ω. L'écriture du code fait appel aux méthodes d'analyse numérique et de programmation. Il n'y a pas forcément identité entre $\mathcal{G}$ et G, notamment quand on discrétise, lors du codage, le temps et l'espace géographique utilisés dans le modèle théorique. Ces questions relèvent de l'analyse numérique. On supposera dans ce qui suit, et c'est en général souhaitable, que G fournit une bonne approximation numérique des sorties de $\mathcal{G}$. Deux grands types de FC se rencontrent en modélisation. Une conception mécaniste aboutit à une FC dont la sortie est entièrement déterminée par des entrées contrôlées. On parlera de FC déterministe. Les entrées étant fixées, la FC déterministe produit un scalaire fixe. Dans certaines problématiques liées à l'environnement, on doit prendre en compte l'aspect aléatoire des processus (dispersion de pollen, épidémie dans un verger, défaillance aléatoire,...). Dans le corps même de la FC, le tirage de nombres pseudo-aléatoires permettra de mimer ces processus. On parlera alors de FC stochastique. Les entrées étant fixées, la sortie n'est plus invariante avec la répétition de la simulation. La FC stochastique fournit dans ce cas un ensemble potentiel de résultats représenté au moyen d'une distribution de probabilité. Ce chapitre aborde essentiellement l'analyse d'une FC déterministe. Supposons le modélisateur, chercheur ou ingénieur, convaincu de la nécessité de comprendre et de valider une construction mathématico-informatique qui mime les propriétés attendues ou pas du système réel mais peut aussi contenir des artefacts. L'analyse de sensibilité globale propose au modélisateur une vision synthétique permettant d'appréhender le comportement du modèle dans un grand nombre de circonstances. Basée sur des méthodes rationnelles, l'analyse de sensibilité ouvre la voie à la critique raisonnée du modèle mais aussi à la critique des hypothèses sur lesquelles l'analyse de sensibilité est elle même fondée. La pratique de l'analyse de sensibilité est envisageable de la conception à l'utilisation du modèle. L'accent sera mis ici sur l'analyse du comportement du modèle en tant que tel et non sur la qualité d'ajustement à des observations. La définition de l'espace Ω sous forme d'hypercube peut s'avèrer trop générale quand certaines combinaisons des facteurs ne sont pas physiquement possibles ou bien lorsque des dépendances existent entre les facteurs. Des méthodes sont disponibles quand la dépendance des facteurs est analysée par corrélation ([**?**]). Une définition plus complexe de Ω nécessite alors une adaptation ou la mise en oeuvre de méthodes dont certaines relèvent de la recherche. L'analyse de sensibilité avec des facteurs dépendants n'est pas présentée dans ce texte d'initiation.

1.1.3 Analyse statistique de la Fonction Code

Il semble paradoxal d'analyser statistiquement, et donc de façon partielle, un modèle théorique et un code dont on peut connaître tous les éléments. La compréhension d'un système où s'entrecroisent concepts bien formalisés et représentations approchées de pro-

cessus suscite des questions. De plus, le concepteur a besoin d'une synthèse pertinente pour la réflexion et la critique. Les utilisateurs du modèle ont notamment besoin d'identifier les facteurs les plus influents pour élaborer des décisions. Une analyse rigoureuse de $\mathcal{G}$ et de son incarnation informatique G serait fort appréciée si on pouvait la mener à bien. L'analyse des propriétés mathématiques de $\mathcal{G}$ consiste à rechercher les points de discontinuité ou singuliers et à étudier les variations au moyen de sa dérivée. Cette analyse est difficile à mener avec l'aide du calcul formel dans le cas de modèles volumineux. Les méthodes récentes de dérivation automatique ([**?**]) sont prometteuses mais nécessitent une intrusion dans le code qui peut se révéler délicate si la FC est dérivable par morceaux du fait de l'utilisation de fonctions à seuil. Il serait aussi utile, notamment pour l'utilisateur, de définir l'ensemble $\mathcal{A} \subset \Omega$ des antécédents par la FC d'un ensemble de valeurs d'intérêts $\mathcal{B}$. Ce problème correspond notamment à la recherche de l'inverse de $\mathcal{G}$. Si on disposait de façon formelle des inverses par morceaux de $\mathcal{G}$, le problème serait résolu. A ce jour, on est obligé de résoudre ce problème de façon approchée à l'aide d'algorithme. Faute d'une analyse théorique des propriétés de la FC, le modélisateur est conduit à effectuer des simulations. La collection de simulations sans stratégie d'analyse peut s'avérer coûteuse en temps et insatisfaisante. Le choix qui est pris dans cet ouvrage est d'utiliser un cadre statistique pour construire et analyser les simulations. Une telle approche peut sembler à première vue superficielle pour analyser une FC constituée d'un système d'équations représentant des processus aux multiples facettes. En effet, cette approche ne permet pas de connaître des caractéristiques théoriques telles la continuité ou la dérivabilité du modèle. Par contre, elle permet de répondre à des questions liées à l'utilisation de la FC. Une autre problématique est l'utilisation d'une FC pour améliorer la prédiction d'un phénomène. Un argument en faveur de l'approche statistique résulte de l'incertitude inhérente à la FC. Cette incertitude provient de l'imprécision des estimations des paramètres de la FC mais aussi des choix pris pour écrire les équations constitutives de la FC. Beaucoup de concepts biologiques n'ont pas le niveau de formalisation mathématique que celui atteint par les concepts physiques. Une certaine liberté en résulte dans l'écriture des équations. Cette incertitude dans la formalisation des connaissances mérite d'être abordée. Les démons du hasard sont de retour dans la démarche de modélisation qui pensait peut-être en avoir réduit la place.

Dans ce qui suit, la FC est considérée comme une boîte noire dans laquelle on fait entrer des signaux dont on analyse l'effet sur les sorties. Les méthodes d'analyse seront dites non-intrusive dans le sens où on n'intervient pas sur le code de la FC (a contrario, la dérivation automatique est une méthode intrusive). Le modèle statistique est certes moins ambitieux dans sa forme mais il permet de fournir une vision synthétique et de hiérarchiser l'importance des facteurs. Ce dernier objectif s'avère utile pour l'expertise de la FC et pour élaborer des conseils destinés aux utilisateurs de la FC. Un modèle construit *sur* une FC est appelé métamodèle ou représentation d'un point de vue sur le modèle. Le métamodèle statistique cherche à approximer au mieux des sorties de la FC. La recherche d'un métamodèle est au coeur de certaines méthodes d'analyse de sensi-

bilité. Dans ce chapitre, le métamodèle permettra de définir et calculer des indices de sensibilité. La métamodélisation a d'autres utilisations (cf chapitre ?). Un modèle statistique classique consiste à écrire la sortie de la FC comme la somme des effets (principaux et d'interaction) des facteurs de la FC et d'un bruit. Le modèle statistique privilégié ici est linéaire (en les paramètres) et est estimé par la méthode d'analyse de la variance, en abrégé anova (**An**alysis **o**f **va**riance). On cherchera à quantifier l'importance relative des termes du modèle d'anova. La notion de criblage fait référence au tri de certains facteurs. Les indices de sensibilité globale de la FC définis comme des parts de variabilité permettront d'effectuer un criblage. Les méthodes statistiques d'inférence trouveront leur place du fait de l'utilisation de stratégies d'échantillonnage aléatoire et de metamodèles ayant une composante stochastique. L'analyse statistique de la FC sera liée au choix de variable de sortie correspondant à l'objectif de l'étude. Ce choix est fondamental et peut être appelé à s'enrichir au gré des questionnements des modélisateurs et des utilisateurs de la FC.

On ne limitera pas dans cet exposé l'analyse de sensibilité au calcul d'indices. Une analyse exploratoire graphique des résultats de simulations sera utile pour détecter des anomalies, aider à formaliser un meta-modèle statistique des sorties de la FC, nuancer l'interprétation ou susciter des voies nouvelles d'exploration. Ce texte peut être considéré comme une introduction aux méthodes. On trouvera des présentations et des compléments théoriques ou appliquées sur les méthodes de ce chapitre dans [**?**], [**?**], [**?**], [**?**] et [**?**]. La mise en pratique des méthodes étant un objectif important du texte, un paragraphe sera consacré au traitement d'un exemple réel à l'aide du logiciel R. Le code des fonctions nécessaires pour la mise en oeuvre étant fourni, le lecteur pourra expérimenter les méthodes avant de passer à l'analyse de ses propres modèles.

1.1.4 Exploration d'une Fonction Code

Pratiques élémentaires

Le balayage de tous les points de Ω, espace continu, est impossible, même dans sa version informatisée. En pratique, les autres facteurs limitants du balayage sont la dimension de Ω, i.e. le nombre de facteurs, ainsi que le temps de calcul pour obtenir une simulation (cf chapitre 1). Comment explorer de façon optimale un espace de grande dimension pour un objectif donné ? Cette question relève d'une problématique majeure en analyse de sensibilité. L'idée de faire varier dans une expérimentation les facteurs un par un, les autres étant fixés à une valeur moyenne ou de référence, est une idée qui séduit de par sa simplicité. Cette méthode peut certes donner un ordre de grandeur des effets des facteurs pris isolément, et revient à estimer dans le meilleur des cas un développement limité à l'ordre un de la FC. Cette méthode présente le sérieux inconvénient de ne pas aborder l'interaction. Son utilisation en analyse de sensibilité peut correspondre à un objectif d'analyse

locale. Or, même dans ce cas, il peut y avoir des interactions. Cette approche est trop limitée quand on veut connaître le comportement d'une FC dans un espace plus large que le voisinage immédiat d'un point de référence. En pratique, la qualité d'approximation d'une FC par un modèle additif du premier ordre est loin d'être suffisante.

