COMPARAISON DE METHODES DE CRIBLAGE POUR LA SIMULATION NUMERIQUE

Michelle SERGENT[1], Delphine DUPUY[2], Bernard CORRE[3], Magalie CLAEYS-BRUNO [1]
[1] *Aix-Marseille Université, Institut des Sciences Moléculaires de Marseille, AD2EM, UMR-CNRS-6263, Campus St Jérôme, Service D 52, 13397 Marseille Cedex 20, France*
[2.] *Ecole des Mines de Saint-Etienne, Département 3MI, 42000 SAINT-ETIENNE.*
[3.] *Total EP/GSR/COP/EM, CSTJF Avenue Larribau 64018 PAU Cedex.*

Résumé
Les méthodes de criblage permettent de repérer parmi un ensemble de variables d'entrée potentiellement influentes, celles qui le sont effectivement dans un domaine de variation fixé. Il existe actuellement différentes méthodes de criblage adaptées à cette problématique, comme les matrices d'expériences de Plackett et Burman, les matrices d'expériences supersaturées, le criblage par groupes, la bifurcation séquentielle,…
Il nous a semblé intéressant de comparer ces différentes méthodes sur un cas réel dans le domaine de la simulation numérique. En effet, les phénomènes très complexes sont souvent abordés au moyen de codes de calcul, lourds et coûteux en temps. Dans ce type d'approche, les modèles sont alors de très grande dimension et la première étape consiste à réduire l'espace par identification des paramètres réellement influents, ce qui ressort précisément du domaine d'application des méthodes de criblage. L'étude présentée porte sur la simulation d'un gisement d'hydrocarbures exploité par 9 puits avec pour objectif de simuler le plus précisément possible le comportement d'un réservoir. Parmi les milliers de variables du modèle, 51 facteurs ont été retenus comme potentiellement influents sur les variables de sortie, à savoir la production cumulée d'hydrocarbure à différents temps.
Les résultats obtenus pour chaque méthode seront présentés et comparés à ceux obtenus lors de l'étude préliminaire par la méthode OAT (one factor at a time).

Mots clés : Plans d'expériences

Abstract
In simulation models or in experimental programs involving a large number of controllable factors, one of the aims is the identification of the subset of factors that have substantial influence on the responses of interest. In this paper we provide a review of different screening methods that are useful to eliminate negligible factors so that efforts may be concentrated upon just the important ones. These different methods are described and are applied on a simulation model with 51 factors. The comparison of the results is presented with the advantages and disadvantages for each method

Introduction

L'objet de ce travail est :

1. de tester sur un cas réel une série de méthodes de criblage basées sur l'utilisation de matrices prédéfinies plus ou moins coûteuses en termes de simulations et des méthodes plus globales, basées sur de l'algorithmique.
2. de les comparer entre elles.

Ainsi, sont comparées entre elles les méthodes mettant en œuvre :

- des matrices d'Hadamard de résolution III
- des matrices d'Hadamard de résolution IV
- des matrices supersaturées
- le criblage par groupes
- le criblage par groupes multiples
- la bifurcation séquentielle.

1. PRESENTATION DU CAS D'ETUDE

Le cas réel qui a servi de support a été proposé par B. Corre (TOTAL) : cette étude s'est appuyée sur l'utilisation d'un simulateur d'écoulement en milieu poreux. Il s'agit d'un cas "spatial", c'est-à-dire tridimensionnel, discrétisé en 80 x 80 x 6 = 38400 mailles. Le nombre de facteurs considérés comme incertains ou variables est 51, sachant que le jeu de données lui-même en compte quelques centaines de milliers. Les réponses du simulateur sont multiples et temporelles : en l'occurrence, seules les **productions cumulées d'hydrocarbure à différents temps** ont été étudiées.

La résolution du système au sein du simulateur est basée sur la technique des lignes de courant.

