REVUE DE STATISTIQUE APPLIQUÉE

A. JOURDAN

Approches statistiques des expériences simulées

Revue de statistique appliquée, tome 50, n° 1 (2002), p. 49-64

<http://www.numdam.org/item?id=RSA_2002__50_1_49_0>

Rev. Statistique Appliquée, 2002, **L** (1), 49-64

APPROCHES STATISTIQUES DES EXPÉRIENCES SIMULÉES

A. JOURDAN

Institut Français du Pétrole, 92852 Rueil-Malmaison, France
astrid.jourdan@ifp.fr

RÉSUMÉ

De nombreux phénomènes physiques sont étudiés à l'aide de simulateurs très coûteux. L'intérêt des approches statistiques présentées dans ce travail est alors de prédire la réponse du simulateur à partir de quelques simulations. Il existe essentiellement deux méthodes statistiques pour aborder les expériences simulées. Après une rapide présentation et analyse de chacune d'elles, nous proposons une nouvelle approche statistique intégrant les points forts des deux approches déjà existantes et qui a pour objectif d'en atténuer les inconvénients.

Mots-clés : *Expériences simulées, Krigeage, DACE, Plans d'échantillonnage.*

ABSTRACT

Many Scientific phenomena are investigated by expensive computer models. The interest of the statistics approaches presented in this work is to predict the response of the computer model from a few runs of the code. The two main statistics approaches for computer experiments are reviewed. We propose our own approach for computer experiments which includes the advantages offer by the two existing methods and fill their gaps.

Keywords : *Computer experiments, Kriging, DACE, Sampling designs.*

1. Introduction

Cet article traite des expériences simulées, c'est-à-dire menées à l'aide d'un code informatique, et plus particulièrement de la prédiction de la réponse d'un simulateur à partir d'un petit nombre de simulations. Ce sujet devient d'une grande importance pratique dans toutes les disciplines où il est fait appel à des modèles déterministes complexes. Les phénomènes physiques ainsi modélisés sont ensuite étudiés à l'aide de simulateurs dont, d'une part, la fiabilité est mise en question du fait même de la complexité des modèles, et dont d'autre part, les temps de calcul sont souvent prohibitifs. L'utilisateur souhaite alors disposer d'un modèle simple et rapide pour résumer la réponse du simulateur.

Koehler et Owen (1996) présentent les outils statistiques utilisés pour l'analyse et la planification des expériences simulées. Ils opposent notamment deux types d'approches statistiques. La première utilise des techniques de prédiction linéaire sans biais, *i.e.* le krigeage des géostatisticiens (puis de prédiction bayésienne). La

deuxième quant à elle, consiste à adapter à la prédiction des techniques d'intégration numérique de fonctions simulées par méthodes de Monte Carlo ou Quasi-Monte Carlo. Ces deux types d'approches s'articulent autour des questions suivantes. Quel modèle peut-on proposer pour la réponse du simulateur qui tienne compte du caractère déterministe des expériences simulées? Où effectuer les simulations afin de récolter un maximum d'information pour un coût minimal?

La première partie (sections 2 et 3) présente brièvement ces deux approches classiques des expériences simulées. L'objectif ici est de mettre en évidence, à partir d'arguments empiriques, les avantages de chacune d'elles tout en précisant certaines de leurs limites. Cette première partie nous amène tout naturellement à proposer une nouvelle approche statistique des expériences simulées (section 4) qui permet d'intégrer les points forts des deux approches classiques, avec pour but, d'atténuer les faiblesses de chacune d'elles et ainsi repousser certaines de leurs limites.

2. Approche par résidu aléatoire

Cette première approche statistique des expériences simulées est la plus couramment utilisée. Elle a été proposée par Sacks, Welch, Mitchell et Wynn (1989) et Sacks, Schiller et Welch (1989).

2.1. Modélisation et prédiction

Supposons que les d paramètres d'entrée du simulateur varient entre 0 et 1. Alors pour tout $x \in [0,1]^d$, la réponse déterministe du simulateur, $y(x)$, est considérée comme la réalisation d'une fonction aléatoire, $Y(x)$, qui se décompose en deux parties :

$$Y(x) = X(x)\beta + \Gamma(x),\ \forall x \in [0,1]^d, \tag{1}$$

• $X(x)\beta = E[Y(x)]$ est la partie déterministe du modèle. Il s'agit en fait d'une régression linéaire classique, où $X(x)$ est le vecteur formé des m fonctions de la régression et β est le vecteur inconnu des paramètres du modèle,

• $\Gamma(x)$ est la partie aléatoire du modèle. La méthode statistique traditionnelle consiste à considérer $\Gamma(x)$ comme un bruit blanc correspondant à une erreur de mesure. Cependant, dans le cas d'un simulateur déterministe, les expériences sont nécessairement sans erreur, $\Gamma(x)$ représente alors l'écart systématique entre la réponse du code et la régression linéaire présumée (*cf.* figure 1). $\Gamma(x)$ est appelé le ***processus résiduel***. Il est choisi tel que $e[\Gamma(x)] = 0$ et $\mathrm{cov}(\Gamma(x), \Gamma(u)) = \sigma^2 R(x,u)\ \forall x \in [0,1]^d$, $\forall u \in [0,1]^d$ où σ^2 est la variance et R est la ***fonction de corrélation*** du processus.

