PUB. IRMA, LILLE 2011
Vol. 71, N° III

Pratique de l'analyse de sensibilité : comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique*

Julien JACQUES[a]

Abstract

Global sensitivity analysis (SA) analyze a mathematical model by studying the impact of the variability of model input factors on the output variable. Determining the inputs responsible for this variability with sensitivity indices, SA allows to reduce the variance of output if it is synonymous with vagueness, or ease the model by setting the entries whose variability does not influence the output variable. We present in this paper the main sensitivity indices, based on the assumption of independent input variables, their estimates, and then address the case of non-independent input models. Two numerical applications illustrate the intepretation of sensitivity indices in the case of models with independent and dependent inputs.

Résumé

L'analyse de sensibilité globale (AS) permet d'analyser un modèle mathématique en étudiant l'impact de la variabilité des facteurs d'entrée du modèle sur la variable de sortie. Déterminant les entrées responsables de cette variabilité à l'aide d'indices de sensibilité, l'AS permet de prendre les mesures nécessaires pour diminuer la variance de la sortie si celle-ci est synonyme d'imprécision, ou encore d'alléger le modèle en fixant les entrées dont la variabilité n'influe pas la variable de sortie. Nous présentons dans ce document les principaux indices de sensibilité, basés sur l'hypothèse d'indépendance des variables d'entrée, leurs estimations, puis abordons le cas des modèles à entrées non indépendantes. Deux applications numériques illustrent l'inteprétation des indices de sensibilité dans le cas de modèle à entrées indépendantes et dépendantes.

MSC 2009 subject classifications. 49Q12.

Mots clés. Analy e de en ibilité, décompo ition de la variance, indice de Sobol, entrée dépendante .

*Preprint.

[a] Laboratoire P. Painlevé, UMR 8524 CNRS Université Lille I, Bât M2, Cité Scientifique, F-59655 Villeneuve d'Ascq Cedex, France.

1 Introduction : les objectifs de l'analyse de sensibilité

Con idéron un modèle mathématique qui, à un en emble de variable d'entrée aléatoire $\mathbf{X}$, fait corre pondre, via une fonction f détermini te, une variable de ortie Y (ou répon e) aléatoire :

$$\begin{aligned} f : \mathbb{R}^p &\to \mathbb{R} \\ \mathbf{X} &\mapsto Y = f(\mathbf{X}) \end{aligned} \tag{1}$$

La fonction f du modèle peut être trè complexe (y tème d'équation différentielle ...), et e t en pratique évaluée à l'aide d'un code informatique, plu ou moin onéreux en temp de calcul. L'en emble de variable d'entrée $\mathbf{X} = (X_1, \ldots, X_p)$ regroupe toute le entité con idérée comme aléatoire dan le modèle.

L'analy e de en ibilité étudie comment de perturbation ur le variable d'entrée du modèle engendrent de perturbation ur la variable répon e. L'auteur intere é par un ouvrage de référence pourra e référer à [11]. Il e t po ible de grouper le méthode d'analy e de en ibilité en troi cla e : le méthode de *screening*, qui con i tent en une analy e qualitative de la en ibilité de la variable de ortie aux variable d'entrée, le méthode d'analy e locale [17], qui évaluent quantitativement l'impact d'une petite variation autour d'une valeur donnée de entrée , et enfin le méthode d'analy e de en ibilité globale, qui 'intére ent à la variabilité de la ortie du modèle dan l'intégralité de on domaine de variation. L'analy e de en ibilité globale étudie comment la variabilité de entrée e répercute ur celle de la ortie, en déterminant quelle part de variance de la ortie e t due à telle entrée ou tel en emble d'entrée . Si l'analy e de en ibilité locale 'intére e plu à la valeur de la variable répon e, l'analy e de en ibilité globale 'intére e quant à elle à a variabilité. Nou nou intére on dan ce document à l'analy e de en ibilité globale et omettron donc par la uite l'adjectif global.

Le enjeux de l'analy e de en ibilité peuvent être multiple : validation d'une méthode ou d'un code de calcul, orientation de effort de recherche et développement, ou encore ju tification en terme de ûreté d'un dimen ionnement ou d'une modification d'un y tème. Nou décrivon ci-aprè le principale que - tion auxquelle l'analy e de en ibilité permet d'apporter de élément de répon e.

