Analyse de sensibilité : comparaison entre les plans d'expérience et la méthod

Magalie Claeys-Bruno, M. Dobrijevic, Michelle Sergent

ANALYSE DE SENSIBILITE : COMPARAISON ENTRE LES PLANS D'EXPERIENCES ET LA METHODE MONTE CARLO

Magalie CLAEYS-BRUNO [1], Michel DOBRIJEVIC[2,3], Michelle SERGENT[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] *Université de Bordeaux, Laboratoire d'Astrophysique de Bordeaux.*

[3] *CNRS/INSU, UMR 5804, BP 89, F-33271 Floirac CEDEX.*

Résumé

Depuis une dizaine d'années, les codes de simulation numérique deviennent de plus en plus réalistes et complexes, rendant leur compréhension délicate. Une attention toute particulière est accordée aux problèmes des incertitudes sur les variables d'entrée, à l'étude de leur propagation et de leur conséquence sur l'interprétation des résultats, ce qui définit les études de sensibilité. Pour tenter de répondre à ces questions, il est nécessaire de déterminer une stratégie efficace et la moins coûteuse possible, mais digne de confiance.

Ce travail présente une étude de sensibilité permettant de détecter les réactions clés dans un modèle chimique complexe modélisant l'atmosphère de Titan, en utilisant la technique des plans d'expériences qui, dans l'élaboration d'une stratégie expérimentale, vise à choisir au mieux les expériences les plus informatives. Cette méthodologie ainsi que les différentes stratégies envisagées (matrice de criblage de Plackett et Burman et matrice supersaturée) sont présentées et les résultats obtenus sont comparés à ceux issus de la méthode Monte Carlo.

Abstract

We compare two global sensitivity analysis methods dedicated to the determination of the relation between a given uncertain input and the output: Monte Carlo based method and experimental design in the field of hydrocarbons photochemistry for giant planets and Titan. These methods can be applied to the search for key reactions in a complex chemical scheme since these determination is a primordial importance for experimenters that will focus their study on these few reactions, in conditions adapted to the atmosphere of Titan.

The aim of these global methods is to determine how the system react to a global perturbation. All the input factors, the reaction rates, are perturbed simultaneously within a variation domain given by an arbitrary factor or by the uncertainty factor attached to each reaction. The principle is then to scan the space of input factor the most efficiently in order to have a good estimation on the effect on each output factors (concentrations), according to two different ways: Monte Carlo based method and design of experiments. The result for each method will be presented and discussed.

Mots-clés : Plans d'expériences

Introduction

1. Problématique

Dans le domaine de la modélisation de la photochimie des atmosphères planétaires, un nouveau paradigme est en cours d'émergence. En effet, une importance particulière est accordée à l'exactitude des prévisions des modèles et à la quantification des incertitudes sur les variables d'entrée du modèle. Pour améliorer ces prévisions, il est nécessaire d'identifier dans le schéma chimique, les réactions clés qui doivent être étudiées en priorité et de manière précise, dans des conditions pertinentes, c'est-à-dire proches de celles des atmosphères planétaires.

Dans cette étude, le modèle photochimique a été réduit aux composés en C_2 (composés de la forme CnHp avec n = 0, 1, 2), avec 12 composés et 114 réactions possibles (figure 1).

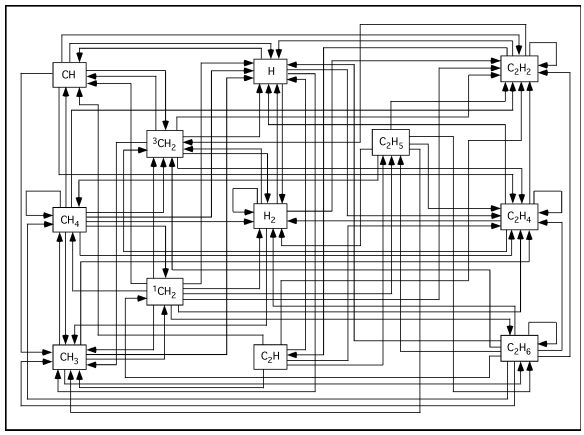

Figure 1 : Exemple possible de schéma réactionnel, pour un modèle photochimique

On considère ici, pour simplifier, un modèle dit 0D (sans dimension verticale) où il n'y a pas de transport. On s'intéresse donc à l'évolution des composés dans une "boîte atmosphérique" isolée où seuls les processus chimiques sont inclus. Les molécules sont dissociées par le rayonnement UV solaire et les composés réagissent entre eux pour former de nouvelles molécules. Pour simuler l'évolution des composés, il faut résoudre numériquement un système d'équations différentielles couplées et non linéaires auquel il faut ajouter les conditions initiales, à savoir les concentrations des différents composés. La concentration totale et la température sont des données connues au départ et sont supposées constantes dans le temps.