Planification des simulations

La planification des simulations présentée est inspirée de deux méthodes d'échantillonnage connues : stratification et échantillonnage aléatoire. La stratification consiste à décomposer la population des unités statistiques en sous-groupes ou catégories distinctes et dans la mesure du possible homogènes. Un tirage aléatoire dans la population dans sa globalité ne garantit pas la présence d'unités dans tous les groupes et ce avec un effectif suffisant. Les strates étant définies on tirera donc au hasard dans chaque strate un certain nombre d'unités. Un argument important dans l'usage de cette stratégie réside dans le fait que l'échantillonnage stratifié permet d'obtenir des estimations plus précises que l'échantillonnage aléatoire. De plus, si la variabilité entre les strates est forte et si les variables qui servent à construire les strates sont fortement corrélées avec la caractéristique, la méthode gagnera en efficacité. La méthode perd en efficacité si les strates ne sont pas très homogènes mais une pré-étude peut aider à construire des strates homogènes. Cela dit, une stratification médiocre apporte toujours un plus pour la précision des estimateurs. Une certaine optimalité sur la précision est obtenue si la taille de l'échantillon d'une strate est proportionnelle à la taille de la strate ou bien à la variabilité de la caractéristique étudiée à lintérieur de la strate. Dans le contexte de l'expérimentation virtuelle, la stratification est élémentaire et consistera à structurer Ω, espace des facteurs de la FC à partir d'une discrétisation uniforme de la gamme des facteurs. La question de la qualité du remplissage de l'espace Ω se pose en expérimentation numérique. La stratification, même imparfaite, permet d'y répondre. Les strates ont la forme de pavés (hyper-cubes ou rectangles). Le pavage de Ω ainsi obtenu présente des analogies avec le plan d'expérience complet dans lequel toutes les combinaisons des facteurs sont réalisées. Le traitement statistique est simple si les nombres de répétitions des combinaisons des facteurs sont identiques. Le plan est alors dit complet et équilibré. Le plan complet et équilibré présente une géométrie garantissant l'orthogonalité des facteurs. Cette propriété est cohérente avec l'hypothèse d'indépendance ou, dans un sens plus faible, de non corrélation des facteurs. Si le découpage des gammes des facteurs n'est pas uniforme et si le nombre de tirages dans chaque pavé est proportionnel à son volume, le plan sera déséquilibré. La décomposition de la variabilité à la base de l'anova n'est alors plus unique. L'analyse de sensibilité est plus difficile et nécessite une méthodologie de comparaison de modèles basée entre autres sur la différence des sommes de carrés résiduels de deux modèles emboîtés. L'analyse est rendue compliquée par l'inflation de modèles possibles lorsque la FC a beaucoup de facteurs. Dans ce texte, l'équilibre du plan est choisi afin de préserver l'orthogonalité des facteurs.

Une approche originale proposée par Morris ne fait pas appel à l'écriture explicite d'un modèle statistique paramétrique. La méthode d'échantillonnage est définie sur une grille régulière superposée sur Ω. L'analyse des résultats de simulation est basée sur l'étude des variations de la FC en des points de Ω positionnés sur des trajectoires aléatoires définies sur la grille. L'avantage de cette méthode est de ne pas faire d'hypothèse restrictive et donc de mettre en évidence des comportements non-linéaires de la FC. La connaissance des principes de l'anova restera aussi une aide à l'interprétation. Pour dépasser la limitation inhérente à l'exploration des seuls noeuds d'une grille, une extension récente prenant en compte la continuité de l'espace sera signalée.

1.2 Discrétisation de l'espace des facteurs

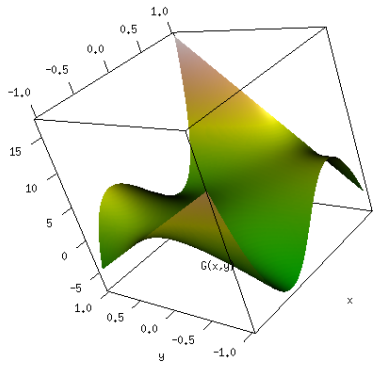
1.2.1 Fonction Code jouet

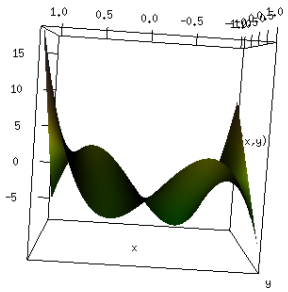
On considère le modèle jouet $\mathcal{G}$ déterministe suivant. Les arguments x et y de la fonction sont associés à deux facteurs d'intérêt X et Y. Le modèle est définie par une combinaison de polynômes de Legendre.

$\Omega = [-1, 1] \times [-1, 1] \to \mathbb{R}$,

$\mathcal{G}(x, y) = \mu + \alpha_1 \Psi_1(x) + \alpha_2 \Psi_2(x) + \beta_1 \Psi_1(y) + \gamma\, \Psi_3(x)\Psi_1(y)$

avec $\Psi_1(x) = x\sqrt{3}$, $\Psi_2(x) = \frac{\sqrt{5}}{2}\,(3\,x^2 - 1)$, $\Psi_3(x) = \frac{\sqrt{7}}{2}\,(5\,x^3 - 3\,x)$. Les scalaires μ, α_1, α_2, β_1 et γ sont des constantes. Par construction, l'intégrale de tout polynôme Ψ est nulle et l'intégrale de son carré est égale à un (pour la mesure 1/2). Ce modèle est simple et facile à manipuler sur le plan théorique mais il présente l'intérêt d'avoir une forme non triviale (fig. 1). Il permettra illustrer et confronter les calculs théoriques aux résultats de deux méthodes (anova et Morris). La variable de sortie analysée est l'image de la fonction.





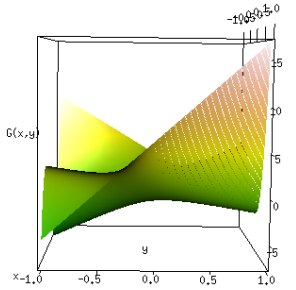


Figure 1 : le modèle jouet $\mathcal{G}(x, y)$ (cf définition dans le texte) sous plusieurs facettes.

La FC G associée à $\mathcal{G}(x, y)$ est écrite dans le langage du logiciel R et se présente sous la forme suivante :

```
G = function(x,y){
   Psi1 = function(u){x*sqrt(3)}
   Psi2 = function(u){sqrt(5)*(3*x^2-1)/2}
   Psi3 = function(u){sqrt(7/32)*(5*x^3-3*x)}
   mu =0 ; alpha1 = 1 ; alpha2 = 2;
   beta1=1 ; gamma=2;
   mu + alpha1*Psi1(x) +  alpha2*Psi2(x)+ beta1*Psi1(y) + gamma*Psi3(x)*Psi1(y)
}
```

L'analyse statistique procèdera comme si G était une boîte noire sans supposer la moindre information sur sa forme géométrique.

1.2.2 Discrétisation de Ω

Deux possibilités sont envisagées, la grille ou le pavage de Ω. Le choix d'une grille implique que les calculs seront effectués uniquement sur les noeuds. Il est souhaitable que la FC ne présente pas de trop fortes variations entre le noeuds de la grille. En toute rigueur, seule une analyse mathématique permet de l'affirmer. Par exemple, si la FC est définie par un rapport de polynômes, les variations peuvent être importantes au voisinage des racines du dénominateur. Cela dit, si les problèmes de passage à l'infini ont pu être évités et si le maillage est fin, l'hypothèse semble réaliste. Par contre, un maillage fin risque de faire exploser le nombre de noeuds et la possibilité en pratique de faire les calculs. D'où la nécessité de se contenter d'un maillage relativement grossier quand on a beaucoup de

facteurs. Cela dit, les limites de mise en oeuvre dépendent aussi de la puissance de calcul à disposition.

Le pavage de Ω constitue une réponse au côté restrictif de la grille. C'est l'extention naturelle de la grille, un pavé étant défini par l'espace entre les noeuds. Aucun point de l'espace n'est exclu a priori dans le dispositif en pavés. Les découpages de $[-1, 1]$ en I intervalles pour X et en J intervalles pour Y définissent un ensemble de pavés $P_{i,j}$ (ici des carrés) disjoints tel que $\Omega = \underset{i,j}{\cup} P_{i,j}$. L'analyse statistique qui suit est présentée sur un pavage de Ω. L'analyse sur une grille est très proche, mise à part l'inférence.

1.2.3 Plan d'expérience et métamodèle associé

Dans chaque pavé $P_{i,j}$ on échantillonne au hasard R points selon une loi uniforme. On suppose de plus que les facteurs X et Y peuvent varier de façon indépendante (Fig. 2).

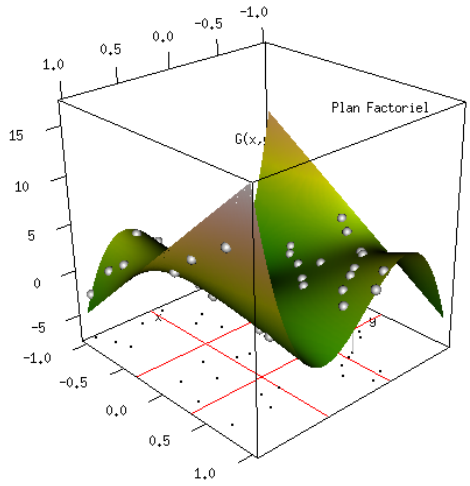


Figure 2 : pavage de Ω en neuf pavés avec un tirage aléatoire de quatre points par pavé.

Le métamodèle utilisé est le modèle classique d'anova à effets *fixes* sur un plan complet et équilibré. On en rappelle les principaux éléments. Soit,

$$G_{ijr} = M + X_i + Y_j + X{:}Y_{i,j} + \epsilon_{i,j,r} \quad \forall i = 1, I \;\; j = 1, J \;\; r = 1, R \tag{1.1}$$

G_{ijr} est la valeur de la FC pour la répétition r située dans le pavé $P_{i,j}$. M est une constante. X_i (resp. Y_j) est appelé i^{eme} effet principal de X (resp. de Y). $X{:}Y_{ij}$ représente un effet d'interaction. Les effets sont exprimés dans l'unité de la sortie de la FC. L'erreur $\epsilon_{i,j,r}$

est la réalisation d'une variable aléatoire ϵ supposée d'espérance nulle et de variance σ^2. Pour obtenir des estimations uniques, on utilise les contraintes suivantes : $\sum_{i=1}^{I} X_i = 0$, $\sum_{j=1}^{J} Y_j = 0$, $\forall j$, $\sum_i X:Y_{ij} = 0$ et $\forall i$, $\sum_j X:Y_{ij} = 0$. Il y a IJ effets à estimer. Cette définition des contraintes correspond à l'idée que les effets de plusieurs traitements (i.e. les niveaux d'un facteur) appliqués simultanément se compensent.

La constante M est estimée par la moyenne $\bar{G}_{\bullet\bullet\bullet} = \sum_{i,j,r} \frac{G_{i,j,r}}{IJR}$ des résultats de simulation. Le $\bullet$ indiquant une sommation sur l'indice de même position, les effets X_i (resp. Y_j) sont définis par $\widehat{X_i} = \bar{G}_{i\bullet\bullet} - \bar{G}_{\bullet\bullet\bullet}$ (resp. $\widehat{Y_j} = \bar{G}_{\bullet j\bullet} - \bar{G}_{\bullet\bullet\bullet}$ pour Y_j). L'effet d'interaction $X{:}Y_{ij}$ est estimé par $\widehat{XY}_{ij} = \bar{G}_{ij\bullet} - \bar{G}_{i\bullet\bullet} - \bar{G}_{\bullet j\bullet} + \bar{G}_{\bullet\bullet\bullet}$

Remarque : l'othogonalité résulte de la géométrie sous jacente au modèle d'anova et notamment de l'écriture matricielle du modèle d'anova que l'on verra plus bas.