L'étude ici présentée est avant tout une étude dont l'objectif est de réduire les dimensions de l'espace étudié par la mise en œuvre de différentes techniques de criblage. Du jeu de données initial, **51 facteurs** ont été retenus comme potentiellement influents sur les réponses, donc susceptibles de varier dans un intervalle d'intérêt, tous les autres facteurs étant fixés à leur valeur nominale. L'espace considéré dans cette étude comporte donc 51 dimensions. Les facteurs choisis pour le présent problème sont de différents types :

- des données d'environnement portant sur les propriétés pétrophysiques du réservoir et thermodynamiques des fluides en place. Ces données sont affectées d'incertitudes.
- des données de contrôle, à savoir l'emplacement des différents puits autour d'une position initiale (ou de base), leurs contraintes de fonctionnement en débit et pression, X_{16} à X_{33}.
- une donnée de simulation, qui est le nombre total de lignes de courant utilisées dans la résolution du système.

Plus précisément, ce cadre étant fixé, le domaine de variation des facteurs a été défini de telle sorte qu'un nombre limité de facteurs soit actif.

2. ETUDE DE CRIBLAGE

Rappelons qu'une **étude de criblage** peut être définie comme une étape permettant de **repérer rapidement** dans un grand nombre (k) de facteurs potentiellement influents, les quelques facteurs (f) qui le sont effectivement dans un domaine expérimental fixé. Cette étude va permettre de déterminer le "**poids**" de chaque niveau de chaque facteur, pour ensuite **les classer** par ordre d'importance. Selon le principe de parcimonie, ou principe de "Pareto", ou du "rasoir d'Ockham", le nombre de facteurs effectivement actifs est faible : $f << k$

Les principales méthodes de criblage disponibles sont présentées brièvement ci-dessous et illustrées sur ce cas d'étude.

2.1. Matrices de criblage

2.1.1. Matrice d'HADAMARD de résolution III

Les matrices d'expériences de criblage les plus connues pour des facteurs à 2 niveaux sont les matrices d'Hadamard ou de Plackett et Burman (1946), $2^k//N$. Une **matrice d'Hadamard** est une matrice carrée dont les éléments sont +1 ou -1 et dont les colonnes sont toutes orthogonales entre elles et pour laquelle la matrice d'information $X'X$ est telle que :$\mathbf{X'X^*} = N\,\mathbf{I}_N$ avec, $\mathbf{I}_N$: matrice identité d'ordre N.

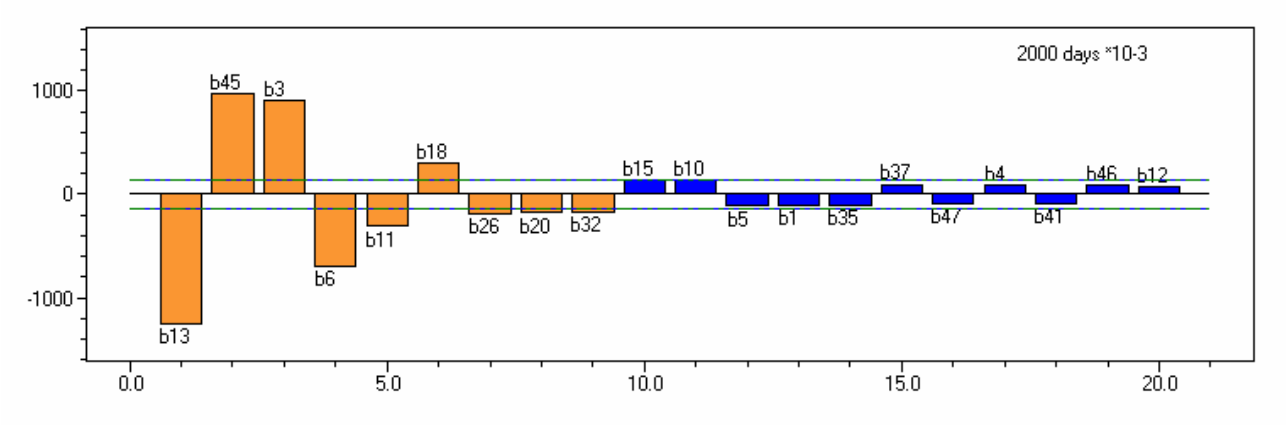
Nous devons rappeler qu'une matrice d'expériences d'Hadamard est dite de **résolution III** lorsque tous les coefficients monoindicés, qui représentent les "poids" (b_i) des facteurs, peuvent être connus indépendamment des autres coefficients monoindicés mais que certains - tous, dans le cas présent - seront aliasés avec des coefficients bi-indicés, c'est-à-dire les effets d'interaction du 1^{er} ordre.