Les ambitions de l'analyse de sensibilité Au cour de l'élaboration, de la con truction ou de l'utili ation d'un modèle mathématique, l'analy e de en ibilité peut 'avérer être un outil précieux. En effet, en étudiant comment la répon e du modèle réagit aux variation de e variable d'entrée, l'analy e de en ibilité permet de répondre à un certain nombre de que tion .

1. Le modèle e t-il bien fidèle au phénomène/proce u modéli é ?
   En effet, i l'analy e exhibe une forte influence d'une variable d'entrée

habituellement connue comme non influente, il era néce aire de remettre en cau e la qualité du modèle ou (et) la véracité de no connai ance ur l'impact réel de variable d'entrée.

2. Quelle ont le variable qui contribuent le plu à la variabilité de la répon e du modèle ?
   Si cette variabilité e t ynonyme d'impréci ion ur la valeur prédite de la ortie, il era alor po ible d'améliorer la qualité de la répon e du modèle à moindre coût. En effet, la variabilité de la ortie du modèle pourra être diminuée en concentrant le effort ur la réduction de variabilité de entrée le plu influente . Il doit être préci é que cela n'e t pa toujour po ible, notamment lor que la variabilité d'une variable d'entrée e t intrin èque à la nature de la variable et non due à un manque d'information ou à de impréci ion de me ure .

3. Quelle ont au contraire le variable le moin influente ?
   Il era po ible de le con idérer comme de paramètre détermini te , en le fixant par exemple à leur e pérance, et obtenir ain i un modèle plu *léger* avec moin de variable d'entrée. Dan le ca d'un code informatique, il era po ible de upprimer le partie de code qui n'ont aucune influence ur la valeur et la variabilité de la répon e.

4. Quelle variable , ou quel groupe de variable , interagi ent avec quelle (quel) autre ?
   L'analy e de en ibilité peut permettre de mieux appréhender et comprendre le phénomène modéli é, en éclairant le relation entre le variable d'entrée.

Bon nombre de publication ur le ujet explicitent et illu trent ce objectif . On pourra e référer notamment aux travaux de Saltelli et al. [12, 13, 15].

La ection uivante pré ente le indice de en ibilité défini pour de modèle à variable d'entrée indépendante , ain i que leur méthode d'e timation. La ection 3 'intére e aux modèle à entrée non indépendante , et pré ente deux type d'indice utili able dan ce ca . Enfin, la ection 4 pré ente deux application imulée illu trant l'intérêt et l'interprétation de indice de en ibilité , dan le ca de modèle à entrée indépendante et non indépendante .

2 Indicateurs de sensibilité pour modèles à entrées indépendantes

Nou uppo on dan cette ection que le variable d'entrée $\mathbf{X} = (X_1, \ldots, X_p)$ du modèle ont indépendante .

2.1 Préambule : cas du modèle linéaire

Suppo on que le modèle étudié oit linéaire, et qu'il 'écrive ou la forme uivante :

$$Y = {}_0 + \sum_{i=1}^{p} {}_i X_i : \tag{2}$$

Comme le variable X_i ont uppo ée indépendante , la variance de Y 'écrit alor :

$$\mathrm{V}(Y) = \sum_{i=1}^{p} {}_i^2 \mathrm{V}(X_i);$$

où ${}_i^2\mathrm{V}(X_i)$ e t la part de variance due à la variable X_i. La en ibilité de Y à X_i peut donc implement être quantifié par le rapport de la part de variance due à X_i ur la variance totale. On définit ain i l'indice de en ibilité SRC (*Standardized Regression Coefficient*) :

$$SRC_i = \frac{{}_i^2\mathrm{V}(X_i)}{\mathrm{V}(Y)} : \tag{3}$$

Il exprime la part de variance de la répon e Y due à la variance de la variable X_i. Cet indice SRC, toujour po itif ($SRC \in [0; 1]$), e t en outre le carré du coefficient de corrélation linéaire entre la répon e du modèle et e variable d'entrée.

2.2 Les indices de Sobol

Plaçon nou dé ormai dan le ca d'une fonction f dont la forme analytique n'e t pa connue. Pour apprécier l'importance d'une variable d'entrée X_i ur la variance de la ortie Y, nou étudion à combien la variance de Y décroît i on fixe la variable X_i à une valeur x_i : $\mathrm{V}(Y|X_i = x_i)$. Le problème de cet indicateur e t le choix de la valeur x_i de X_i, que l'on ré out en con idèrant l'e pérance de cette quantité pour toute le valeur po ible de x_i : $\mathbf{E}[\mathrm{V}(Y|X_i)]$. Ain i, plu la variable X_i era importante vi -à-vi de la variance de Y, plu cette quantité era petite. Etant donné la formule de la variance totale $\mathrm{V}(Y) = \mathrm{V}(\mathbf{E}[Y|X_i]) + \mathbf{E}[\mathrm{V}(Y|X_i)]$, nou pouvon utili er de façon équivalente la quantité