1.2.4 Commentaires sur le modèle d'anova

Ce métamodèle simple revient à représenter la FC par une fonction en escalier. Comparé à la représentation exacte, le modèle d'anova présente des discontinuités et semble peu efficace pour effectuer des interpolations en particulier si le pavage est grossier. Toutefois l'anova ajuste au mieux (i.e. au sens des moindres carrés des écarts) les moyennes des simulations dans chaque pavé. Les effets principaux sont représentés à l'aide de deux fonctions à une variable. Se contenter de la somme des effets principaux et de la constante M comme métamodèle revient à écrire une fonction à deux variables par la somme de deux fonctions uni-variables. Cela revient à dire que la différence entre les effets du facteur X en deux valeurs distinctes du facteur Y est constante. On notera que $\bar{G}_{i..}$ est une estimation de $\int_{\bigcup_j P_{i,j}} \mathcal{G}(x,y)dxdy$. La stochasticité du métamodèle provient du caractère aléatoire de l'échantillonnage et non de la FC qui est déterministe. De plus, le modèle d'anova suppose que la variance du bruit est invariante sur l'ensemble des pavés. Cela revient à supposer que $\int_{P_{i,j}} \mathcal{G}^2(x,y)dxdy - \left(\int_{P_{i,j}} \mathcal{G}(x,y)dxdy)\right)$ est une constante $\forall i,j$ ce qui est en général faux. Cela ne remet nullement en cause le principe de l'utilisation de l'anova pour une FC mais nécessitera de bien définir une méthode d'inférence.

1.2.5 Premiers calculs

Dans notre exemple, on choisit $I = J = 3$ et $R = 2$. On obtient à partir de la table des simulations (Table 1) les estimations des effets (Table 2) pour le modèle jouet. De plus, on obtient $\bar{G}_{...} = 0.24$ pour la moyenne et $\sigma^2_G = 8.67$ pour la variance de la sortie. Différents modèles d'anova sont représentés par des surfaces en escalier (figure 3). On constate que les deux modèles d'anova prennent mal en compte les niveaux élevés de la FC sur les bords de Ω.

simulation	i	j	r	G_{ijr}
1	1	1	1	-5.9
2	1	1	2	-5
3	1	1	3	-2.3
4	1	1	4	-1.8
5	1	2	1	1.97
6	1	2	2	-2.13
7	1	2	3	1.16
8	1	2	4	-0.65
⋮	⋮	⋮	⋮	⋮
33	3	3	1	9.89
34	3	4	2	-1.9
35	3	3	3	-0.6
36	3	3	4	-0.9

Table 1 : extrait des résultats de simulation sur le modèle jouet. L'échantillonnage est basée sur un plan complet et équilibré à deux facteurs dont les gammes de variation ont été découpées en trois classes numérotées par i et j. r est le numéro de répétition. G_{ijr} est la sortie associée.

i ou j	X_i	Y_j
1	-0.82	-1.76
2	-0.79	0.44
3	1.61	1.324

$i \setminus j$	1	2	3
1	-1.40	0.23	1.17
2	1.07	-1.25	0.18
3	0.33	1.02	-1.35

table 2 : tables des effets principaux (à gauche) et d'interaction (à droite) obtenus à partir du plan complet et équilibré. La sommes des effets est nulle.

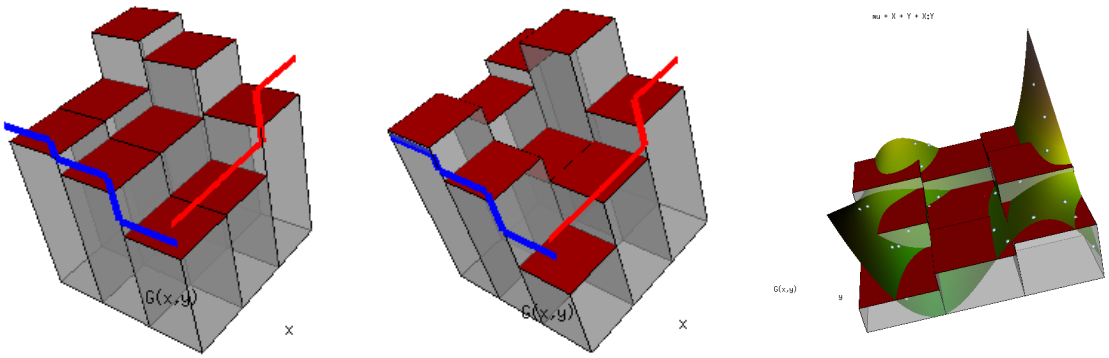


Figure 3 : représentation de l'ajustement du modèle jouet $\mathcal{G}$ avec uniquement les effets principaux (à gauche) et l'ensemble des effets (au milieu) issus de l'anova. Les moyennes marginales des facteurs sont représentées en bleu (Y)et rouge (X). Le graphique de droite superpose, dans une échelle différente, la surface du modèle jouet, l'ajustement du modèle d'anova complet et les points échantillonnés.

1.3 Analyse de sensibilité par anova

1.3.1 Introduction à l'analyse conditionnelle

Le modèle d'anova représente la réponse moyenne de la sortie de la FC ou, de façon plus théorique, l'espérance mathématique. L'espérance $\mathbb{E}$ d'une variable aléatoire est un opérateur linéaire. Sous les hypothèses de l'anova, on a :

$$\mathbb{E}(G_{ijk}) = M + X_i + Y_j + X{:}Y_{i,j}$$

et

$$V(G_{ijk}) = \sigma^2$$

Le modèle d'anova servira de base aux calculs formels. L'espérance de la sortie *sachant* que X prend la modalité i, notée $\mathbb{E}(G_{ijr}|X = i)$, est donnée par :

$$\mathbb{E}(G_{ijr}|X = i) = \frac{1}{JR}\sum_{jr}(M + X_i + Y_j + X{:}Y_{i,j}) + \mathbb{E}(\epsilon_{ijr}) = M + X_i = \mathbb{E}(\bar{G}_{i\cdot})$$

Cette quantité sera notée en abrégé $\mathbb{E}(G|i)$. La relation $i \rightarrow \mathbb{E}(G_{ijr}|X = i)$, notée $\mathbb{E}[G|X]$, représente l'effet principal du facteur X. Elle est appelée *espérance conditionnelle* de G sachant X. Une courbe de régression est un exemple de représentation de l'espérance conditionnelle. La variance de la réponse pour un niveau spécifique du facteur explicatif nous informe de l'information non prise en compte par le facteur pris isolément. Du fait de l'orthogonalité des facteurs, on a :

$$V(G_{ijr}|X = i) = V(G|i) = \mathbb{E}\left[M + X_i + Y_j + X{:}Y_{i,j} + \epsilon_{ijr} - M - X_i\right]^2$$
$$= \frac{1}{JR}\sum_{jr}(Y_j + X{:}Y_{i,j})^2 + \sigma^2 = \frac{1}{J}\sum_{j} Y_j^2 + \frac{1}{J}\sum_{j}(X{:}Y_{i,j})^2 + \sigma^2$$

La relation $i \rightarrow V(G|X = i)$définit la variance conditionnelle de G sachant X. Cette variance peut être uniquement celle d'un bruit comme c'est le cas, par exemple, pour un modèle de régression linéaire *simple* (i.e. à une variable). Ici, elle intègre les variances des effets des autres facteurs.

L'effet principal d'un facteur X est nul si la moyenne des réponses (i.e. en théorie l'espérance conditionnelle) de la FC est constante à tous les niveaux de X. D'où l'idée d'utiliser la variance de ces moyennes comme indicateur de l'effet de X pris isolément. On a :

$$V[\mathbb{E}(G|X)] = \frac{1}{I}\sum_{i=1}^{I} X_i^2$$

La moyenne des variances de la sortie en chaque niveau de X est associée à l'information non prise en compte par le facteur X. Soit,

$$\mathbb{E}[V(G|X)] = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{J}\sum_{j} Y_j^2 + \frac{1}{J}\sum_{j} X{:}Y_{i,j}^2 + \sigma^2\right) = \sigma^2 + \frac{1}{J}\sum_{j} Y_j^2 + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j}(X{:}Y_{i,j})^2$$

A partir des équations qui précèdent, on retrouve in fine l'équation classique de décomposition de la variance d'une variable expliquée par des facteurs orthogonaux :

$$V(G) = \mathbb{E}[V(G|X)] + V[\mathbb{E}(G|X)] = V[\mathbb{E}(G|X)] + V[\mathbb{E}(G|Y))] + V[\mathbb{E}(G|X{:}Y)] + \sigma^2$$

.

La FC test permet de calculer analytiquement ces quantités pour un échantillonnage uniforme sur Ω. Par construction, deux polynômes de Legendre Ψ_p et Ψ_q ont une intégrale nulle sur $[-1,1]$ (et donc une espérance nulle par rapport à la loi uniforme sur $[-1,1]$ dont la densité vaut $1/2$). L'orthogonalité de deux polynômes est définie par $\frac{1}{2}\int_{-1}^{1}\Psi_p(x)\Psi_q(x)dx = \delta_{p,q}$ ($=0$ si $p \neq q$, 1 sinon). On déduit aisément que $V(G) = \alpha_1^2+\alpha_2^2+\beta_1^2+\gamma^2 = 10$. De plus, $E(G|X=x) = \frac{1}{4}\int_{-1}^{1} G(x,y)dy = \mu + \alpha_1\Psi_1(x) + \alpha_2\Psi_2(x)$ (le coefficient $1/4$ représente la densité uniforme sur un carré de surface 4). D'où $V(E(G|X)) = \alpha_1^2 + \alpha_2^2 = 5$. De même, $V(E(G|Y)) = \beta_1^2 = 1$. On définit la contribution de chaque facteur à la variance totale par $I_x = \frac{5}{10} = 0.5$ et $I_y = \frac{1}{10} = 0.1$. Le terme contenant le produit de deux polynômes a pour contribution $\frac{4}{10} = 0.4$. Ce terme sera assimilé à une interaction entre X et Y.