Dans notre cas, la matrice d'expériences de screening optimale, réalisée pour estimer les poids des **51 facteurs** à 2 niveaux en un minimum d'essais, est une **matrice d'Hadamard à N = 56 expériences**.

➡ **Outils d'aide à l'interprétation :**

Les divers outils graphiques d'aide à l'interprétation (graphe des effets, Normal plot, Pareto …) permettent de mettre évidence les coefficients significatifs. Pour plus de lisibilité, sur le graphe ci-dessous,

seuls les 20 coefficients les plus significatifs ont été représentés, en les classant par ordre d'importance décroissante :



Etude des effets classés de la réponse : 2000 jours.10^{-3}

L'ensemble de ces outils graphiques montre un sous-ensemble de **6 variables** ayant très probablement une influence sur la réponse étudiée :
X_3, X_6, X_{13}, X_{45}, et dans une moindre mesure, **X_{11} , X_{18}**

2.1.2. Matrice d'HADAMARD de résolution IV

L'hypothèse d'additivité que nous avons énoncée précédemment est une hypothèse forte qui peut se justifier si l'hypothèse du "criblage" est respectée. Néanmoins, pour réduire les risques liés à cette hypothèse, il est possible d'adapter la stratégie expérimentale et pour cela, la démarche la plus utilisée consiste à reproduire la matrice d'Hadamard **H** en inversant tous les signes. Le nombre d'expériences est alors doublé mais la nouvelle matrice d'expériences est de résolution **IV**.

H
- H

Une matrice d'expériences est dite de résolution **IV**, si tous les coefficients monoindicés sont connus indépendamment des coefficients bi-indicés (effets d'interactions du premier ordre) mais certains (ou tous) seront aliasés avec des coefficients tri-indicés (effets d'interaction du deuxième ordre). D'autre part, certains effets d'interaction du premier ordre (ou tous) seront aliasés avec d'autres effets d'interaction du premier ordre. Cette technique est très souvent désignée sous le terme "*repliement*" ou "***Fold over***".

Dans notre cas, selon la technique du "foldover", la matrice d'Hadamard à 56 expériences a été reproduite en inversant tous les signes de la matrice. Ainsi nous obtenons une nouvelle matrice à **N = 112 expériences,** qui est de résolution **IV**, ce qui signifie que les effets principaux sont estimés indépendamment des effets d'interaction d'ordre 1 et les effets d'interaction du 1er ordre sont aliasés avec d'autres effets d'interaction du 1er ordre.

L'exploitation des résultats de cette nouvelle matrice met en évidence plus explicitement les facteurs influents et montre que les variables **X_{13}, X_{45}, X_3, X_6, X_{11}** et **X_{18}** sont les plus influentes sur la réponse : production à 2000 jours.

2.1.3. Matrices supersaturées

Les matrices supersaturées sont caractérisées par un nombre d'expériences inférieur au nombre de facteurs étudiés. Ces matrices ne peuvent être utilisées que si, dans le grand nombre de facteurs étudiés, la probabilité qu'un facteur de cet ensemble soit influent est très faible. D'autre part, ces matrices reposent, comme toute matrice de criblage, sur l'hypothèse d'additivité ce qui sous-entend que tous les effets d'interaction sont supposés nuls.

Il existe dans la littérature de nombreuses méthodes de construction de ces matrices supersaturées. Nous ne développerons ici que la méthode utilisée pour notre cas d'étude, la méthode proposée par Lin en 1993. Lin propose de sélectionner une matrice d'Hadamard (N × k) et de choisir une colonne de cette

matrice désignée sous le nom de colonne "branchée". Il regroupe les N lignes de cette matrice en 2 matrices de N/2 lignes, l'une correspondant aux lignes comportant le signe – dans la colonne "branchée", l'autre correspondant aux lignes comportant le signe + dans la colonne "branchée". On supprime dans ces deux matrices la colonne "branchée" et on obtient alors deux matrices supersaturées comportant N/2 lignes et (k-1) colonnes, ce qui conduit dans notre étude à une matrice de 28 expériences.