Pour un système chimique de *n* composés, le système d'équations différentielles donnant l'évolution de chaque type *i* de concentration $[X_i]$ est de la forme :

$$\begin{cases} \frac{d[X_1]}{dt} = f_1([X_1](t),[X_2](t),...,[X_n](t)) \\ \frac{d[X_2]}{dt} = f_2([X_1](t),[X_2](t),...,[X_n](t)) \\ ... \\ \frac{d[X_n]}{dt} = f_n([X_1](t),[X_2](t),...,[X_n](t)) \end{cases}$$

où les fonctions f_i sont non linéaires.

Ces fonctions non linéaires dépendent des taux de production et de perte chimique de chaque composé. Les constantes de réactions (vitesses de réaction et coefficients de photodissociation) sont obtenues à partir de mesures en laboratoire dans des conditions physico-chimiques qui miment partiellement les atmosphères planétaires. Ces constantes sont donc généralement très imprécises, la plupart ne sont connues qu'à un facteur 2 près, au mieux (cela va souvent jusqu'à un facteur 10 !). Pour réduire cette " incertitude théorique", il est nécessaire d'identifier les réactions prépondérantes dans le système chimique, pour la production d'un composé donné, afin de proposer aux expérimentateurs de faire des mesures plus précises dans les conditions optimales de ces réactions.

2. Analyse de sensibilité

L'objectif des méthodes globales d'analyse de sensibilité est de déterminer comment réagit le système face à des perturbations globales. Pour cela, tous les facteurs d'entrée, les constantes de réactions, sont perturbés simultanément par une variation arbitraire imposée à chaque réaction. Le principe est alors d'explorer l'espace des variables d'entrée le plus efficacement possible pour avoir une bonne estimation de l'effet des perturbations imposées aux variables d'entrée sur les variables de sortie, les concentrations des composés dans cette étude. Deux méthodes globales d'analyse de sensibilité sont présentées : la méthode Monte Carlo et la méthodologie des plans d'expériences.

2.1. Méthode Monte Carlo

L'espace des constantes de réactions est balayé en utilisant une approche de type Monte Carlo. A chaque itération, un ensemble de constantes est choisi au hasard à partir de leur fonction de densité de probabilité (pdf pour Probability Density Function). Si le nombre de simulations est statistiquement suffisant, une pdf des concentrations peut être obtenue. Des calculs de corrélation sont ensuite réalisés pour identifier les constantes de réactions qui influencent fortement la concentration d'un composé donné. Cette technique a été récemment utilisée dans différents cas d'études par Carrasco (2007a), (2007b) et Dobrijevic (2008). Pour l'étude présentée, chaque constante de réaction k_i peut être considérée comme une variable aléatoire distribuée de manière log-normale sur une gamme d'incertitudes. La variable aléatoire log(k_i) suit une distribution normale centrée sur la valeur nominale log(k_{0i}) et générée par : $\log(k_i) = \log(k_{0i}) + \varepsilon_i \log(F_i)$

avec ε_i un nombre aléatoire, normalement distribué, avec une moyenne μ=0 et un écart type σ=1, F_i est le facteur d'incertitude dépendant de la température. Dans notre cas, nous avons choisi de réaliser 10000 simulations afin d'obtenir des résultats représentatifs et d'estimer précisément la moyenne et les écarts types des concentrations.

2.2. Plans d'expériences

Dans une approche par plans d'expériences, les matrices d'expériences de criblage sont utilisées pour identifier rapidement les quelques facteurs actifs parmi un grand ensemble de facteurs, avec un nombre de simulations limité. Elles sont donc très utiles pour l'étude de modèles de simulation comprenant un grand nombre de paramètres et vont permettre, dans notre cas, de repérer les réactions clés du système chimique.

2.2.1. Matrices d'Hadamard

Les matrices d'expériences de criblage les plus connues sont les matrices d'Hadamard ou matrices de Plackett et Burman (1946) pour lesquelles le nombre de simulations est proche du nombre de facteurs étudiés. Une **matrice d'Hadamard**, du nom du mathématicien français Jacques Hadamard, est une matrice carrée dont les éléments x_{ij} sont +1 ou -1 correspondant aux deux niveaux étudiés et pour laquelle la matrice d'information $X'X$ est telle que $(X'X) = N\, I_N$ avec X la matrice du modèle, X' la transposée de la matrice X, N le nombre de simulations et I_N la matrice identité.