1.3.2 Décomposition de la somme des carrés

La décomposition de la variabilité est au coeur de la méthode anova. Un peu d'algèbre est nécessaire pour préciser les calculs nécessaires. La décomposition de la somme des carrés des écarts à la moyenne générale des résultats de simulation G_{ijr} $i = 1, I$, $j = 1, J$ et $r = 1, R$ du plan complet et équilibré est ici obtenue à partir de l'équation suivante :

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{r=1}^{R}(G_{ijr} - \bar{G}_{\bullet\bullet\bullet})^2 = SS_x + SS_y + SS_{x:y} + SS_\epsilon = 303.31$$

avec

$$SS_x = JR\sum_{i=1}^{I} X_i^2 = 6\sum_{i=1}^{3}(\bar{G}_{i\bullet\bullet} - \bar{G})^2 = 46.72$$

représentant la variabilité des I effets du facteur X,

$$SS_y = IR\sum_{j=1}^{J} Y_j^2 = 6\sum_{j=1}^{3}(\bar{G}_{\bullet j\bullet} - \bar{G})^2 = 60.45$$

la variabilité des effets du facteur Y, et

$$SS_{x:y} = R\sum_{i=1}^{I}\sum_{j=1}^{J}(X{:}Y_{ij})^2 = 2\sum_{i=1}^{3}\sum_{j=1}^{3}(\bar{G}_{ij\bullet} - \bar{G})^2 = 36.37$$

la variabilité des effets d'interaction. Enfin,

$$SS_\epsilon = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{r=1}^{R}(G_{ijr} - \bar{G}_{ij\bullet})^2 = 159.58$$

quantifie la variabilité de l'erreur. *Remarque* : la variance est plus utilisée que la somme des carrés des écarts. On constate aisément que, si $n = IJR$ est grand, l'estimation de la variance $V(G)$ de la sortie est proche (à $n/(n-1)$) près de $\frac{SS_T}{n}$. D'où

$$V(G_{ijk}) \simeq \frac{\sum_{i=1}^{I} X_i^2}{I} + \frac{\sum_{j=1}^{J} Y_j^2}{J} + \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(X{:}Y_{ij})^2}{IJ} + \frac{SS_\epsilon}{n}$$

Cette formule approchée de la décomposition de la variance ne tient pas compte des degrés de liberté réels utilisés pour définir des estimateurs non-biaisés.

1.3.3 Indices de sensibilité

La décomposition théorique de la variance et, en pratique, de la somme des carrés permet de définir des indices caractérisant l'importance des différents facteurs à partir de la table d'anova (table 3). L'indice I_x (resp. I_y) principal associé à X (resp. Y)est défini par : $I_x = \frac{SS_x}{SS_T} = 0.15$ (resp. $I_y = \frac{SS_y}{SS_T} = 0.2$). On notera que SS_T prend en compte l'erreur résiduelle. Une autre notion rencontrée en analyse de sensibilité globale est celle d'indice total associé à un facteur. Un facteur contribue isolément mais aussi par l'intermédiaire des interactions. Les indices de sensibilité totaux IT_x et IT_y du facteur X et Y sont définis par : $IT_x = \frac{SS_x + SS_{x:y}}{SS_T} = 0.27$ et $IT_y = \frac{SS_y + SS_{x:y}}{SS_T} = 0.32$. Reste, dans notre cas, la variabilité du bruit quantifiée par le rapport $\frac{SS_\epsilon}{SS_T} = 0.53$. L'explication globale de la FC par le modèle d'anova est quantifiée par le coefficient $R^2 = 1 - \frac{SS_x+SS_y+SS_{xy}}{SS_T}$. Les sommes des carrés sont présentées dans la table d'anova (table 3).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	2	46.715	23.357	3.9520	0.03123 *
Y	2	60.648	30.324	5.1307	0.01292 *
X :Y	4	36.371	9.093	1.5385	0.21928
Residuals	27	159.578	5.910		

Table 3 : table d'anova sur le modèle jouet. Le test de Fisher avec une erreur de première espèce de 0.05 est donc significatif pour X et Y.

1.3.4 Inférence

Si on analyse la FC déterministe sur une grille correspondant à un dispositif complet, la population des unités statistique est constituée des noeuds de la grille. On a toute l'information nécessaire pour calculer de façon exacte les indices de sensibilité de la FC à partir d'un modèle d'anova contenant toutes les interactions possibles avec le dispositif. L'inférence statistique n'a alors pas de sens. Si l'on fait l'hypothèse a priori que certaines interactions d'ordre élevé sont négligeables et assimilables à un bruit, on a alors la possibilité de réaliser une inférence sur les indices à partir de la procédure qui sera expliquée ci-dessous dans le contexte du pavage.

Dans le cas de l'échantillonnage sur pavage avec tirage aléatoire dans chaque pavé, l'inférence sur les effets est possible à partir de la table d'anova. L'importance d'un effet ne se résume pas à sa significativité. La taille de l'échantillon augmente la puissance du test et donc la capacité à discriminer des moyennes. En expérimentation numérique, une taille d'échantillon élevée peut induire des différences significatives même si la valeur de ces différences est faible.

L'inférence sur les indices de sensibilité sera abordée à l'aide de la méthode de re-échantillonnage bootstrap ([**?**], [**?**]). On recherche l'intervalle de confiance ayant la probabilité 0.95 de contenir la vraie valeur d'un indice. La première stratégie de bootstrap consistera à générer ν nouveaux échantillons en effectuant dans chaque pavé un tirage avec remise parmi les R points dont on dispose. Cela est possible si $R > 1$. Pour chacun de ces échantillons simulés on calcule la table d'anova et les indices de sensibilité. In fine, les ν valeurs d'indices obtenues définissent les distributions des indices sous l'hypothèse du modèle d'anova. Les quantiles d'ordre 0.025 et 0.975 fourniront un intervalle de confiance de probabilité 0.95 des indices. Cette façon de procéder présente l'avantage de prendre en compte l'hétéroscédasticité, ou hétérogénéité des variances entre pavés, de la sortie de la FC.

Une autre stratégie de bootstrap consiste à re-échantillonner avec remise dans la population des résidus du modèle d'anova initial. Cela suppose en particulier que la variance de l'erreur est constante sur tous les pavés. Un nouvel échantillon est obtenu en sommant les résidus re-échantillonnés et l'ajustement du modèle d'anova initial. On calcule ensuite les indices de sensibilité à partir de la table issue de l'anova sur cet échantillon. On calcule enfin les intervalles de confiance à l'aide des quantiles des distributions.

Quand la FC a beaucoup de facteurs, il peut s'avérer difficile d'effectuer des répétitions dans chaque pavé. Dans le cas d'une seule répétition par pavé, il est possible, au vue de la table d'anova, de considérer certaines interactions d'ordre élevé équivalente à un bruit. Le modèle d'anova estimé en ne tenant pas compte de ces interactions fournira l'ensemble des résidus utilisé pour le bootstrap. Mais on se gardera d'une généralisation hâtive de cette hypothèse consistant à négliger des interactions. Par exemple, le produit de plusieurs facteurs inclus dans la définition de la FC est une source potentielle d'interaction d'ordre

élevé et interprétable.

Indices principaux	X	Y
Boot. méthode 1		
b_1	0.027	0.1
b_2	0.36	0.37
Boot. méthode 2		
b_1	0.04	0.05
b_2	0.37	0.44

Indices totaux	X	Y
Boot. méthode1		
b_1	0.18	0.25
b_2	0.65	0.74
Boot. méthode2		
b_1	0.16	0.19
b_2	0.58	0.62

Table 4 : intervalles de confiance $[b_1, b_2]$ de probabilité 0.95 sur les indices de sensibilité des facteurs principaux et totaux des facteurs X et Y du modèle jouet. Les indices sont calculés à partir du modèle d'anova avec effet principaux et d'interaction. Les intervalles sont obtenus par deux méthodes de bootstrap : re-échantillonnage par pavé (méthode 1) et re-échantillonnage sur la distribution globale des résidus d'une anova (méthode 2).

Les intervalles de confiance ainsi obtenus ont une grande amplitude. Cette constatation incite à augmenter le nombre de répétitions R. Les contributions des facteurs principaux (SS_x/SS_T et SS_y/SS_T) et d'interaction ($SS_{x:y}/SS_T$) à la somme des carrés totale donne un ordre de grandeur médiocre si on les compare avec les vrais valeurs (fig. 4). Cette observation confirme que la taille de l'échantillon est insuffisante (dans des pratiques réelles, on ne dispose évidemment pas de la surface de réponse réelle, on se basera donc sur les intervalles de confiance). Deux possibilités se présentent : choisir une discrétisation plus fine des gammes des facteurs ou augmenter le nombre de répétitions dans chaque pavé en gardant le même niveau de discrétisation. Les intervalles de confiance suivant sont obtenus à partir de ces deux stratégies (table 5 et 6).

Indices principaux	X	Y
Boot. méthode 1		
b_1	0.39	0.15
b_2	0.53	0.27

Indices totaux	X	Y
Boot. méthode1		
b_1	0.69	0.83
b_2	0.83	0.89

Table 5 : intervalles de confiance $[b_1, b_2]$de probabilité 0.95 par bootstrap dans chaque pavé. Discrétisation en 5 niveaux, R=2. Indices calculés par anova.

Indices principaux	X	Y
Boot. méthode 1		
b_1	0.27	0.002
b_2	0.52	0.1

Indices totaux	X	Y
Boot. méthode1		
b_1	0.40	0.1
b_2	0.75	0.37

Table 6 : intervalles de confiance $[b_1, b_2]$de probabilité 0.95 par bootstrap dans chaque pavé. Discrétisation en 3 niveaux, R=6. Indices calculés par anova.

Dans les deux cas, les niveaux des indices sont mieux estimés, mais la stratégie consistant à affiner le pavage donne des résultats plus précis.

On a pu constater sur le modèle jouet que l'échantillonnage aléatoire standard peut mal prendre mal les bords du domaine de définition de la FC (fig. 3 à droite). Un meilleur

contrôle des marges à l'intérieur d'un pavé (par exemple, à l'aide d'un hyper-cube latin) améliorerait probablement l'échantillonnage. Compléter l'échantillonnage initial sur les frontières du domaine de définition de la FC est aussi envisageable.

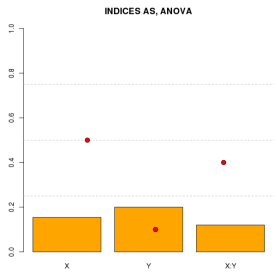


Figure 4 : comparaison des contributions de chaque facteur (principal et interaction) calculés à partir du modèle d'anova (barres) et des valeurs théoriques exactes (points rouges) sur le modèle jouet.

Cet exemple relativement simple a permis de mettre l'accent sur deux contraintes à prendre en compte pour avoir une information statistique fiable : la taille et la qualité de remplissage de l'espace de l'échantillon. L'inférence est primordiale pour avoir une idée de la précision sur les indices statistiques et constitue un garde fou pour l'interprétation.

1.3.5 Variante d'échantillonnage

Dans les paragraphes qui précèdent, les tirages des points dans chaque pavé sont indépendants et suivent une loi uniforme. En pratique, si le nombre K des facteurs est grand la discrétisation des gammes $[-1, 1]$ ne pourra pas être très fine afin de ne pas faire exploser le nombre calculs. Le souci d'un bon remplissage des pavés se pose alors. La méthode des hypercubes latin (LHS) (cf chapitre 4) permet de mieux contrôler le remplissage d'un pavé tout en respectant l'aspect aléatoire et la distribution uniforme des marges de dimension unitaire. De plus, un tri dans les LHS quant à leur qualité de remplissage permet d'améliorer la procédure (réf. ? ?).