➡ **Traitement et Interprétation :**

Dans ces études de criblage, utilisant des matrices d'expériences supersaturées, le nombre d'expériences étant largement inférieur au nombre de facteurs, les méthodes classiques de traitement des résultats (méthode des moindre carés,..) ne peuvent plus être appliquées. Il est donc nécessaire de faire appel à d'autres méthodes permettant d'identifier les quelques facteurs actifs. Nous ne développerons que la méthode utilisée dans notre étude, c'est-à-dire la méthode combinant une **régression Step-Wise et toutes les régressions possibles** pour un nombre de facteurs donné.

En conclusion, l'interprétation simultanée de ces outils montre que **5 variables** apparaissent comme déterminantes : **X_3, X_6, X_{13}, X_{33}** et **X_{45}**.

2.2. Criblage par groupes

2.2.1. Méthode du criblage par groupes

Le principe de cette méthode proposée par Watson (1961) est de passer en revue à chaque test, un maximum de facteurs, sous la condition, bien entendu, que cela puisse se faire physiquement et techniquement. La démarche est la suivante :

- l'ensemble des facteurs ou variables indépendantes de départ est divisé en groupes de tailles adéquates, égales ou non,
- chaque groupe est alors considéré comme un seul facteur appelé facteur-groupe et on définit les deux niveaux du facteur-groupe ainsi :
  - le facteur-groupe est au niveau (+) lorsque tous les facteurs du groupe sont au niveau (+),
  - le facteur-groupe est au niveau (-) lorsque tous les facteurs du groupe sont au niveau (-).

  Les facteur-groupes sont traités comme des variables simples dans une première étape commune aux stratégies à deux étapes et à plusieurs étapes. Dans cette étape, on teste et on élimine les groupes non influents,
- les facteurs appartenant aux groupes influents sont ensuite traités,
  - → soit en les testant individuellement et ceci constitue un processus à deux étapes,
  - → soit dans un processus à plusieurs étapes où les groupes qui se sont révélés influents dans la première étape sont redivisés en groupes de taille plus petite, qui sont alors testés de la même manière dans les étapes suivantes.

Ce processus de rassemblement des facteurs en groupes, considérés comme de simples facteurs, pourra se répéter aussi longtemps que nécessaire.

Néanmoins, la démarche décrite précédemment demeure généralement théorique car les cas réels d'application mettent en évidence un besoin d'adaptation de la théorie de cette méthode à la réalité du problème. En effet, on peut facilement concevoir que tous les facteurs n'ont probablement pas la même probabilité (**p**) *a priori* d'être influent et il est intéressant de prendre en compte la connaissance *a priori* dans la construction des groupes. Cette approche va entraîner une division des facteurs en groupes de taille inégale et une répartition judicieuse des facteurs.

➡ **Etape 1 :**

Dans notre étude, les **51 facteurs** ont été partitionnés en **plusieurs groupes de taille non égale**. Pour cela, il a été demandé de classer les facteurs selon une idée *a priori* de l'impact de chaque facteur sur la réponse (et/ou de sa probabilité). Pour chaque facteur, 2 niveaux ont été attribués en choisissant comme niveau (+) celui qui serait sensé augmenter la réponse si le facteur était influent et il a été attribué, si possible, une probabilité d'être influent. Les différents groupes ont été construits à partir de cette connaissance *a priori*. Ainsi, **19 groupes** ont été créés et une matrice de screening optimale a été construite pour étudier ces 19 facteur-groupes. La matrice choisie est une matrice d'Hadamard à **20 expériences**.

➡ **Etape 2 :**

Selon la démarche décrite précédemment, les facteurs appartenant aux facteur-groupes influents seront étudiés dans une deuxième étape. Ainsi, les 10 facteurs des 2 groupes actifs (G_7 et G_8) seront étudiés dans une deuxième étape alors que les 41 autres facteurs sont déclarés comme non-influents. Pour l'étude de ces **10 facteurs** potentiellement influents, X_2, X_3, X_4, X_6, X_{10}, X_{11}, X_{13}, X_{18}, X_{19} et X_{45}, une matrice d'Hadamard à N = **12 expériences** a été construite.