Pour analyser un tel modèle, une possibilité consiste à utiliser une méthode de krigeage (Matheron (1963)). Soit $D = \{x_1, ..., x_N\}$ un plan d'échantillonnage, où N représente le nombre de simulations à effectuer et les x_i l'endroit du cube unité où on effectue ces simulations. On suppose que l'on dispose des réponses du simulateur aux points du plan que l'on note dans un vecteur Y $(N \times 1)$. On peut alors construire

le meilleur prédicteur linéaire sans biais (BLUP) (Sacks Welch, Mitchell et Wynn (1989))

$$\widehat{Y}(x) = X(x)\widehat{\beta} + {}^t r(x) R^{-1} \lfloor Y - X\widehat{\beta} \rfloor, \ \forall x \in [0,1]^d \tag{2}$$

- $X = (X(x_i))_{i=1,\ldots,N} : N \times m$ ***matrice du modèle*** aux points du plan
- $R = (R(x_i, x_j))_{i,j=1,\ldots,N} : N \times N$ ***matrice de corrélation***, $\mathrm{cov}(Y) = \sigma^2 R$, supposée régulière
- $r(x) = {}^t[R(x_1, x), \ldots, R(x_N, x)] : N \times 1$ ***vecteur de corrélation***, $\mathrm{cov}(Y(x), Y) = \sigma^2 r(x)$.

où $\widehat{\beta} = ({}^t X R^{-1} X)^{-1} {}^t X R^{-1} Y$, l'estimateur de Gauss-Markov de β, existe si et seulement si les colonnes de la matrice du modèle sont indépendantes.

Le prédicteur minimise donc l'erreur quadratique moyenne qui est égale à, $\forall x \in [0,1]^d$

$$\mathrm{MSE}(x) = E[Y(x) - \widehat{Y}(x)]^2 = \sigma^2 (1 - {}^t r(x) R^{-1} r(x) + K(x) ({}^t X R^{-1} X)^{-1} {}^t K(x)),$$

où $K(x) = \lfloor X(x) - {}^t r(x) R^{-1} X \rfloor$ (Christensen (1990)).

La matrice de corrélation R et le vecteur de corrélation $r(x)$ permettent de prendre en compte le fait que plus le point x (où l'on cherche à prédire la réponse du simulateur) est loin des points de simulation x_i, et plus l'erreur de prédiction est grande.

La caractéristique première du prédicteur vient du fait qu'il interpole les réponses du simulateur aux points du plan :

$$\forall x_i \in D \ \widehat{Y}(x_i) = Y(x_i).$$

Ce comportement est essentiellement dû au processus résiduel. En effet, on peut remarquer que le prédicteur (2) se décompose en deux parties : un ajustement en moyenne classique plus un terme de correction dans lequel interviennent la matrice et le vecteur de corrélation. Comme on peut le constater sur la figure 1, c'est ce terme qui permet de corriger l'ajustement en moyenne de façon à atteindre l'interpolation.

Cette caractéristique du prédicteur semble *a priori* intéressante puisqu'il est légitime de vouloir interpoler les observations dans le cas d'un phénomène déterministe. Il est cependant à noter que cette contrainte d'interpolation devient très forte lorsque le nombre d'observations augmente, et entraîne notamment des problèmes d'irrégularités du prédicteur.

Dans le modèle (1) plusieurs fonctions sont à fixées *a priori* : la régression linéaire et la fonction de corrélation. Il semble clair d'après ce qui précède, que le processus résiduel joue un rôle prépondérant dans le modèle alors que la partie déterministe n'a qu'une contribution mineure. En conséquence, la régression linéaire est souvent en pratique fixée constante afin de simplifier les calculs (Welch *et al.* (1992)), alors que la fonction de corrélation est choisie de façon à refléter au mieux la nature présumée du simulateur, parmi différentes familles de fonctions, chacune

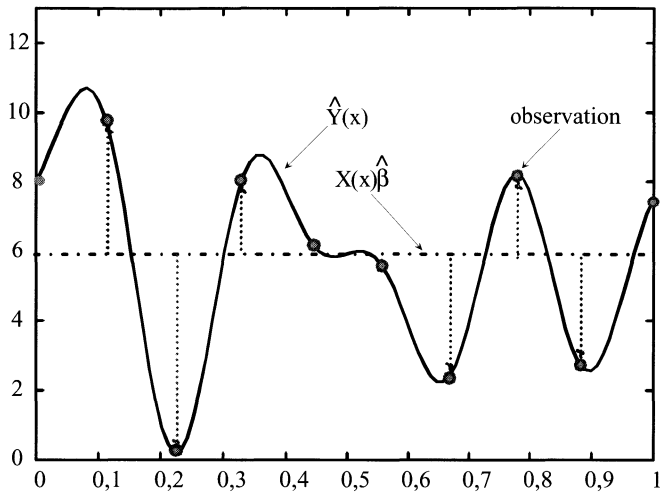

FIGURE 1

Prédicteur et ajustement en moyenne construits à partir de 10 observations correspondant à 10 réponses présumées d'un simulateur à 1 paramètre d'entrée. La régression linéaire est fixée par $X(x) = 1$ (d'où un ajustement en moyenne constant) et le processus résiduel est gaussien.

ayant ses particularités (Sacks, Welch, Mitchell et Wynn (1989), Christensen (1990) ou Koehler et Owen (1996)). Dans cet article on suppose que le processus résiduel est stationnaire gaussien et que sa fonction de corrélation dépend d'un ***paramètre de corrélation*** θ :

$$R(x,u) = \exp\left(-\theta \sum_{j=1}^{d} |x_j - u_j|^2\right) = \exp(-\theta \|x-u\|^2). \qquad (3)$$

Cette fonction de corrélation permet de prendre en compte le fait que si deux points du domaine d'échantillonnage sont proches, alors nécessairement les réponses du simulateur en ces deux points sont fortement corrélées et *a contrario*, si deux points sont très éloignés les réponses sont considérées comme quasi indépendantes. Les paramètres de la fonction de corrélation (ici θ) sont bien souvent estimés par maximum de vraisemblance (Koehler et Owen (1996)). Cette estimation s'obtient grâce à des outils numériques (Mardia et Marshall (1984)) qui malheureusement coûtent très chers et aboutissent parfois à un maximum local de vraisemblance. Or la figure 2 montre que le comportement du prédicteur est très différent suivant la valeur du paramètre de corrélation. Il semble donc que la prédiction ne soit pas très robuste aux variations de θ.