1.3.6 Intérêt et limites du plan complet équilibré

On a pu noter la facilité de construction d'un plan complet et équilibré. Il en est de même de l'analyse statistique des résultats. Le modèle d'anova est interprétable et les interactions d'ordre élevé caractérisées. La représentation graphique des effets complète l'analyse et est utile pour définir un metamodèle continu. De plus, il est possible de faire intervenir dans un modèle d'anova des facteurs de types différents, quantitatifs et

qualitatifs (ordonnés ou pas). Les facteurs quantitatifs seront discrétisés. Les indices sont facilement calculables et l'inférence est possible par simulation. Limite majeure de ce plan, il s'avère coûteux voire impossible à réaliser pour une FC comportant un nombre de facteurs élevé, même si le pavage n'est pas très fin.

Exemple : soit une FC à 30 facteurs. La gamme de chaque facteur et découpée en trois intervalles et un seul tirage est effectué dans chaque pavé. Le plan complet exigera $3^{30} \sim 2 \times 10^{14}$ simulations. Si le temps d'exécution de la FC est $10^{-4}s$ et si on dispose d'une ferme de calcul composée de 1000 processeurs alors 7.8 mois en temps de calcul sont nécessaires !

Il ne semble pas réaliste d'utiliser cette méthode dans toute sa généralité dans tous les cas. Par contre, si le nombre des facteurs d'intérêt a pu être réduit a priori ou a posteriori en utilisant des méthodes exploratoires (cf Morris), la méthode apporte des informations intéressantes.

1.4 Plan fractionnaire

1.4.1 Matrice d'incidence d'un plan factoriel

La définition du plan fractionnaire nécessite de préciser l'écriture matricielle du modèle d'anova. On utilisera un exemple simple pour introduite les notations. Soit une FC à trois facteurs $G(X, Y, Z)$ définie sur un hypercube. On discrétise les facteurs en deux niveaux -1 et $+1$. Les facteurs discrétisés seront notés A, B et C. Le plan complet sans répétition engendre $N = 2^3$ combinaisons présentées dans la table suivante :

A	B	C	G
-1	-1	-1	0.367
-1	-1	$+1$	0.532
-1	$+1$	-1	0.495
-1	$+1$	$+1$	0.489
$+1$	-1	-1	0.310
$+1$	-1	$+1$	0.485
$+1$	$+1$	-1	0.476
$+1$	$+1$	$+1$	0.440

Table 7 : plan factoriel complet à 3 facteurs ayant deux niveaux.

Soit un modèle d'anova ne comportant que les effets principaux des trois facteurs. Compte tenu de la contrainte imposée sur les effets ($A_+ + A_- = B_+ + B_- = C_+ + C_- = 0$), le modèle d'anova comprend 4 paramètres notés dans un vecteur $\theta = (\mu, A_+, B_+, C_+)'$. La forme matricielle du modèle est alors :

$$G = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & +1 \\ 1 & -1 & +1 & -1 \\ 1 & -1 & +1 & +1 \\ 1 & +1 & -1 & -1 \\ 1 & +1 & -1 & +1 \\ 1 & +1 & +1 & -1 \\ 1 & +1 & +1 & +1 \end{pmatrix} \begin{pmatrix} \mu \\ A_+ \\ B_+ \\ C_+ \end{pmatrix} = D\theta$$

Le résultat du modèle associé à la simulation 3 s'écrit : $G_{-,+,-} = \mu - A_+ + B_+ - C_+$

La matrice D est appelée matrice d'incidence. Le produit scalaire de deux colonnes de D est égal à zéro. Cette propriété caractérise un plan complet et équilibré. L'interaction est définie par le produit terme à terme des colonnes des facteurs associés. Les contraintes imposées sur les paramètres de l'interaction d'ordre 2 entre A et B impliquent l'estimation d'un seul paramètre. En effet, on a $AB_{++} + AB_{+-} = AB_{-+} + AB_{--} = AB_{-+} + AB_{++} = AB_{--} + AB_{+-} = 0$ soit $AB_{+-} = AB_{-+} = -AB_{--} = -AB_{++}$. Il y a donc un seul paramètre AB_{++} à estimer pour l'interaction entre A et B. Le nombre de degrés de liberté correspond au nombre de paramètres estimables associés à un facteur. La partie du modèle contenant les effets principaux et toutes les interactions d'ordre 2 s'écrit :

$$\begin{pmatrix} \mu & A & B & C & A:B & A:C & B:C \\ 1 & -1 & -1 & -1 & +1 & +1 & +1 \\ 1 & -1 & -1 & +1 & +1 & -1 & -1 \\ 1 & -1 & +1 & -1 & -1 & +1 & -1 \\ 1 & -1 & +1 & +1 & -1 & -1 & +1 \\ 1 & +1 & -1 & -1 & -1 & -1 & +1 \\ 1 & +1 & -1 & +1 & -1 & +1 & -1 \\ 1 & +1 & +1 & -1 & +1 & -1 & -1 \\ 1 & +1 & +1 & +1 & +1 & +1 & +1 \end{pmatrix} \begin{pmatrix} \mu \\ A_+ \\ B_+ \\ C_+ \\ AB_{++} \\ AC_{++} \\ BC_{++} \end{pmatrix}$$

1.4.2 Effets factoriels

La représentation du dispositif à trois facteurs précédent est possible à l'aide d'un cube, chaque dimension étant associée à un facteur. Chaque sommet du cube correspond à une combinaison des niveaux des facteurs (fig. 5).

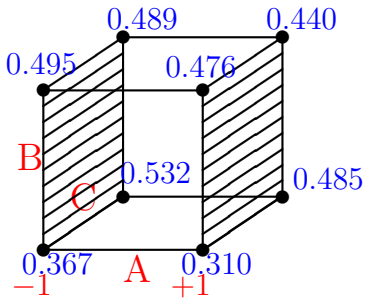


Figure 5 : représentation du plan factoriel comprenant trois facteurs A, B et C à 2 niveaux (-1 et +1). Les valeurs de la sortie G sont indiquées sur les sommets du cube. Un plan hachuré regroupe les sommets associés au même niveau du facteur A.

Le plan étant équilibré, les effets des facteurs B et C s'annulent quand on calcule la moyenne des simulations situées sur les sommets d'une face du cube. L'effet factoriel $e(A)$ de A est défini par la *demie-différence* de ces moyennes. Cette définition diffère de celle introduite plus haut dans le cadre du modèle d'anova. Dans le modèle d'anova classique, un effet A_i est défini par l'écart entre la moyenne de G à un niveau d'un facteur et la moyenne générale. Pour un facteur à deux niveaux, on a donc $A_+ - A_- = 2e(A)$. D'où l'usage de la notation $e(.)$ dans ce qui suit.

1.4.3 Confusion d'effets

Les colonnes de la matrice d'incidence réservées aux interactions présentent un codage équivalent à celui d'un facteur à deux niveaux. Sous l'hypothèse d'une interaction négligeable, on peut donc associer un nouveau facteur à cette interaction. Par exemple, si on suppose les interactions du second ordre négligeables entre les facteurs A, B et C, on peut leur associer respectivement trois nouveaux facteurs, D, E et F, en réservant toujours l'interaction d'ordre 3 au bruit.

M	A	B	C	$D = A:B$	$E = A:C$	$F = B:C$	$A:B:C$
+	−	−	−	−	+	+	+
+	−	−	+	+	+	−	−
+	−	+	−	+	−	+	−
+	−	+	+	−	−	−	+
+	+	−	−	+	−	−	+
+	+	−	+	−	−	+	−
+	+	+	−	−	+	−	−
+	+	+	+	+	+	+	+

Le plan orthogonal ainsi obtenu permet d'analyser les effets principaux de 6 facteurs avec seulement huit expériences (fig. 7). Un plan complet à 6 facteurs nécessiterait $2^6 = 64$ simulations. Le plan défini est une fraction 1/8 du plan complet et comprend 2^{6-3} simulations. D'où la notation générale 2^{K-q} du plan fractionnaire correspondant à la q^{ieme} fraction d'un plan complet comprenant K facteurs à deux niveaux.

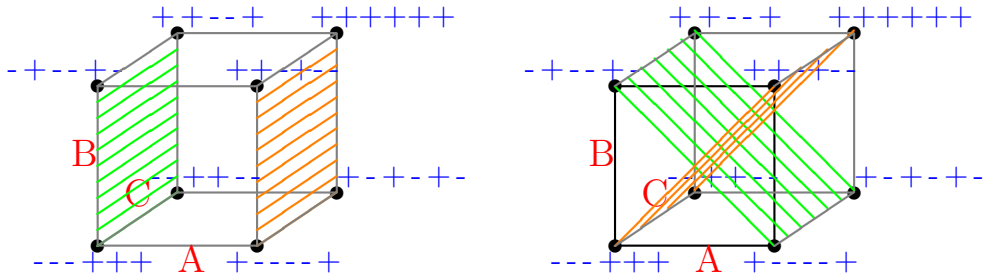


Figure 6 : $(-+--+-) = (A=-1, B=+1, C=-1, D=-1, E=+1, F=-1)$. Les plans hachurés regroupent les sommets correspondant au niveau $-$ (couleur verte) et $+$ (couleur orange) du facteur A (à gauche) et D (à droite).

Par construction, les nouveaux facteurs sont donc confondus ou aliasés avec des interactions du plan complet à 3 facteurs. Ne pas chercher à interpréter des interactions d'ordre élevé (au delà de 4) est pratique courante. Le coût du plan complet provient notamment de sa capacité à estimer un grand nombre d'interactions. D'où l'idée de tirer profit d'interactions supposées négligeables pour construire des plans qui se contentent de ne mettre en évidence que des interactions interprétables. Cette stratégie a fait ses preuves dans un contexte expérimental classique. La validation des hypothèses sur les interactions en jeu dans une FC sera étudiée dans la mesure du possible.

1.4.4 Générateurs d'un plan fractionnaire

Ce paragraphe sera développé dans la prochine version du texte. Il n'est pas fondamental pour la compréhension de la suite.

1.4.5 Résolution d'un plan fractionnaire

Un plan fractionnaire sera dit de résolution $\mathcal{R}$ s'il permet d'analyser sans confusions les interactions d'ordre $(\mathcal{R}-1)/2$ dans le cas où $\mathcal{R}$ est impair ou les interactions d'ordre $(\mathcal{R}-2)/2$ dans le cas où $\mathcal{R}$ est pair. Un plan fractionnaire de résolution R sera noté 2^{K-q}_R, R étant écrit en chiffres romains. Dans notre exemple, le plan 2^{6-3}, de résolution $\mathcal{R}=3$, est noté 2^{6-3}_{III}. Aborder l'interaction d'ordre 2, et donc utiliser un plan de résolution V, est pour nous un minimum dans l'analyse d'une FC. La relation entre le nombre K de facteurs et la résolution d'un plan fractionnaire permet d'établir le coût en nombre de simulations N :

- si $\mathcal{R}=III$ alors $N \geq 1+K$,
- si $\mathcal{R}=IV$ alors $N \geq 2K$,
- si $\mathcal{R}=V$ alors $N \geq 1+K+K(K-1)/2$.