$$\mathrm{V}(\mathbf{E}[Y|X_i]);$$

qui era d'autant plu grande que la variable X_i era importante vi -à-vi de la variance de Y. Afin d'utili er un indicateur normali é, nou défini on l'indice de en ibilité de Y à X_i :

$$S_i = \frac{\mathrm{V}(\mathbf{E}[Y|X_i])}{\mathrm{V}(Y)} : \tag{4}$$

Cet indice e t appelé **indice de sensibilité de premier ordre** par Sobol [16], *correlation ratio* par McKay [5], ou encore *importance measure*. Il quantifie la

en ibilité de la ortie Y à la variable d'entrée X_i, ou encore la part de variance de Y due à la variable X_i.

Remarque. Dan le ca du modèle linéaire (2), cet indice de en ibilité e t égal à l'indice SRC, pui que $\mathrm{V}(\mathbf{E}[Y|X_i]) = \mathrm{V}(\ _iX_i) = \ _i{}^2\mathrm{V}(X_i)$.

Sobol [16] a introduit cet indice de en ibilité en décompo ant la fonction f du modèle en omme de fonction de dimen ion roi ante :

$$\begin{aligned} Y &= f(X_1;:::;X_p) && (5)\\ &= f_0 + \sum_{i=1}^{p} f_i(X_i) + \sum_{1 \le i<j \le p} f_{ij}(X_i;X_j) + :::+ f_{1,...p}(X_1;:::;X_p): && (6) \end{aligned}$$

où

$$\begin{aligned} f_0 &= \mathbf{E}[Y];\\ f_i(X_i) &= \mathbf{E}[Y|X_i] - \mathbf{E}[Y];\\ f_{i,j}(X_i;X_j) &= \mathbf{E}[Y|X_i;X_j] - \mathbf{E}[Y|X_i] - \mathbf{E}[Y|X_j] + \mathbf{E}[Y];\\ f_{i,j,k}(X_i;X_j;X_k) &= \mathbf{E}[Y|X_i;X_j;X_k] - \mathbf{E}[Y|X_i;X_j] - \mathbf{E}[Y|X_i;X_k]\\ &\quad -\mathbf{E}[Y|X_j;X_k]:::\end{aligned}$$

La variance de Y, V, peut alor e décompo er elon le théorème uivant.

Théorème. Décompo ition de Sobol de la variance.
La variance du modèle à entrée indépendante (1) e décompo e en :

$$V = \sum_{i=1}^{p} V_i + \sum_{1 \le i<j \le p} V_{ij} + :::+ V_{1...p}; \qquad (7)$$

où

$$\begin{aligned} V_i &= \mathrm{V}(\mathbf{E}[Y|X_i]);\\ V_{ij} &= \mathrm{V}(\mathbf{E}[Y|X_i;X_j]) - V_i - V_j;\\ V_{ijk} &= \mathrm{V}(\mathbf{E}[Y|X_i;X_j;X_k]) - V_{ij} - V_{ik} - V_{jk} - V_i - V_j - V_k;\\ &:::\\ V_{1...p} &= V - \sum_{i=1}^{p} V_i - \sum_{1 \le i<j \le p} V_{ij} - ::: - \sum_{1 \le i_1<...<i_{p-1} \le p} V_{i_1...i}\end{aligned}$$

expriment la en ibilité de la variance de Y aux variable X_i, X_j et X_k qui n'e t pa pri e en compte dan l'effet de variable eule et de interaction deux à deux. Et ain i de uite ju qu'à l'ordre p.
L'**interprétation de ces indices est facile**, pui que grâce à (7), **leur somme est égale à 1**, et étant tou po itif , plu l'indice era grand (proche de 1), plu la variable aura d'importance.