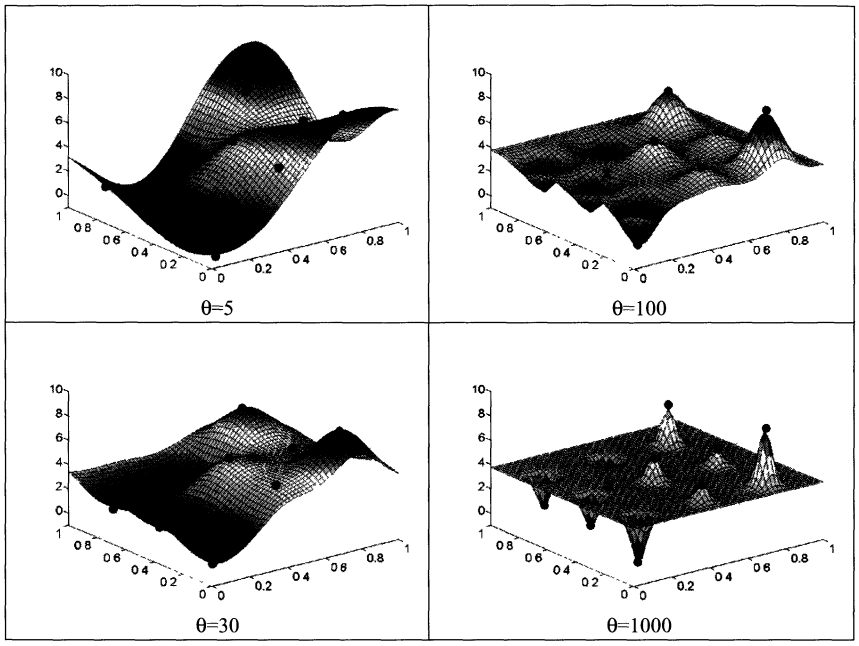

FIGURE 2

Evolution de la prédiction d'un simulateur à 2 paramètres d'entrée en fonction de θ. Le prédicteur est construit à partir d'une régression constante $(X(x) = 1)$ et d'un processus résiduel de fonction de corrélation (3). Les points représentent les 9 simulations effectuées.

2.2. Les plans d'échantillonnage

Les plans d'échantillonnage couramment utilisés dans cette approche statistique n'ont, en général, pas de structure particulière. Ils sont construits de façon à optimiser des critères de qualité tels que :

- L'erreur quadratique moyenne intégrée (Sacks *et al.* (1989 a et b)) qui va, en quelque sorte, sélectionner le plan qui minimise l'écart entre la réponse du simulateur et sa prédiction,

$$\min_D \frac{1}{\sigma^2} \int_{[0,1]^d} E[Y(x) - \widehat{Y}(x)]^2 dx.$$

- Le critère d'entropie (Schwery et Wynn (1987) ou Currin *et al.* (1991)) qui permet de mesurer la quantité d'information fournie par la simulation,

$$\max_D \det(R).$$

- La distance maximin (Johnson *et al.* (1990)) qui vise à choisir le plan qui maximise la distance minimale entre deux points du plan

$$\max_{D} \min_{x,y \in D} d(x, y).$$

où d est la distance qui intervient dans la fonction de corrélation R.

Les plans optimaux construits à l'aide de ces critères coûtent très chers puisqu'ils nécessitent l'optimisation numérique d'une fonction à $N \times d$ variables (Sacks *et al.* (1989a et b)). Il est possible de diminuer le temps de calcul en limitant la recherche à certaines classes de plans tels que les hypercubes latins (Park (1994)), les tableaux orthogonaux (Tang (1994)), ou les suites de faible discrépance (Bates *et al.* (1996)). On peut de plus noter que les trois critères ci-dessus dépendent de la modélisation choisie. Les plans sélectionnés ne sont donc optimaux que pour un paramètre de corrélation fixé.

2.3. Bilan

La modélisation de cette première approche statistique est particulièrement bien adaptée aux expériences simulées grâce au processus résiduel qui permet à la fois de corriger l'ajustement en moyenne et d'introduire une corrélation spatiale entre les paramètres du simulateur. Cependant, le fait d'imposer au prédicteur de passer par tous les points d'observation entraîne des problèmes d'irrégularité. De plus, nous pouvons émettre quelques inquiétudes sur le fait que la prédiction repose entièrement sur l'*a priori* fait sur la fonction de corrélation, étant donné que le prédicteur n'est pas robuste aux variations du paramètre de corrélation. Les deux inconvénients majeurs de cette modélisation viennent du rôle prépondérant du processus résiduel. Il serait alors intéressant de proposer un nouveau modèle permettant d'équilibrer l'importance des parties déterministe et aléatoire.

Les plans utilisés pour cette approche sont efficaces, puisque construits suivant un critère de qualité, mais uniquement pour un modèle fixé. Ce qui entraîne, outre des problèmes de robustesse du plan, une mauvaise répartition spatiale des points dans le domaine de simulation. Les points d'échantillonnage ont en effet tendance à se concentrer aux bord du cube unité et laissent le centre du domaine non testé par les simulations (*cf.* Koehler et Owen (1996)).

3. Approche par échantillon aléatoire

Cette approche statistique est proposée par Koehler et Owen (1996). Elle repose essentiellement sur les techniques d'échantillonnage utilisées en intégration numérique.

3.1. Modélisation et estimation

On considère ici que la réponse du simulateur peut se décomposer sur une base de fonctions et que seuls les termes prédominants de la décomposition suffisent à la

modéliser.

$$Y(x) = X(x)\beta, \ \forall x \in [0,1]^d \tag{4}$$

où $X(x)$ est le vecteur des fonctions de la base retenues pour la modélisation et β est le vecteur inconnu des paramètres du modèle.