D'autre part, le nombre *maximal* K_{max} de facteurs d'un plan de résolution V nécessitant $N = 2^s$ simulations et donnant des estimateurs de variance minimale, est indiqué dans la table suivante :

s	4	5	6	7	8	9
N	16	32	64	128	256	512
K_{max}	5	6	8	11	17	≥ 23

1.4.6 Modèle statistique d'un plan fractionnaire

Le choix de la définition des effets dans un plan fractionnaire fut motivé dans le passé par la facilité des calculs et de l'interprétation. Le souci d'une interprétation aisée reste aujourd'hui toujours valable. Le modèle statistique associé au plan fractionnaire peut être défini à partir des effets factoriels spécifiques définis plus haut. Soit un plan comprenant trois facteurs A B et C à deux niveaux -1 et $+1$. L'effet d'une combinaison des facteurs est noté de façon générale $e(A^a, B^bC^c)$ avec $a, b, c \in \{0, 1\}$. Par exemple, l'effet d'interaction $e(AB)$ correspond à $a = b = 1$ et $c = 0$. Cette notation permet d'établir la réponse moyenne $G(A, B, C)$ sous la forme :

$$G\,(A, B, C) = \sum_{a,b,c} A^a B^b C^c e(A^a B^b C^c)$$

$$= \mu + Ae(A) + Be(B) + Ce(C) + ABe(AB) + ACe(AC) + BCe(BC) + ABCe(ABC)$$

Pour $(A, B, C) = (+1, -1, +1)$ on obtient :

$$G\,(+1, -1, +1) = \mu + e(A) - e(B) + e(C) - e(AB) + e(AC) - e(BC) - e(ABC)$$

De façon générale, un effet est défini à partir des colonnes A, B et C de la matrice d'incidence à l'aide du produit scalaire

$$e(A^a B^b C^c) = \frac{1}{N}\langle A^a B^b C^c, G\rangle$$

, N désignant la taille du plan. Dans l'exemple, on a pour l'effet de A :

$$e(A) = \frac{1}{2}(\bar{G}_{+\bullet\bullet} - \bar{G}_{-\bullet\bullet}) = \frac{1}{8}[(G_{+--} + G_{+-+} + G_{++-} + G_{+++}) - (G_{---} + G_{--+} + G_{-+-} + G_{-++})] = \frac{1}{8}\langle A, G\rangle$$

L'effet de l'interaction entre A et B est défini par la différence des effets de A aux deux niveaux de B :

$$2e(AB) = e(A|B = +1) - e(A|B = -1)$$

d'où

$$e(AB) = \frac{1}{2}\left(\frac{1}{4}(G_{++-} + G_{+++} - G_{-+-} - G_{-++}) - \frac{1}{4}(G_{+--} + G_{+-+} - G_{---} - G_{--+})\right)$$

$$= \frac{1}{8}\langle AB, G\rangle$$

On peut aussi utiliser le modèle classique d'anova rencontré dans le cas d'un plan complet qui donnera des résultats équivalents. Ce modèle classique permettra de plus de calculer les indices de sensibilité.

Remarque : les plans fractionnaires sont générés avec le logiciel SAS au moyen de la procédure factex.

1.4.7 Plan fractionnaire et pavage

On peut bien évidemment construire un pavage de l'espace des facteurs à l'aide d'un plan fractionnaire. Cela permet de ne négliger a priori aucune zone de l'espace. Plutôt que de faire des répétitions dans chaque pavé, le remplissage de l'espace sera amélioré à l'aide de tirages de plusieurs plans fractionnaires en affectant au hasard les facteurs au plan fractionnaire de référence. Comme dans le cas du plan complet, l'inférence pourra être effectuée avec une méthode de ré-échantillonage.

1.5 Méthode de Morris

L'exploration d'une FC basée sur un modèle d'anova est coûteuse ou bien nécessite des hypothèses sur l'ordre des interactions prises en compte. De plus, les plans fractionnaires, du fait de contraintes algébriques, ne peuvent être définis dans toutes les configurations de FC. Pour justifier des hypothèses ou bien faire un tri parmi les facteurs, une étude préalable de la FC peut s'avérer utile. La méthode de Morris permet de répondre à cet objectif pour un coût de calcul somme toute acceptable. La méthode de Morris est présentée dans sa forme originelle qui consiste à analyser la FC déterministe sur une grille discrète superposée sur l'espace des facteurs. Des variantes seront signalées qui prennent en compte la continuité de l'espace Ω.

1.5.1 Plan d'échantillonnage de Morris

La méthode est basée sur l'étude des variations de la sortie, les facteurs variant successivement. A la différence des méthodes qui font varier les facteurs un par un autour d'un point de référence, la méthode de Morris est globale. Soit une FC comprenant K facteurs continus et définie sur l'hypercube $\Omega = [0,1]^K$. Une transformation linéaire de la gamme $[0,1]$ permettra un retour éventuel aux unités des facteurs. L'espace des facteurs Ω est exploré aléatoirement et de façon uniforme. De plus, les lois d'échantillonnage des facteurs sont supposées indépendantes entre elles. La premère étape consiste à discrétiser

la gamme de variation de chaque facteur en Q niveaux notés $\{0, \frac{1}{Q-1}, \frac{2}{Q-1}, .., 1\}$. Le croisement de tous ces niveaux définit une grille $\widehat{\Omega}$ qui se superpose sur Ω. L'échantillonnage consiste à générer une suite de N trajectoires $\mathcal{T}^{(i)}$, $i = 1, N$ dans $\widehat{\Omega}$. Le point de départ $A_0^{(i)}$ de la trajectoire $\mathcal{T}^{(i)}$ est choisi au hasard dans $\widehat{\Omega}$. Une trajectoire est composée de $K+1$ points $(A_0^{(i)}, .., A_K^{(i)})$ définis en faisant varier sur la grille chacun des K facteurs d'un pas $\delta \propto \frac{1}{Q-1}$. L'ordre des facteurs dans une trajectoire est aléatoire (Fig.7).

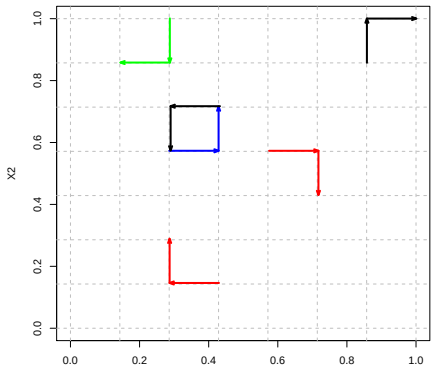


Figure 7 : Chaque dimension du carré $\Omega = [0, 1]^2$ est discrétisée en $Q = 7$ niveaux. $N = 6$ trajectoires aléatoires $\mathcal{T}^{(i)}$, $i = 1, 6$, passent par les noeuds de la grille régulière $\widehat{\Omega}$ superposée sur le carré. Le pas de déplacement est $\delta = 1/7$ dans chaque dimension. La direction de la flèche indique le sens du déplacement.

Pour assurer une distribution uniforme des niveaux échantillonnés des facteurs (Fig. 8), la règle suivante sera utilisée pour définir la grille et les pas δ des déplacements. Le niveau de discrétisation Q est *pair* et le pas de déplacement δ vaut $\frac{Q}{2(Q-1)}$.

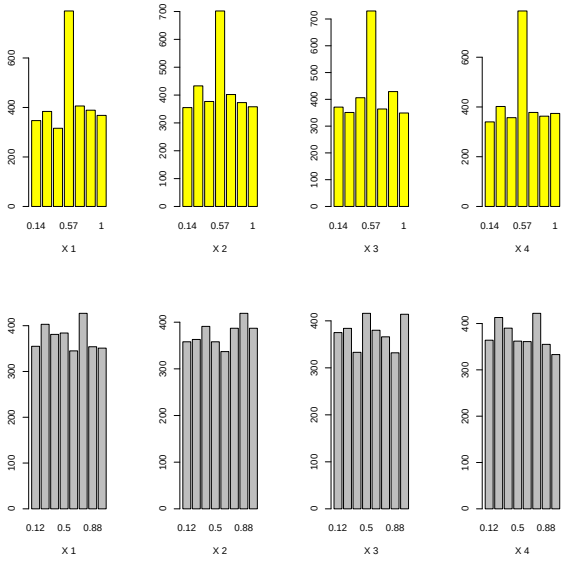


Figure 8 : diagramme des fréquences des valeurs des coordonnées des points obtenues avec un plan de Morris à 4 facteurs. La ligne du haut (en jaune) contient les diagrammes des fréquences des quatre coordonnées X_i, $i = 1, 4$, pour $Q = 7$, $\delta = 3$ et 1000 trajectoires. La ligne du bas (en gris) contient les diagrammes des quatre coordonnées pour $Q = 8$, $\delta = 4$ et 600 trajectoires. Cette deuxième simulation correspond à la règle qui garantit une distribution uniforme des quatre coordonnées X_i.

1.5.2 Indices de sensibilité

La méthode de Morris repose sur l'étude des variations de la sortie de la FC pour des variations des facteurs sur la grille. Soit la trajectoire $\mathcal{T}^{(i)}$. Le déplacement entre deux points successifs d, notés $A_j^{(i)}$ et $B_j^{(i)}$, e $\mathcal{T}^{(i)}$ correspond à la variation d'un pas δ du facteur X_j. Cette variation du facteur X_j induit une variation $\Delta_j^{(i)}G$ de la FC. La direction du déplacement de longueur $\delta \times \frac{1}{Q}$ est prise en compte au moyen du signe $\pm$ dans la définition suivante :

$$\Delta_j^{(i)}G = \frac{G[A^{(i)}] - G[B^{(i)}]}{\delta} = \frac{G[.., X_j^{(i)} \pm \delta \times \frac{1}{Q}, ..] \; - \; G[.., X_j^{(i)}, ..]}{\delta}$$

L'unité de cette variation est celle de la sortie de la FC car δ est sans dimension (contrairement au cas de la dérivé). Un modèle d'anova aidera à l'interprétation de cette quantité. Soient une FC à deux facteurs $G(X, Y)$ dont un modèle d'anova s'écrit $G_{ij} = M + X_i + Y_j + X : Y_{ij}$, en notant X_i, Y_j et $X : Y_{ij}$ les effets associés aux facteurs discrétisés. Le facteur Y restant au niveau j, la variation de G suite à la variation de X entre deux modalités i et i', est égale (au résidu éventuel près) à $X_{i'} + X : Y_{i'j} - X_i - X : Y_{ij}$.