Le nombre d'indice de en ibilité ain i con truit, de l'ordre 1 à l'ordre p, e t égale à $2^p - 1$. Lor que le nombre de variable d'entrée p e t trop important, le nombre d'indice de en ibilité explo e. L'e timation et l'interprétation de tou ce indice deviennent vite impo ible. Homma et Saltelli [2] ont alor introduit de indice de en ibilité totaux, qui expriment la en ibilité totale de la variance Y à une variable, c'e t-à-dire la en ibilité à cette variable ou toute e forme (en ibilité à la variable eule et en ibilité aux interaction de cette variable avec d'autre variable).
L'**indice de sensibilité total** S_{T_i} à la variable X_i e t défini comme la omme de tou le indice de en ibilité relatif à la variable X_i :

$$S_{T_i} = \sum_{k \# i} S_k. \tag{8}$$

où $\#i$ repré ente tou le en emble d'indice contenant l'indice i.
Exemple : pour un modèle à troi variable d'entrée $S_{T_1} = S_1 + S_{12} + S_{13} + S_{123}$.

2.3 Estimation des indices de Sobol

Estimation de Monte Carlo Dan beaucoup de problème cientifique , on e t amené à calculer une intégrale du type

$$I = \int_D f(\mathbf{x}) d\mathbf{x},$$

où D e t un e pace de plu ou moin grande dimen ion, et f une fonction (intégrable). Soit $x_1, \ldots, x_N$ la réali ation d'un N-échantillon d'une variable aléatoire uniforme ur D. Nou uppo on cet échantillon pri de manière totalement aléatoire (échantillonnage aléatoire). Une approximation de I par la méthode de Monte Carlo e t faite par :

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i).$$

La convergence (pre que ûre) de I_N ver I découle directement de la loi forte de grand nombre . Cette méthode d'e timation permet alor d'e timer l'e pérance de toute fonction d'une variable aléatoire de den ité quelconque par

$$\hat{\mathbf{E}}[f(X)] = \frac{1}{N} \sum_{i=1}^{N} f(x_i),$$

où $(x_i)_{i=1..N}$ e t un N-échantillon de réali ation de la variable aléatoire X. Le taux de convergence d'une méthode de Monte Carlo e t en $\mathcal{O}(N^{\frac{1}{2}})$.

Bon nombre de méthode alternative ont été propo ée pour améliorer la convergence, parmi le quelle le méthode de imulation p eudo-probabili te [1], comme l'échantillonnage tratifié ou par hypercube latin (*LHS*) [6], le méthode de Qua i-Monte Carlo [7], ou encore le méthode de Qua i-Monte Carlo Randomi é [9]. L'échantillonnage tratifié con i te à découper l'e pace de variable d'entrée en petit e pace di joint , pui à échantillonner au ein de chacun de ce ou e pace . L'échantillonnage *LHS* e t ba é ur le même principe, en 'a - urant que le découpage a défini de e pace équiprobable , et que chaque e pace e t bien échantillonné ; le quadrillage e fait dan le cube unité, pour un tirage aléatoire d'échantillon uniforme, pui ce échantillon ont tran formé via la fonction de répartition inver e. Le méthode de Qua i-Monte Carlo ont de ver ion détermini te de méthode de Monte Carlo. Ce méthode défini ent de équence d'échantillon détermini te qui ont une di crépance plu faible que le équence aléatoire , c'e t-à-dire qu'elle ont une meilleure répartition uniforme dan l'e pace de variable d'entrée. Ce méthode de qua i-Monte Carlo permettent d'obtenir une convergence plu rapide en $\mathcal{O}(N^{-1}(\log N)^{p-1})$ (ou de condition relativement faible de régularité de f). Le méthode de Qua i-Monte Carlo Randomi é, ou certaine condition peu re trictive ur f, ont un taux de convergence en $\mathcal{O}(N^{-\frac{3}{2}}(\log N)^{\frac{p-1}{2}})$, et permettent une approximation de l'erreur d'e timation. Owen [8] pré ente ce méthode comme une ré-randomi ation de équence utili ée dan le méthode de qua i-Monte Carlo : on prend le équence détermini te a_i de ce dernière , et on le tran - forme en variable aléatoire x_i. Cette tran formation e fait par exemple par $x_i = a_i + U \mod 1$; où $U \sim U[0;1]^p$.

Estimation des indices de sensibilité par Monte Carlo

Con idéron un N-échantillon $\tilde{X}_{(N)} = (x_{k1};\dots;x_{kp})_{k=1..N}$ de réali ation de variable d'entrée $(X_1;\dots;X_p)$. L'e pérance de Y, $\mathbf{E}[Y] = f_0$, et a variance, $\mathrm{V}(Y) = V$, ont e timée par :

$$\hat{f_0} = \frac{1}{N}\sum_{k=1}^{N} f(x_{k1};\dots;x_{kp}); \qquad \text{et} \qquad \hat{V} = \frac{1}{N}\sum_{k=1}^{N} f^2(x_{k1};\dots;x_{kp}) - \hat{f_0}^2; \quad (9)$$

L'e timation de indice de en ibilité néce ite l'e timation d'e pérance de variance conditionnelle. Nou pré enton une technique d'e timation due à Sobol [16].