Il existe plusieurs méthodes pour déterminer le meilleur vecteur de paramètres β, mais la plus naturelle est d'utiliser le critère moindres carrés par rapport à une distribution F sur $[0,1]^d$. Afin de simplifier les notations, supposons que F soit la distribution uniforme sur le cube unité, on a alors

$$\widehat{\beta} = \left(\int_{[0,1]^d} {}^t X(x) X(x) dx \right)^{-1} \int_{[0,1]^d} {}^t X(x) Y(x) dx.$$

On remarque que le premier terme de l'estimateur disparaît si on choisit une base orthonormée de fonctions. Autrement dit, estimer β revient à intégrer simultanément plusieurs fonctions définies par ${}^t X(x)Y(x)$. La réponse du simulateur $Y(x)$ n'étant pas connue analytiquement, ces intégrales sont évaluées numériquement à partir de quelques simulations.

La technique la moins coûteuse pour évaluer l'intégrale I d'une fonction f sur $[0,1]^d$ est la méthode de Monte Carlo. Elle consiste à estimer sans biais l'intégrale I par la somme

$$\widehat{I} = \frac{1}{N} \sum_{i=1}^{N} f(x_i), \tag{5}$$

où les x_i, $i = 1, , N$ sont des points choisis au hasard dans le cube unité. On peut noter que contrairement à l'approche précédente l'aléa n'est pas introduit au niveau du modèle mais de l'échantillon.

On sait de façon classique que cet estimateur converge et que pour toute fonction de carré intégrable, sa vitesse de convergence est égale à $\sigma N^{-1/2}$ et sa variance à $\sigma^2 N^{-1}$ (où σ^2 est la variance de $f(xi)$).

Si on applique cette méthode au modèle, on obtient alors l'estimateur de β

$$\widehat{\beta} = \left(\int_{[0,1]^d} {}^t X(x) X(x) dx \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} X(x_i) Y(x_i). \tag{6}$$

Il existe différentes techniques pour améliorer la méthode de Monte Carlo. La plupart agissent sur l'échantillonnage de façon à contraindre les points x_i à bien représenter toutes les parties du cube unité (*space filling designs*). On trouve essentiellement deux types de techniques : les méthodes quasi-Monte Carlo basées sur des suites de faible discrépance (Sloan et Joe (1994), Fang *et al.* (2000)) qui permettent une répartition uniforme des points (suivant une mesure donnée), et les méthodes de réduction de la variance basées sur des techniques d'échantillonnage à 2 degrés et que nous présentons maintenant.

3.2. Echantillonnage à 2 degrés

Chaque arête du domaine de simulation $[0,1]^d$ est découpée en un nombre fini de segments de même longueur que l'on numérote de 0 à $q_i - 1$ où i désigne la $i^{\text{ème}}$ arête. On munit chaque ensemble $\{0, ..., q_i - 1\}$ de la loi d'addition modulo q_i. On obtient ainsi une partition du cube unité en $q_1 \times ... \times q_d$ cellules, indexées par les éléments du groupe abélien fini $G = (\mathbb{Z}/q_1) \times ... \times (\mathbb{Z}/q_d)$ (Figure 3).

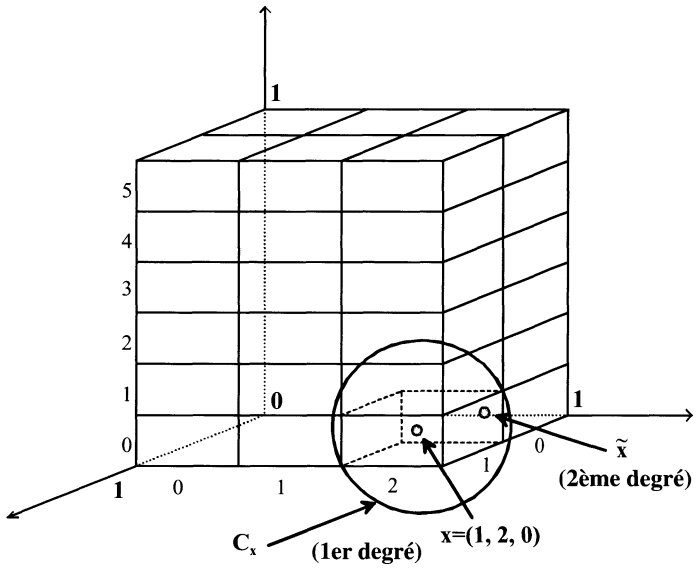

FIGURE 3

Découpage de $[0,1]^3$ *avec* $G = (\mathbb{Z}/2) \times (\mathbb{Z}/3) \times (\mathbb{Z}/2)$

L'échantillonnage à deux degrés consiste alors à

1er degré : choisir les cellules du cube unité représentées par le plan

2ème degré : tirer au hasard un point dans chaque cellule sélectionnée au 1$^{\text{er}}$ degré.

Un point d'échantillonnage $\widetilde{x} = (\widetilde{x}_1, \ldots, \widetilde{x}_d)$, est alors la réalisation d'un vecteur aléatoire de la forme $\widetilde{x}_j = (x_j + \varepsilon_j)/q_j, j = 1, ..., d$, où $x = (x_1, ..., x_d) \in G$ représente la cellule dans laquelle se trouve le point $\widetilde{x}$, et $\varepsilon = (\varepsilon_1, ..., \varepsilon_d)$ est un vecteur aléatoire de distribution uniforme sur $[0,1]^d$, qui désigne à quel endroit le point se trouve dans la cellule. Dans cette technique d'échantillonnage seul le premier degré est contrôlable. L'objectif est alors de choisir les cellules représentées par le plan, de façon à s'assurer que toutes les parties de $[0,1]^d$ soient bien testées bien par les simulations. On utilise pour cela des sous-ensembles de G appelés tableaux orthogonaux de force t, $1 \leqslant t \leqslant d$, qui ont la particularité d'avoir une répartition uniforme de leurs points en projection sur t faces du cube unité. Par exemple sur la figure 4 sont représentés en projection deux tableaux orthogonaux de $(\mathbb{Z}/3)^3$ à 9 points. Celui de droite est de