L'indice μ_j relatif au facteur X_j proposé par Morris consiste à faire la moyenne des effets élémentaires :

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} \Delta_j^{(i)}G$$

Si la relation entre X_j et G est linéaire et si X_j n'interagit pas avec les autres facteurs, alors la moyenne des effets élémentaires est proportionnelle à l'intensité de la liaison. En présence d'interaction, l'effet moyen est un mélange des effets principaux et d'interaction. Cet indice présente une analogie avec l'indice total rencontré plus haut. En cas de périodicité de la FC, un pas Δ correspondant à une demie période de la FC peut induire un indice μ_j nul ou très faible. De plus, si la FC n'est pas monotone, les variations élémentaires vont se compenser en partie, d'où un indice μ également peu élevé. Une interprétation rapide en déduirait l'absence d'effet d'un tel facteur. Pour éviter ce phénomène, on préfère utiliser l'indice $\mu^\star$ basé sur les valeurs absolues des variations

élémentaires. Soit donc l'indice $\mu_j^\star$ pour le facteur X_j :

$$\mu_j^\star = \frac{1}{N}\sum_{i=1}^{N} |\Delta_j^{(i)} G|$$

Cet indice présente comme μ_j une analogie avec l'indice de sensibilité total.

L'autre indice proposé par Morris est défini à partir de la variance des variations élémentaires. Soit, pour le facteur j :

$$\sigma_j = \sqrt{\frac{1}{N-1}\sum_{k}\left(\Delta_j^{(i)} G - \mu_j\right)^2}$$

L'écart type σ_j sera utilisé pour caractériser la présence d'interaction de X_j avec d'autres facteurs. L'indice n'indique pas toutefois avec quel facteur X_j interagit.

Une forte relation non-linéaire peut induire un $\mu_j^\star$ élevé mais aussi une forte valeur pour l'indice σ_j. L'interprétation d'une interaction de X_j avec d'autres facteurs de la FC est dans ce cas erronée (Fig. 9).

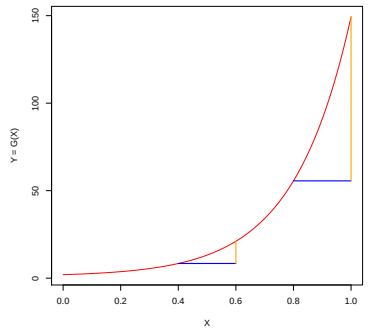


Figure 9 : exemple de relation induisant un indice σ_j non nul sans présence interaction.

Le diagramme de corrélation entre un facteur et la sortie de la FC permet ici aussi de préciser l'interprétation. Une synthèse graphique des indices utile pour l'interprétation consiste à représenter les points $(\mu_j^\star, \sigma_j)$ dans le plan (Fig 10).

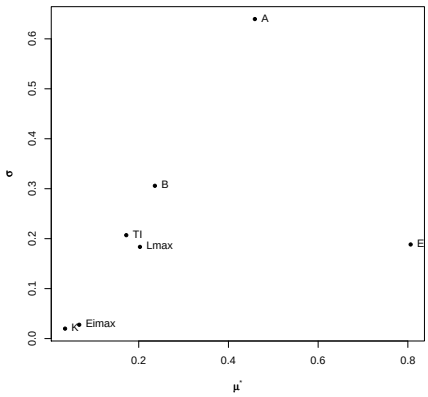


Figure 10 : sortie graphique de la méthode de Morris pour la sortie d'une FC à 7 facteurs (A, B, E, TI, Lmax, K, Eimax). Un point a pour coordonnées le couple $(\mu_j^\star, \sigma_j)$ des indices de Morris pour le facteur j (cf texte). Les facteurs K et $Eimax$ n'ont pas d'effet et n'interagissent pas. Le facteur A présente un effet moyen sensible ainsi qu'une forte interaction. Le facteur E a plutôt un effet moyen et interagit moins fortement avec les autres. Les facteurs B, TI et $Lmax$ sont dans une situation intermédiaire.

On peut proposer à partir du graphique de Morris une typologie des facteurs en trois groupes :

- facteurs ayant un effet négligeable (points proches de l'origine)
- facteurs ayant un effet linéaire important (points situés à droite sur l'axe des abscisses)
- facteurs ayant un effet non-linéaire ou interagissant (points situés en haut à droite)

Le coût de la méthode de Morris en nombre de simulations de la FC est $N(K+1)$.

1.5.3 Inférence

La méthode de Morris n'informe pas sur la variabilité des indices obtenus sur un échantillon de trajectoires. Il est possible de calculer les intervalles de confiance des indices sous l'hypothèse d'indépendance des variations élémentaires. Cette hypothèse est vérifiée car les trajectoires sont échantillonnées de façon indépendante. L'intervalle de confiance contenant la vraie valeur de $\mu_j^\star$ avec la probabilité $(1-\alpha)$ est obtenu par la méthode de Student :

$$\widehat{\mu^\star}_j \pm T_{N-1}^{(1-\alpha/2)} \widehat{\sigma}_j^* / \sqrt{N}$$

avec $\widehat{\sigma}_j^*$ estimation de l'écart type des valeurs absolues des variations élémentaires et $T_{N-1}^{(\alpha)}$ quantile de la loi de Student à $N-1$ degrés de liberté. L'intervalle de confiance de

probabilité $(1-\alpha)$ pour σ_j est donné par :

$$\left[\sqrt{\frac{(N-1)}{\chi^2(N-1,\alpha/2)}}\,\widehat{\sigma}_j\,,\ \sqrt{\frac{(N-1)}{\chi^2(N-1,1-\alpha/2)}}\,\widehat{\sigma}_j\right]$$

avec $\chi^2(N-1,\alpha)$ quantile d'une loi du χ^2 à $N-1$ degrés de liberté.

Le calcul des intervalles de confiance peut aussi être réalisé au moyen de la procédure de re-échantillonnage bootstrap. Elle consiste à simuler N_b plans de Morris en tirant avec remise N trajectoires parmi les trajectoires du plan de Morris originel. Les quantiles des distributions des N_b valeurs des indices ainsi obtenues permettent de construire un intervalle de confiance de probabilité $(1-\alpha)$.

Il est difficile de déterminer a priori le nombre de trajectoires N nécessaire pour avoir une analyse fiable sur le plan statistique. Le nombre N dépend de la complexité de la FC et de la précision souhaitée sur les indices. Si les intervalles de confiance sur les indices ont une amplitude trop grande, un tirage supplémentaire de trajectoires sera nécessaire. La mise en oeuvre de cette méthode sera présentée dans le paragraphe ci-dessous consacré à un exemple d'analyse d'un modèle de culture comprenant sept facteurs.

1.5.4 Analyse exploratoire

L'analyse de Morris peut être utilement complétée par une analyse exploratoire des simulations. Par exemple, l'analyse graphique des distributions des variations élémentaires permettra de détecter des zones de fortes variations. De plus, l'examen des diagrammes de corrélation associé à une estimation non-paramétrique de la courbe de régression (cf lowess sous le logiciel R) permettra d'inférer une forme sur la liaison entre sortie de la FC et facteurs. Une analyse de variance sur les sorties de la méthode Morris est aussi possible à partir des résultats de simulation utilisés par la méthode. Mais le plan ainsi construit a de fortes chances d'être déséquilibré voire incomplet pour certaines combinaisons des facteurs. Une procédure d'anova adéquate devra donc être mise en oeuvre. Elle permettra d'identifier dans une certaine mesure les interactions dont on pourra effectuer une représentation graphique.

1.5.5 Distribution d'échantillonnage non uniforme

Une adaptation simple de la méthode de Morris permet d'aborder le cas d'une distribution d'échantillonnage non uniforme des facteurs. Cela consiste à découper les gammes de variation des facteurs à partir des quantiles des distributions non uniformes en jeu. Des intervalles equi-probables mais d'amplitudes différentes sont ainsi définis. La règle de Morris qui garantit une distribution uniforme des points des trajectoires dans le cas

usuel garantira, par le jeu de la transformation utilisée, la distribution souhaitée pour les coordonnées des points placés sur les trajectoires. Dans cette adaptation, deux variations égales de la sortie peuvent correspondre à des variations équi-probable mais d'amplitudes différentes d'un facteur.

1.5.6 Analyse de Morris du modèle jouet

L'analyse est réalisée avec un niveau de discrétisation relativement élevé et égal à 16 afin d'augmenter le nombre potentiel de trajectoires et éviter les doublons. On compare les intervalles de confiance obtenues par bootstrap avec 15, 40 et 100 trajectoires (table 8). La méthode montre bien la présence d'interaction entre X et Y.

Indices		X	Y
$\mu^\star$	b_1	2.97	2.27
	b_2	6.35	5
σ	b_1	3.83	2.5
	b_2	7.23	6.12

indices		X	Y
$\mu^\star$	b_1	3.95	2.7
	b_2	6.04	4.4
σ	b_1	4.83	3.4
	b_2	7.2	5.3

indices		X	Y
$\mu^\star$	b_1	3.71	3.05
	b_2	4.76	4.11
σ	b_1	4.2	3.39
	b_2	5.21	4.45

Table 8 : intervalles de confiance $[b_1, b_2]$ de probabilité 0.95 des indices de Morris calculés selon la méthode bootstrap sur le modèle jouet à 2 facteurs X et Y. Les nombres de trajectoires sont 15 (table à gauche), 40 (table au milieu) et 100 (table à doite) trajectoires.

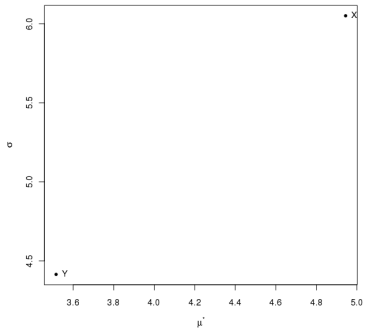


Figure 11 : graphique de Morris pour le modèle jouet à deux facteurs X et Y (cf texte) : discrétisation des gammes en 16 intervalles, 40 trajectoires.

1.5.7 Variantes

Le plan proposé par Morris définit les trajectoires sur les noeuds d'une grille. De plus, la direction de la variation d'un facteur est parallèle à l'axe attribué au facteur. Une variante (Réf.) prend en compte la continuité de l'espace de facteurs et ne privilégie par

des directions. Le point de départ d'une trajectoire est tiré dans l'espace des facteurs. Les autres points de la trajectoire sont définis par les sommets d'un triangle en dimension deux (Fig. 12), d'un tétrahèdre en dimension 3 ou d'un simplex au delà. Le simplex s est tiré au hasard. Un polynôme du premier degré passant par les $K+1$ points $(A_{s,l}, G(A_{s,l}))$, $l = 0, K$, avec $A_{s,l} = (X_1^{(s,l)}, .., X_K^{(s,l)})$ est défini par le système de $K+1$ équations à $K+1$ inconnues suivant

$$G(X_1^{(s,1)}, .., X_K^{(s,l)}) = \theta_0^{(s)} + \sum_{j=1}^{K} \theta_j^{(s)} \, X_j^{(s,l)} \quad , l = 0, K+1$$

La variation ΔG vérifie $\Delta G = \sum_{j=1}^{K} \theta_j \Delta X_j$, avec ΔX_j variation élémentaire du facteur. L'effet élémentaire du facteur X_j pour le simplex s est défini par le coefficient $\theta_j^{(s)}$. Les indices de Morris seront calculés à partir de ces coefficients.