2.2.2. Matrices supersaturées

Le nombre de simulations peut être encore réduit en utilisant des matrices supersaturées. On emploie le terme de matrices supersaturées lorsque le nombre d'expériences N est bien inférieur au nombre de facteurs étudiés. Leur utilisation repose sur deux hypothèses :

- le nombre de facteurs effectivement influents est très faible,
- les interactions sont négligeables.

Développées, dans les années 1950 par Satterwaithe (1959) de manière aléatoire et par Booth et Cox (1962) de manière systématique, on trouve depuis les années 1990 de nombreuses autres méthodes de construction permettant de réduire considérablement le nombre d'expériences.

3. Résultats

3.1. Méthode Monte Carlo (sampling-based method, Helton (2004))

Le principe général de la méthode Monte Carlo est décrit dans la figure 2. Dans une première étape, nous calculons la distribution des concentrations en suivant le modèle développé par Parisot (1998) et Dobrijevic (2003). Chaque constante chimique possède sa propre pdf construite à partir de sa valeur nominale et de son facteur d'incertitude Fj. A chaque itération, toutes les constantes de réactions sont aléatoirement choisies selon leur pdf. Ensuite le système d'équations différentielles est résolu et les concentrations à un temps donné sont enregistrées. Après un nombre suffisant d'itérations, nous obtenons les distributions pour chacun des composés.

Dans une seconde étape, nous calculons les corrélations entre les composés et les réactions pour déterminer les réactions clés. Cette méthode est basée sur le calcul des rangs de corrélations de Spearman qui permet de convertir des relations non linéaires mais monotones en une relation linéaire en remplaçant les valeurs d'entrée et de sortie par leur rang respectif. Ces coefficients de corrélation

peuvent varier entre -1 et +1. Une valeur positive indique que les deux paramètres augmentent ou diminuent simultanément.

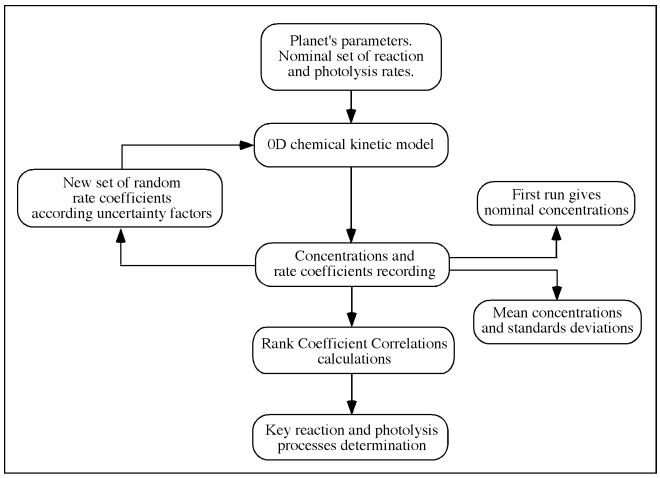

Figure 2 : Principe de la détermination des réactions-clés dans un modèle photochimique en utilisant la méthode Monte Carlo.

Les principaux coefficients de corrélations des réactions avec le composé XX sont regroupés dans le tableau ci-dessous :

Réactions	Coefficients de corrélation
2	0.123656
4	0.352642
6	0.231242
27	0.605503
28	0.454201
45	-0.322368
62	-0.137916
97	0.180021

Le calcul du coefficient de détermination de la régression est en cours de réalisation.
Au final la méthode Monte Carlo identifie **6 réactions très influentes : R_{27}, R_{28}, R_4, R_{45}, R_6, R_{97} et deux plus faibles, R_{62} et R_2**

3.2. Matrice d'Hadamard

Dans ce cas, la matrice d'expériences de criblage optimale, permettant d'estimer les "poids" des 114 facteurs à 2 niveaux (114 constantes de réactions) en un minimum d'essais, est une matrice d'Hadamard à N = 120 expériences. A partir de ces résultats, les estimations moindres carrés des coefficients du modèle polynomial de degré 1 ont été calculées. Pour plus de lisibilité, sur la figure 3, seuls les 20 coefficients les plus significatifs ont été représentés, en les classant par ordre d'importance décroissante.