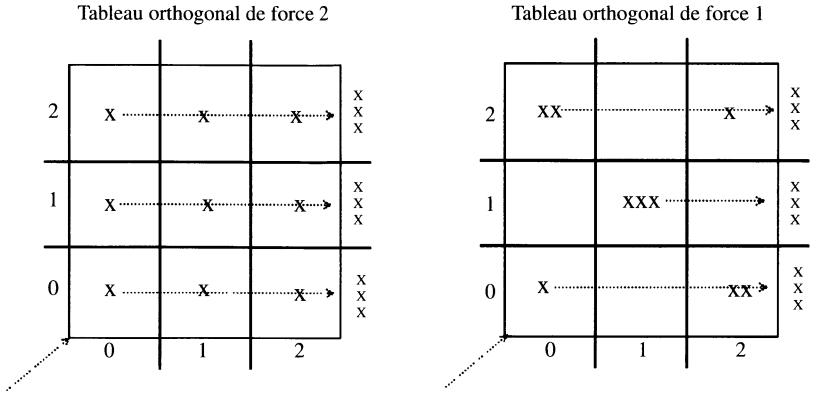

FIGURE 4

Projection de tableaux orthogonaux sur une arête et une face du cube unité

force 2 et celui de gauche est de force 1 (lorsque $t = 1$, on parle aussi d'hypercube latin).

On constate sur la figure 4, que les points du tableau de force 2 recouvrent de façon plus uniforme le domaine de simulation que ceux du tableau de force 1. Et en effet, Mc Kay *et al.* (1979), Owen (1992) et Tang (1993) ont montré que plus la force du tableau orthogonal est élevée, plus la variance de l'estimateur $\widehat{I}$ (5) diminue. Ce qui confirme bien que la répartition uniforme des points d'échantillonnage dans le domaine de simulation est un critère de qualité des plans, et ceci quel que soit le modèle choisi.

Remarque : Les propriétés de l'estimateur $\widehat{I}$ aux points d'un échantillonnage à 2 degrés, notamment le fait qu'il soit sans biais, ont été démontrées dans le cas où le tableau orthogonal (1^{er} degré) est randomisé (Owen (1992-1995)). La randomisation n'intervient cependant pas dans le fait que les points du plan sont uniformément répartis dans le cube unité. Elle ne semble donc pas nécessaire lorsqu'on sort du cadre de l'intégration numérique, par exemple pour ajuster un modèle de krigeage.

3.3. Bilan

Contrairement à l'approche par résidu aléatoire, la modélisation ici est assez peu réaliste à moins d'avoir des informations précises sur le phénomène simulé. En revanche les plans de cette deuxième approche sont construits de façon indépendante du modèle puisqu'ils visent à représenter au mieux tout le domaine de simulation. On peut donc supposer qu'un plan de bonne qualité le sera quel que soit le modèle choisi, et notamment qu'il sera robuste aux variations du paramètre de corrélation dans le cas d'une modélisation du type (1).

4. Nouvelle approche statistique des expériences simulées

L'idée pour définir une nouvelle approche statistique des expériences simulées s'inspire directement des bilans des deux approches précédentes et consiste à :

- conserver le processus résiduel en ce qui concerne la partie modélisation pour tenir compte de l'écart systématique entre la réponse du simulateur et la régression linéaire présumée,
- utiliser les méthodes d'échantillonnage à 2 degrés avec un 1^{er} degré qui permet de s'assurer que toutes les parties du domaine d'échantillonnage sont bien testées par les simulations.

4.1. Le modèle

Pour définir un nouveau modèle, il est nécessaire de tenir compte des deux remarques suivantes. Premièrement, nous avons vu dans le paragraphe 2, que le rôle prépondérant tenu par le processus résiduel dans le modèle entraîne des problèmes d'irrégularité et de non robustesse du prédicteur. Il est donc indispensable de contrebalancer l'importance du processus résiduel afin que toute la prédiction ne repose pas entièrement sur lui. Deuxièmement, dans un échantillonnage à 2 degrés, seul le 1^{er} degré est contrôlable. Le modèle doit donc prendre en compte l'effet produit par le 2^{e} degré.

Notons C_x toute cellule du premier degré, où x est le point de $[0,1]^d$ représentant la cellule (par exemple son centre), et $\widetilde{x}$ le vecteur aléatoire appartenant à C_x, obtenu au deuxième degré (*cf.* figure 3). Nous proposons alors de modéliser la réponse du simulateur au point $\widetilde{x} \in C_x$, par un processus indexé par le représentant de la cellule x :

$$Y(x) = X(x)\beta + \Gamma(x) + e(x). \tag{7}$$

• $X(x)\beta = E[Y(x)]$ est la partie déterministe du modèle, précisant l'approximation retenue pour la prédiction en moyenne.

• $\Gamma(x)$ est un processus aléatoire, introduit pour tenir compte du résidu de l'approximation. De même que dans le paragraphe 2, il sert à corriger la prédiction en moyenne.

• $e(x)$ est un terme d'erreur introduit pour prendre en compte l'effet du deuxième degré de l'échantillonnage. Il est indépendant de $\Gamma(x)$, tel que $\forall x \in [0,1]^d$, $\forall u \in [0,1]^d$.

$$E[e(x)] = 0 \text{ et } \operatorname{cov}(e(x), e(u)) = \sigma_e^2 \delta_{xu}.$$

C'est donc un modèle avec terme d'erreur qui intervient ici. Celui-ci n'est pas nouveau, puisqu'il est déjà utilisé dans le cadre des expériences simulées, notamment par Sacks, Schiller et Welch (1989). Seulement leur motivation est tout autre que la notre, puisqu'elle vient de la nature de ce que l'on cherche à prédire. Un prédicteur sans terme d'erreur interpole les observations, il va donc prédire la réponse du simulateur et non pas le phénomène simulé. Or on peut tout à fait considérer que cette réponse est

entachée d'une erreur par rapport au phénomène simulé, due, d'une part aux arrondis machine, et d'autre part aux simplifications de la modélisation. Ils introduisent en fait une *erreur de mesure* pour prédire le phénomène simulé.