Une synthèse de l'étude graphique des effets et de l'approche bayésienne met en évidence la non-influence de 4 variables : X_2, X_4, X_{10} et X_{19} et permet de classer les 6 autres variables par ordre d'importance : **X_{13}, X_{45}, X_3, X_6** et dans une moindre mesure **X_{11}** et **X_{18}**.

2.2.2. Méthode du criblage par groupes multiples

Cette méthode est une généralisation de la méthode du "criblage par groupes" dans le but de détecter rapidement quelques facteurs influents dans un grand nombre de facteurs potentiellement influents. Dans le "criblage par groupes" conventionnel, chaque facteur est affecté à un unique groupe de facteurs et la somme des effets associée à chaque groupe est estimée après la première étape du screening. Dans la méthode de "criblage par groupes multiples", chaque facteur est affecté à plusieurs groupes à la première étape et un facteur est potentiellement influent si chaque groupe auquel le facteur appartient est déclaré "actif" à cette étape.

Dans notre étude, cette méthode met en évidence 3 variables très significatives : **X_3, X_{13}** et **X_{45}**.

2.3. Bifurcation séquentielle

La "bifurcation séquentielle" proposée par Bettonvil (1995) considère aussi des groupes de facteurs mais n'utilise pas de plan d'expériences de screening et pas de méthode d'analyse, du type régression ou analyse de variance. Le principe de la bifurcation séquentielle est similaire au principe de la recherche dichotomique : les expériences sont choisies séquentiellement en utilisant les résultats des expériences précédentes pour orienter la recherche des facteurs influents. Cette méthode permet d'étudier un grand nombre de variables en peu d'expériences, mais suppose, comme dans le criblage par groupes, de connaître *a priori* la direction des effets des facteurs étudiés.

Cette méthode de la bifurcation séquentielle a permis, en 29 simulations, de détecter avec certitude 5 facteurs influents : **X_3, X_6, X_{13}, X_{18}, X_{45}** et 5 autres facteurs dans une moindre mesure : **X_{11}, X_{19}, X_{26}, X_{32}, X_{33}.**

Conclusion

Cette application a permis de comparer les nombreuses méthodes de criblage et met en exergue l'importance des hypothèses qui leur sont propres, à savoir :

- L'hypothèse additivité pour les matrices d'Hadamard de résolution III,
- Un faible nombre de facteurs actifs pour les matrices supersaturées,
- Une connaissance du signe des effets des facteurs pour le criblage par groupe et la bifurcation séquentielle.

Pour conclure, un récapitulatif des résultats est présenté dans le tableau ci dessous.

METHODES	Nombre d'expériences	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15	U16	U17	U18	U19	U20	U21	U22	U23	U24	U25	U26	U27	U28	U39	U30	U31	U32	U33	U34	U35	U36	U37	U38	U39	U40	U41	U42	U43	U44	U45	U46	U47	U48	U49	U50	U51
OFAT	102																																																					
HADAMARD R III	56																																																					
HADAMARD R IV	112																																																				
Matrice supersaturée	28																																																				
Criblage par groupe	32																																																				
groupe multiple	28																																																				
Bifurcation Séquentielle	29																																																				

O

On peut facilement voir que :

- Toutes les méthodes, même avec un faible nombre d'expériences (N=28, N= 29) détectent les facteurs les plus importants : U_3, U_{13}, $U_{45.}$
- La matrice d'Hadamard de résolution IV avec 112 expériences permet une identification plus fiable et plus précise en s'affranchissant des effets d'interaction entre deux facteurs.
- La bifurcation séquentielle conduit inévitablement à l'identification de tous les facteurs influents mais cette approche exige une connaissance *a priori* du sens de l'éventuel effet d'un facteur.

Bibliographie

[1] Plackett, R.L. and J.P. Burman (1946) *Design of Optimum Multifactorial Experiments.* Biometrika, **33**: p. 305-325.

[2] Lin, D.K.J. (1993) *A New class of Supersaturated designs.* Technometrics, **35**: p. 28-31.

[3] Watson G.S. (1961) *A study of the group screening method,* Technometrics, **3**, p. 371-388.

[4] Kleijnen J. P. C. (1975) *Screening designs for poly-factor experimentation,* Technometrics, Vol.18, p 487-493.

[5] Bettonvil, B. (1995) *Factor screening by sequential bifurcation*, Simulation and computation, **24**, p. 165-185