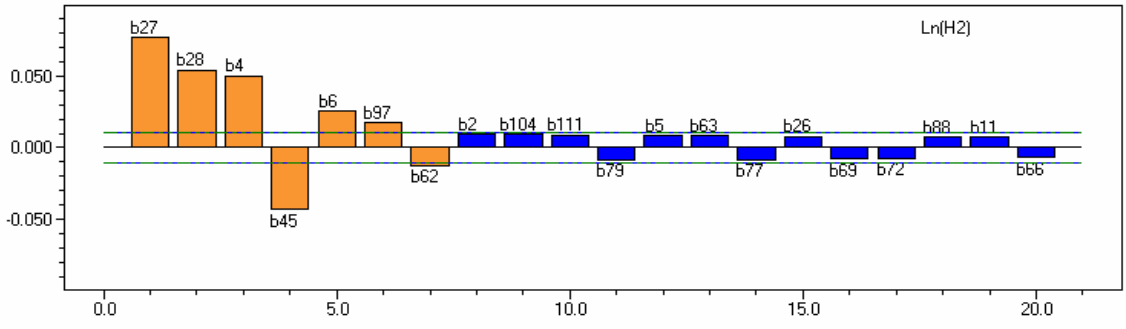

Figure 3 : Graphique des effets pour la réponse Ln(H_2)

Cet outil graphique montre un sous-ensemble de **7 réactions** ayant très probablement une influence sur la concentration étudiée **:R_{27}, R_{28}, R_4, R_{45}, R_6**, et dans une moindre mesure, **R_{97}, R_{62}**

3.3. Matrice supersaturée

Dans cette étude, nous avons construit une matrice supersaturée en utilisant la méthode proposée par Lin (1993), basée sur l'utilisation de fractions de matrices d'Hadamard. Ainsi, nous avons obtenu une matrice supersaturée à 60 expériences.

Le nombre d'expériences étant largement inférieur au nombre de facteurs, les stratégies classiques d'exploitation des résultats issus de plans d'expériences 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 ici que la méthode utilisée dans notre étude, c'est-à-dire la méthode combinant une régression Step-Wise à l'étude de toutes les régressions pour un nombre de variables donné. La première étape (Step-Wise) qui permet de sélectionner k' variables explicatives parmi les k variables étudiées présente l'inconvénient de sélectionner un modèle final qui ne contient pas toujours les facteurs réellement actifs. Westfall (1998) et Abraham (1999) recommandent alors de faire toutes les régressions à 2, 3,…, jusqu'à *f* facteurs. Plusieurs simulations ont montré qu'avec cette approche, les facteurs actifs étaient correctement identifiés. Par conséquent, lors d'une deuxième étape, tous les modèles à k' variables sont construits de manière systématique et comparés suivant leur valeur de R^2, s^2, AIC et BIC. Cependant, cette méthode n'est envisageable que lorsque le nombre de facteurs est faible car les calculs deviennent extrêmement longs et très rapidement irréalisables. Dans notre étude, le nombre de facteurs étant élevé, nous avons choisi d'appliquer cette méthode de traitement sous sa forme séquentielle, qui peut se résumer ainsi :

Sur l'ensemble des facteurs, appelé F, une méthode de type "ridge analysis" est réalisé afin d'obtenir une estimation "biaisée" des "poids" des facteurs, ce qui permet de réaliser un pré-classement des facteurs.

Etape 1 : Extraire un sous-ensemble de F, appelé F1, contenant environ 70 % des facteurs selon le classement précédent. On appelle $\overline{F1}$, le complément de l'ensemble F1.

Etape 2 : Utiliser la procédure de régression Step-Wise classique sur l'ensemble F1 pour déterminer le nombre approximatif, *f'*, de facteurs probablement influents.

Etape 3 : Réaliser toutes les régressions possibles à 2, 3 …. *f'* facteurs et retenir un ensemble F2 de facteurs probablement influents.

Etape 4 : $\overline{F1}$ et F2 sont regroupés et les étapes 2 et 3 sont répétées sur cet ensemble de ($\overline{F1}$+F2) facteurs, pour ne retenir que les facteurs probablement actifs.

Cette méthode qui reprend la combinaison de la régression Step-Wise et de toutes les régressions permet de réduire les temps de calcul et autorise ainsi son utilisation avec un grand nombre de facteurs.

Les résultats obtenus avec cette méthode pour un modèle à 4 variables sont présentés dans le tableau ci-dessous et résumés par un outil graphique appelé "carte des variables". Chaque variable présente dans la régression est coloriée en bleue. Cet outil visuel permet très rapidement de détecter les variables qui vont être conservées dans toutes les régressions (colonne bleutée).