De notre côté, le terme d'erreur représente l'effet du deuxième degré de l'échantillonnage (même si on peut imaginer qu'il englobe aussi une erreur de mesure). On raisonne en fait ici comme si on procédait à un «recalage» des observations, c'est-à-dire en considérant la réponse du simulateur $y(\widetilde{x})$, au point $\widetilde{x} \in C_x$, comme la réalisation de $Y(x)$ bien que $\widetilde{x} \neq x$ presque sûrement. Ceci se justifie par le fait que le plan d'échantillonnage consiste ici en un sous-ensemble de cellules, représentées par des points choisis *a priori*, mais qui ne fixe pas les paramètres des simulations à réaliser, puisque le deuxième degré n'est pas contrôlable. Il les contraint uniquement à appartenir à certaines cellules. Au demeurant, il est tout à fait possible d'envisager que seules soient connues pour la prédiction les cellules C_x et les réponses du simulateurs au point $\widetilde{x} \in C_x$, mais pas les paramètres des simulations. Par exemple, dans le cas où les valeurs de certains paramètres ne peuvent pas être fixées avec précision. On cherche en fait à prendre en compte l'influence sur la réponse observée de l'incertitude sur les niveaux des paramètres de simulation. On dit qu'il y a *transmission des erreurs (sur les niveaux des facteurs)*. Certains travaux sur l'étude des effets des erreurs sur les niveaux des facteurs dans les plans d'expériences ont déjà été publiés (Box (1963), Draper et Beggs (1971)). Dans le cas particulier de l'échantillonnage à 2 degrés des expériences simulées, les erreurs sur les niveaux des facteurs ont une distribution connue, par exemple ici une loi uniforme sur la cellule concernée. Il est alors possible de faire le parallèle avec les problèmes rencontrés en qualité industrielle, lorsqu'on cherche à prédire la réponse d'une caractéristique de la qualité d'un produit quand les niveaux des facteurs sont incertains mais sont au voisinage de valeurs nominales données, ces voisinages étant définis par des tolérances fixées *a priori*. Quelques résultats (Vuchkov *et al.* (1983-1987-1998) concernant l'estimation de l'espérance de la réponse et sur la planification des essais lorsqu'il y a des erreurs sur les niveaux des facteurs, ont été établis dans le cas particulier où : 1) soit la caractéristique est connue, soit elle est estimée par une régression polynomiale; 2) les résidus sont de même variance et non corrélés. Ces travaux offrent une piste intéressante et prometteuse, tant du point de vue théorique que pratique pour l'approche que nous envisageons. L'adaptation de ces résultats à notre cas, à savoir des régressions autres que polynomiales (par exemple trigonométriques) et surtout des résidus corrélés (par exemple gaussien), reste cependant une question ouverte.

4.2. Comportement du nouveau prédicteur

Proposition (Christensen (1990))
Soit le paramètre $\eta^2 = \sigma^2/\sigma_e^2$, on définit alors la fonction de covariance

$$\operatorname{cov}(Y(x), Y(u)) = \sigma_e^2 V(x,u) = \begin{cases} \sigma_e^2(1+\eta^2 R(x,u)) & x = u \\ \sigma_e^2 \eta^2 R(x,u) & x \neq u \end{cases}. \qquad (8)$$

Si le vecteur des paramètres est estimable, alors le meilleur prédicteur linéaire sans biais de $Y(x)$ est de la forme

$$\widehat{T}(x) = X(x)\widehat{\beta} + \eta^{2t} r(x) V^{-1}(Y - X\widehat{\beta}), \quad \forall x \in [0,1]^d \tag{9}$$

avec $\widehat{\beta}$: $({}^tXV^{-1}X)^{-1\,t}XV^{-1}Y$ l'estimateur de Gauss-Markov de β, où V est la matrice carrée d'ordre N à éléments $V(x_i, x_j)$.

L'erreur quadratique moyenne de $\widehat{T}(x)$ est égale à

$$MSE(x) = \sigma_e^2 \left\{1 + K(x)({}^tXV^{-1}X)^{-1\,t}K(x) + \eta^2 - \eta^{4t} r(x) V^{-1} r(x)\right\}$$

où $K(x) = [X(x) - \eta^{2t} r(x) V^{-1} X]$.

Avec ce nouveau modèle, la covariance entre $Y(x)$ et $Y(u)$ ne dépend plus entièrement de la fonction de corrélation R du processus résiduel, mais aussi d'un terme η^2 que l'on appelle *paramètre de contribution relative de la corrélation*. C'est ce dernier qui permet de modérer l'importance du processus résiduel dans le modèle, ainsi que de s'affranchir de la contrainte d'interpolation.

La figure 5 permet d'illustrer le rôle joué par le paramètre de contribution relative de la corrélation. Nous avons représenté le prédicteur (9) obtenu pour différentes valeurs du paramètre η^2, ainsi que le prédicteur (2) de l'approche par résidu aléatoire (§ 2). Il apparaît d'une part que plus η^2 est petit, plus le prédicteur est loin des observations, et pour les grandes valeurs de η^2, il atteint presque l'interpolation. Ceci confirme que le terme d'erreur introduit dans ce nouveau modèle permet de contrôler le rôle du processus résiduel, et de lever la contrainte d'interpolation, ce qui entraîne un lissage du prédicteur. D'autre part, on peut remarquer que le 2^{nd} degré de l'échantillonnage est pris en compte par le modèle. En effet, dans le cas où 2 points de simulation sont très proches, on peut noter que cela entraîne une très forte variation pour le prédicteur sans terme d'erreur (2), alors que les deux nouveaux prédicteurs (9) restent à peu près lisses.