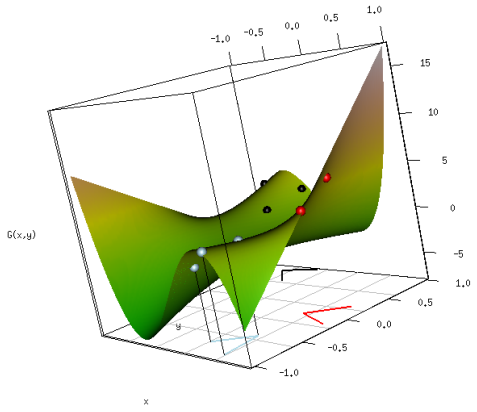


Figure 12 : l'espace des facteurs est le carré. Trois triangles sont obtenus par tirage au hasard d'un point de départ et par transformation d'un triange de référence par une rotation d'angle aléatoire.

La définition des effets élémentaires suppose l'absence d'interaction entre les facteurs. D'où l'importance d'une détermination adéquate de la taille du simplex pour assurer la validité de cette hypothèse. L'approximation de la FC par un modèle linéaire de degré un est souvent suffisante localement : ce n'est pas en général vrai dans ce cas et encore moins à plus grande échelle.

Une autre piste simple résulte de la combinaison de la méthode classique de Morris et du pavage de Ω. Soit la discrétisation utilisant les principes de la méthode de Morris. Le pavage de Ω en est déduit. avec R répétitions par pavé. L'effet élémentaire est alors

défini à partir des valeurs moyennes de la sortie de la FC sur un pavé. Les indices sont définis comme dans la version standard de la méthode de Morris. Cette variante permet de n'exclure a priori aucun point de l'espace des facteurs. De plus on peut utiliser dans chauqe un échantillonnage qui a de meilleures propriétés de remplissage que l'échantillonnage aléatoire usuel. Par exemple un hyper cube latin (cf chapitre ?) de K points dans chaque pavé peut jouer ce rôle. Le souci d'analyser le comportement de la FC plus finement peut conduire à générer dans chaque pavé un plan fractionnaire comprenant K facteurs et ayant une résolution V. Le coût de cette méthode est $R \times N \times (K + 1)$. La première variante est diffusée dans la bibliothèque "sensitivity" de R. La deuxième est une proposition non publiée de l'auteur.

1.6 Exemple d'analyse de sensibilité d'une FC avec le logiciel R

L'objet de l'analyse est le modèle wwdm à 7 facteurs utilisé lors de l'Ecole chercheurs Mexico-INRA. On trouvera sa description dans les documents fournis aux stagiaires. Il s'agit de determiner les facteurs influents sur une variable calculée sur les sorties de la FC. Les trois méthodes présentées dans ce paragraphe seront mises en oeuvre. Comme ce document est envoyé avant le déroulement de l'Ecole chercheurs 2010 (un bon cru selon les personnes bien informées !), nous ne voulons pas totalement priver le lecteur de la découverte des méthodes et de leur mise en oeuvre lors de l'Ecole. Les sections qui suivent resteront donc blanches pour l'instant. A l'issue du stage, l'auteur se fera un plaisir de complèter les parties manquantes en tenant compte des observations des participants de l'Ecole.

1.6.1 Présentation de la FC

1.6.2 Analyse par Anova et plan complet

1.6.3 Analyse par Anova et plan fractionnaire

Génération d'un plan fractionnaire sous R

La génération du plan est effectuée à l'aide de la fonction regular.fraction du logiciel R. La recherche du plan avec cette fonction n'est pas automatique dans cette version du code. Elle devra être effectuée par essai erreur. La définition des indices de sensibilité est basée comme ce fut le cas pour l'anova d'un plan complet sur la décomposition de la somme des carrés obtenus avec la fonction aov.

```
# arguments de regular.fraction:
# Ex: K = 6 facteurs , 2 niveaux, r=3 tq N = p^r unités, résolution R=3
> plan = regular.fraction(6,2,3,3)$plan
> fich = data.frame(A=as.factor(plan[,1]),
           B=..., F=as.factor(plan[,6]),
           Y=c(.367, .532,.495, .489, .310, .485, .476, .440))
> fich.aov = aov(Y ~ A + B + C + D +E + F, fich)
> summary(fich.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	1	0.0037	0.0037	18.49	0.15
B	1	0.0053	0.0053	26.52	0.12
C	1	0.0111	0.0111	55.5	0.085
D	1	0.00016	0.00016	0.8	0.53
E	1	0.00005	0.00005	0.25	0.70
F	1	0.018	0.018	91.203	0.066
Residuals	1	0.0002	0.0002		

Table : anova du plan fractionnaire à 6 facteurs et de résolution III.

1.6.4 Analyse par Morris

```
Q=6
delta = Q/(2*(Q-1)))
morris.out = morris(model = wwdm.simule,
        factors= noms , r = 30,
        design = list(type = "oat", levels = Q,
        grid.jump = delta*(Q-1), scale=F)
plot(morris.out)
```

1.7 Conclusion du chapitre

Les méthodes présentées permettent dans une certaine mesure de détecter les facteurs influents et de déterminer la présence d'interactions. Le rôle du plan d'échantillonnage joue un rôle fondamental. Un certain empirisme est présent dans la phase de discrétisation ou du choix de niveaux des facteurs. La procédure d'échantillonnage et l'inférence sur les indices permettent toutefois de garantir une certaine validité statistique des résultats dans la mesure où la taille de l'échantillon est suffisante et la qualité de remplissage correcte. Ce dernier point est plus délicat à quantifier (Réf.) et on en mesure difficilement les conséquences. L'exemple du modèle jouet a permis de porter l'attention sur les frontières du domaine de définition de facteurs. Ces frontières peuvent correspondre à des zones de

plus fortes incertitudes dans le cas de modèles statistiques mieux estimés au centre de la gamme d'un facteur. L'extrapolation sur les frontières d'un modèle de degré realitevement élevé peut alors produire des artefacts. C'est aussi le rôle de l'analyse de sensibilité de la FC de détecter ce phénomène indésirable quitte à modifier, si nécessaire, le code. Dans la mesure du possible, plutôt que de se limiter a priori sur une méthode, il nous semble nécessaire d'utiliser de façon conjointe plusieurs méthodes afin de profiter des qualités propres à chacune. Dans une pratique réelle, l'analyse de sensibilité n'est pas forcément immédiate et automatique. La définition des gammes de variation des facteurs pose déjà question. C'est une hypothèse fondamentale sur la variation des facteurs qui peut induire des conclusions différentes dans les analyses de sensibilité. L'utilsateur du modèle linéaire s'en convaincra aisément. Comment quantifier cette incertitude et son influence ? L'objectif choisi, souvent caractérisé par un petit nombre de variables de sortie de la FC, ne fournit qu'un aspect parmi beaucoup d'autres. Tel Sisyphe condamné par les Dieux à pousser son rocher au gré des montagnes jusqu'à la fin des Temps, le modélisateur (on espère dans la joie) risque fort de devoir réaliser de nouvelles analyses de sensibilité face à la multiplicité des objectifs et tant que les Dieux (entre autres) prêteront vie au modèle. Les méthodes statistiques présentées dans ce chapitre tournent autour de l'étude des moments d'ordre deux d'une sortie. Que faire pour analyser les valeurs extrêmes d'une FC ? Une valeur extrême dans la sortie de la FC peut correspondre à des erreurs de programmation ou de définition du modèle mais aussi à des phénomènes aux conséquences importantes. Il faudra alors se tourner vers des méthodes d'investigation dont les principes sont moins connus hors de la communauté des statisticiens. Nous avons insisté en introduction sur l'importance d'une démarche critique qui est un des fondements de la méthode scientifique. La critique s'applique aussi à l'analyse de sensibilité dont la qualité dépendra des points de vue complémentaires apportés par les méthodes mises en oeuvre. De plus, la possibilité de bien identifier les hypothèses sous-jacentes à la méthodologie statistique ouvre la porte, dans un contexte scientifique, à de nouvelles investigations et remises en cause ainsi qu'à la modestie des interprétations. C'est donc, selon nous, un bon moyen de faire progresser le questionnement autour des modèles de simulation numérique et faire accepter le raisonnement de l'incertain. Ce raisonnement est adapté aux phénomènes réels qui eux comportent de façon irréductible des caractéristiques incertaines, notamment dans les sciences de l'environnement.

Bibliographie

[1] E.M. Scott A. Saltelli, K. Chan. *Sensitivity Analysis*. Wiley, 2000.

[2] Nicolas Bouleau. *Philosophie des mathématiques et de la modélisation : du chercheur à l'ingénieur*. L'Harmattan, 2000.

[3] Rémy Hordan Bradley Efron, Emmanuel Jolivet. *Le bootstrap et ses applications*. CISIA, 1995.

[4] Petre Enciu. *Dérivation automatique pour le calcul des sensibilités appliqué au dimensionnement en génie électrique*. PhD thesis, Grenoble, 2009.

[5] A. Saltelli et al. *Global Sensitivity Analysis, the primer*. Wiley, 2008.

[6] Michel Armatte et Amy Dahan Dalmedico. Modèles et modélisations 1950-2000 : nouvelles pratiques, nouveaux enjeux. *Rev. Hist. Sci.*, 57 2 :245–305, 2004.

[7] W.G. Hunter G.E.P. Box, J. Stuart Hunter. *Statistics for Experimenters, Desig, Innovation and Discovery*. Wiley, 2005.

[8] J. Goupy. *La méthode des plans d'expériences*. Dunod, 1988.

[9] Gilbert Saporta éditeurs Jean-Jacques Droesbeke, Jeanne Fine. *Plans d'expériences, applications à l'entreprise*. Technip, 1997.

[10] Myanna Lahsen. Seductive simulations ? uncertainty distribution around climate models. *Social Studies of Science*, 35/6 :895–922, 2005.

[11] Bryan F.J. Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd edition*. Chapman & Hall, 1997.

[12] W.J. Conover Ronald L. Iman. A distribution-free approach to inducing rank correlation among input variables. *Commun, Statist.-Simula. Computa.*, 11(3) :311–334, 1982.