Var	Var	Var	Var	$s^2 \times 10^3$	R^2
b4	b27	b28	b45	3.40	0.80
b27	b28	b45	b84	4.93	0.72
b4	b27	b28	b84	4.99	0.71
b4	b27	b45	b73	5.17	0.70
b27	b28	b45	b73	5.18	0.7
b4	b27	b45	b109	5.19	0.70
b4	b27	b28	b73	5.24	0.70
b27	b28	b45	b109	5.39	0.69

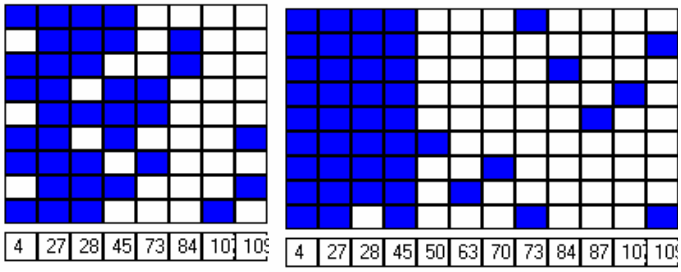

Figure 4 : Cartes des variables pour les modèles à quatre et à cinq variables, respectivement à gauche et à droite.

Finalement, l'interprétation de ces outils montrent que **4 réactions** apparaissent comme déterminantes : **R_{27}, R_{28}, R_{45} et R_4.**

3.4 Comparaison des résultats

Les réactions identifiées comme primordiales sont reportées dans le tableau ci-dessous pour chacune des trois stratégies (Méthode Monte Carlo, Matrice d'Hadamard et Matrice supersaturée).

	Monte Carlo	**Matrice d' Hadamard**	**Matrice Supersaturatée**
Nombre de simulations	10000	120	60
Réactions-clés détectées	R_{27} R_{28} R_4 R_{45} R_6 R_{97} R_{62} R_2	R_{27} R_{28} R_4 R_{45} R_6 R_{97} R_{62}	R_{27} R_{28} R_{45} R_4

La matrice d'Hadamard classique détecte 7 réactions importantes qui sont également identifiées par la méthode Monte Carlo mais avec un nombre de simulations très supérieur. Si l'on diminue encore le nombre de calculs (60 simulations) en utilisant une matrice d'expériences supersaturée, on remarque que seules les réactions les plus importantes sont identifiées. Cet outil ne fournit pas toute l'information mais permet de détecter les variables d'entrée les plus influentes. Il demeure donc un outil très intéressant dans le cas où les temps de calcul sont très longs et où il est nécessaire de repérer rapidement le comportement global du phénomène étudié.

Conclusion

L'utilisation des plans d'expériences dans le domaine de la simulation numérique permet de réaliser des études de sensibilité et d'obtenir des résultats comparables aux méthodes Monte Carlo tout en réduisant considérablement le nombre de simulations. Pour réaliser une étude comparative significative, des études basées sur la technique Monte Carlo avec un nombre de simulations du même ordre de grandeur que celui des plans d'expériences (60 et 120) seront également présentées.

Bibliographie

[1] Booth K. H. V. et D. R. Cox (1962) Some systematic supersaturated designs. *Technometrics,* **4,** 489-495.

[2] J. C. Helton, J. D. Johnson, C.J. Sallaberry, C.B. Storlie (2006) Survey of sampling-based methods for uncertainty and sensitivity analysis, *Reliability Engineering & System Safety*, **91**, 1175-1209

[3a] Carrasco, N., Dutuit, O., Thissen, R., Banaszkiewicz, M., & Pernot, P. (2007a), *Planet. Space Sci*., 55, 141

[3b] Carrasco, N., Hébrard, E., Banaszkiewicz, M., Dobrijevic, M., & Pernot, P. (2007b*), Icarus*, 192, 519

[4] Dobrijevic, M., Carrasco, N., Hébrard, E., & Pernot, P. (2008), *Submitted to Planet. Space Sci*.

[5] Dobrijevic, M., Ollivier J.L., Billebaud F., Brillet J., Parisot J.P. (2003) Effect of chemical kinetics uncertainties on photochemical modeling results: application to Saturn's atmosphere. *Astronomy and Astrophysics*. Vol. 398, 335-344.

[6] Dobrijevic, M.et Parisot, J.P. (1998), Effect of chemical kinetics uncertainties on hydrocarbons production in the stratosphere of Neptune. *Planetary and Space Science*, vol. 46, 491-505.

[7] Abraham B. et H. Chipman (1999). Some risks in the construction and analysis of supersatured designs. *Technometrics,* **41**(2), 135-141.

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

[9] Plackett R.L. et J.P. Burman (1946). Design of optimum multifactorial experiments. *Biometrika*, **33**, 305-325.

[10] Satterwaithe F. (1959). Random balance experimentation (with discussion). *Technometrics,* **1,** 111-137.

[11]Westfall Westfall, P. H. et S. S. Young (1998). Forward selection error control in the analysis of supersatured designs. *Statistica Sinica,* **8,** 101-117.