Avec ce modèle, le processus résiduel a une contribution moindre dans la prédiction, donc *a fortiori*, le rôle de l'ajustement en moyenne se trouve renforcée. Il faut donc maintenant choisir la régression linéaire avec soin, et non plus la fixer constante ($X(x) = 1$) comme cela est préconisé par certains auteurs tels que Welch *et al.* (1992). On peut donc supposer qu'un ajustement en moyenne de bonne qualité permet d'obtenir un prédicteur plus robuste aux variations du paramètre de corrélation θ. La figure 6 illustre bien cette hypothèse. Nous avons représenté l'évolution de deux prédicteurs en fonction de θ, l'un construit à partir d'une régression linéaire constante, l'autre à partir d'une régression linéaire trigonométrique prenant en compte l'effet moyen, les effets simples et l'interaction des deux facteurs (*cf.* Collombier et Jourdan (2001)).

4.3. Les plans d'échantillonnage

Les plans utilisés pour déterminer le 1^{er} degré de l'échantillonnage sont bien entendu ceux de l'approche par échantillon aléatoire (§ 3). Le but recherché étant

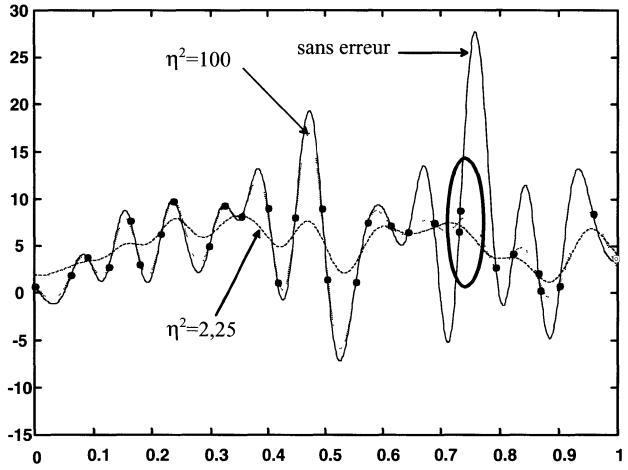

FIGURE 5

Evolution du prédicteur (7) pour différentes valeurs de η^2. Le prédicteur est construit à partir de 30 observations correspondant à 30 réponses présumées d'un simulateur à 1 paramètre d'entrée. La régression linéaire est fixée constante et le processus résiduel est gaussien

qu'ils représentent au mieux le domaine de simulation. Une étude empirique sur la qualité de ces plans (dans le cadre d'une régression trigonométrique) a été effectuée par Collombier et Jourdan (2001). Il ressort de cette analyse deux points importants.

Premièrement, les tableaux orthogonaux sont des plans efficaces quel que soit le paramètre de corrélation choisi. Ce qui n'était pas le cas des plans (§ 2.2) utilisés dans l'approche par résidu aléatoire, et posait un problème étant donné que le paramètre θ n'est pas connu *a priori*.

Deuxièmement, pour un coût fixé, tous les tableaux orthogonaux ne sont pas de même qualité. On utilise alors le ou les tableaux orthogonaux qui optimisent l'un des critères de qualité définis au paragraphe 2.2. On remarque que les plans d'échantillonnage ainsi sélectionnés sont extrêmement efficaces par rapport aux meilleurs plans (*i.e.* ceux qui minimisent la IMSE), et ceci quel que soit le paramètre de corrélation θ (contrairement aux meilleurs plans qui dépendent entièrement de la modélisation choisie).

Pour plus de détail, on pourra donc se référer à Collombier et Jourdan (2001), notamment en ce qui concerne l'influence de la force du tableau orthogonal (lien avec les critères d'entropie et distance maximin – compromis entre finesse de découpage du cube unité et force du tableau).

5. Conclusion

Cette nouvelle approche statistique des expériences simulées semble être intéressante. En effet, en ce qui concerne le modèle, nous obtenons un prédicteur

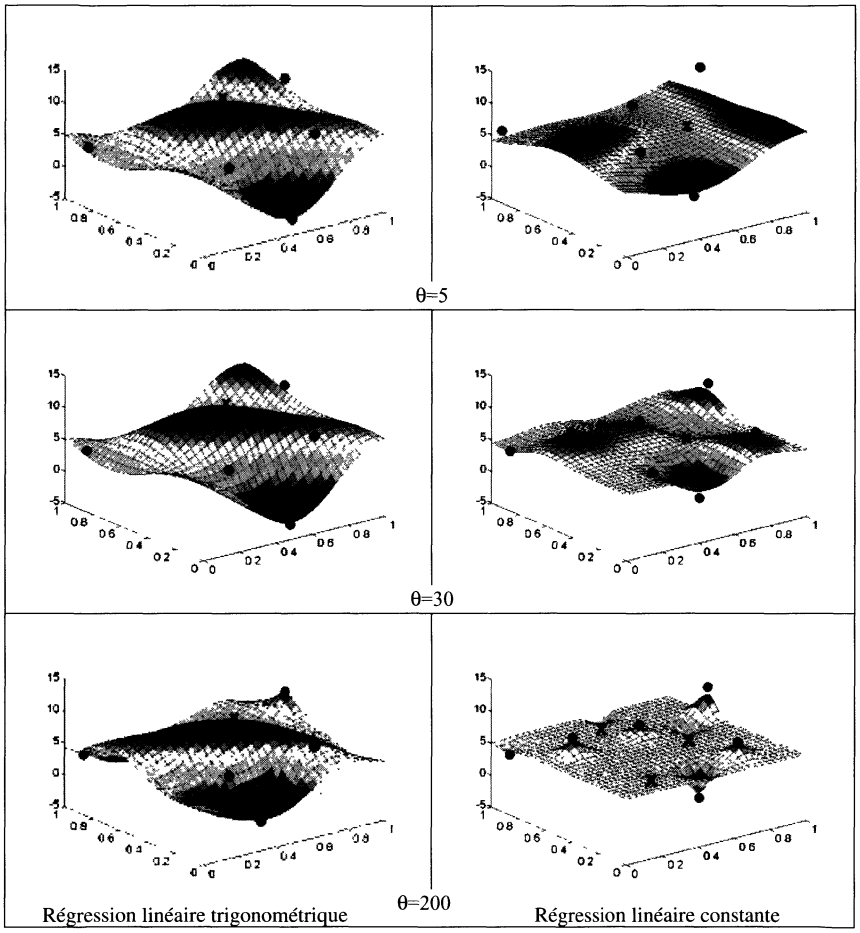

FIGURE 6

Evolution de la prédiction d'un simulateur à 2 paramètres d'entrée en fonction de θ. Le prédicteur est construit à partir d'un processus résiduel gaussien et du paramètre $\eta = 2$. Les points représentent les 9 simulations effectuées

plus lisse et robuste aux variations du paramètre de corrélation. Et pour ce qui est des plans, la technique d'échantillonnage proposée permet d'utiliser des plans de bonne qualité quel que soit le modèle fixé, et notamment θ.

Cet article est un premier travail exploratoire qui permet de faire un rapide état des lieux des différentes techniques statistiques utilisées pour traiter les expériences simulées et de proposer quelques idées nouvelles.

Remerciements

Je remercie les rapporteurs pour leurs remarques constructives, notamment celle concernant la transmission des erreurs sur les niveaux des facteurs qui ouvre des perspectives intéressantes.

Références

BATES R.A., BUCK R.J., RICCOMAGNO E., WYNN H.P. (1996), Experimental design and observation for large systems. *J. R. Statist. Soc. B* **58**, 77-94.

BOX G.E.P. (1963), The effects errors in the factor levels and experimental designs. *Technometrics* **5**, 247-262.

CHRISTENSEN R. (1990), *Linear models for multivariate, time series, and spatial data*. Springer-Verlag.

COLLOMBIER D., JOURDAN A. (2001), Régression trigonométrique et plans d'échantillonnage pour expériences simulées. *Revue de Statistique Appliquée*, **49 (2)**, 5-25.

CURRIN C.T., MITCHELL M., MORRIS M., YLVISAKER D. (1991), Bayesian prediction of deterministic functions, with applications to the design and the analysis of computer experiments. *J. Amer. Statist. Assoc.*, **86**, 953-963.

DRAPER N.R., BEGGS W.J. (1971), Errors in factor levels and experimental designs. *Ann. of Math. Statist.*, **41**, 46-58.

FANG K-T., LIN D.K.J., WINKER P., ZHANG Y. (2000), Uniform design : theory and application. *Technometrics* **42**, 237-248.

JOHNSON M.E., MOORE L.M., YLVISAKER D. (1990), Minimax and maximin distance designs. *J. of Statist. Planning and Inference*, **26**, 131-148.

JOURDAN A. (2000), Analyse statistique et échantillonnage d'expériences simulées. Thèse, Université de Pau et des Pays de l'Adour

KOEHLER JR., OWEN A.B. (1996), Computer experiments. In Ghosh S. and Rao C.R., editors, *Handbook of statistics, 13 : Design and analysis of experiments*, North-Holland, Amsterdam, 261-308.

MARDIA K.V., MARSHALL R.J. (1984), Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**, 135-146.

MATHERON G. (1963), Principles of geostatistics. *Economic Geology*, **58**, 1246-1266.

MC KAY M.D., BECKMAN R.J., CONOVER W.J. (1979), Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239-245.

OWEN A.B. (1992), Orthogonal arrays for computer experiments, integration and visualization. *Statist. Sinica*, **2**, 439-452.

OWEN A.B. (1995), Randomly permuted (t,m,s)-nets and (t,s)-sequences. In Niederreiter H. and Shuie P.J.S. eds, *Monte Carlo and quasi-Monte Carlo Methods in Scientific Computing*, Springer, New York.

PARK J.S. (1994), Optimal Latin hypercube designs for computer experiments. *J. of Statist. Planning and Inference*, **39**, 95-111.

SACKS J., SCHILLER S.B., WELCH W.J. (1989a), Designs for computer experiments. *Technometrics*, **31**, 41-47.

SACKS J., WELCH W.J., MITCHELL T.J., WYNN H.P. (1989b), Design and analysis of computer experiments. *Statist. Science*, **4**, 409-435.

SCHWERY M.C., WYNN H.P. (1987), Maximum entropy sampling. *J. of Appl. Statist.*, **14**, 165-170.

SLOAN I.H., JOE S. (1994), *Lattice methods for multiple integration*. Oxford science Publications.

TANG B. (1993), Orthogonal array-based Latin hypercubes. *J. Amer. Statist. Assoc.*, **88**, 1392-1397.

TANG B. (1994), A theorem for selecting OA-based latin hypercubes using a distance criterion. *Comm. Statist.-Theory Meth.*, **23**, 2047-2058.

VUCHKOV, I.N., BOYADJIEVA L.N. (1983), The robustness of experimental designs against errors in the factor levels. *J. Statistical Computation & Simulation*, **17**, 31-41.

VUCHKOV, I.N., BOYADJIEVA L.N., DAMYANOV P.S. (1987), Off-line quality control without replicates. Paper presented at the *Satellite Conference on Mathematics of Design of Experiments, 17th Europeean Meeting of Statisticians*, Thessaloniki, Greece.

VUCHKOV, I.N., BOYADJIEVA L.N., TZENKOVA (1998), Quality mprovement through mechanistic models. In Atkinson A.C., Pronzato L., Wynn W.P. editors, *MODA 5 – Advances in Model Oriented Data Analysis and Experimental Design* (Proceedings of the 5th International Worshop, Marseille, France, june 22-26 1998), Physica-Verlag.

WELCH W.J., BUCK R.J, SACKS J., WYNN H.P., MITCHELL T.J, MORRIS M.D. (1992), Screening, predicting and computer experiments. *Technometrics*, **34**, 15-25.