L'e timation de indice de en ibilité de premier ordre (4) con i te à e timer la quantité :

$$V_i = \mathrm{V}(\mathbf{E}[Y|X_i]) = \underbrace{\mathbf{E}[\mathbf{E}[Y|X_i]^2]}_{U_i} - \mathbf{E}[\mathbf{E}[Y|X_i]]^2 = U_i - \mathbf{E}[Y]^2;$$

[1] «pseudo» puisqu'elle consiste en un échantillonnage non totalement aléatoire

la variance de Y étant e timée cla iquement par (9). Sobol propo e d'e timer la quantité U_i, c'e t-à-dire l'e pérance du carré de l'e pérance de Y conditionnellement à X_i, comme une e pérance cla ique, mai en tenant compte du conditionement à X_i en fai ant varier entre le deux appel à la fonction f toute le variable auf la variable X_i. Ceci néce ite deux échantillon de réali ation de variable d'entrée, que nou noton $\tilde{X}^{(1)}_{(N)}$ et $\tilde{X}^{(2)}_{(N)}$:

$$\hat{U}_i = \frac{1}{N}\sum_{k=1}^{N} \quad f\left(x^{(1)}_{k1};\dots;x^{(1)}_{k(i\ \ 1)};x^{(1)}_{ki};x^{(1)}_{k(i+1)};\dots;x^{(1)}_{kp}\right)$$
$$f\left(x^{(2)}_{k1};\dots;x^{(2)}_{k(i\ \ 1)};x^{(1)}_{ki};x^{(2)}_{k(i+1)}\dots;x^{(2)}_{kp};\right):$$

Le indice de en ibilité de premier ordre ont alor e timé par :

$$\hat{S}_i = \frac{\hat{V}_i}{\hat{V}} = \frac{\hat{U}_i - \hat{f}_0^{\,2}}{\hat{V}}:$$

Pour le indice de en ibilité de econd ordre $S_{ij} = \frac{V_{ij}}{V}$, où :

$$V_{ij} = \mathrm{V}(\mathbf{E}[Y|X_i;X_j]) - V_i - V_j = U_{ij} - \mathbf{E}[Y]^2 - V_i - V_j;$$

nou e timon le quantité $U_{ij} = \mathbf{E}[\mathbf{E}[Y|X_i;X_j]^2]$ de la même manière, en fai ant varier entre le deux appel à la fonction toute le variable auf X_i et X_j :

$$\hat{U}_{ij} = \frac{1}{N}\sum_{k=1}^{N} f\left(x^{(1)}_{k1};\dots;x^{(1)}_{k(i\ \ 1)};x^{(1)}_{ki};x^{(1)}_{k(i+1)};\dots;x^{(1)}_{k(j\ \ 1)};x^{(1)}_{kj};x^{(1)}_{k(j+1)};\dots;x^{(1)}_{kp}\right)$$
$$\times f\left(x^{(2)}_{k1};\dots;x^{(2)}_{k(i\ \ 1)};x^{(1)}_{ki};x^{(2)}_{k(i+1)};\dots;x^{(2)}_{k(j\ \ 1)};x^{(1)}_{kj};x^{(2)}_{k(j+1)};\dots;x^{(2)}_{kp}\right):$$

L'indice S_{ij} e t alor e timé par :

$$\hat{S}_{ij} = \frac{\hat{U}_{ij} - \hat{f}_0^{\,2} - \hat{V}_i - \hat{V}_j}{\hat{V}}:$$

Et ain i de uite pour le indice de en ibilité d'ordre upérieur.

Remarque. L'e timation de indice de en ibilité d'ordre i, $(1 < i \leq p)$, néce ite l'e timation de indice de en ibilité d'ordre 1 à $i-1$.

Par contre, le indice de en ibilité totaux peuvent être e timé directement. En effet, on remarque facilement que l'indice de en ibilité total peut 'écrire

$$S_{T_i} = 1 - \frac{\mathrm{V}(\mathbf{E}[Y|X_{\ \ i}])}{\mathrm{V}(Y)} = 1 - \frac{V_{\ \ i}}{V}:$$

où $V_{\ \ i}$ e t la variance de l'e pérance de Y conditionnellement à toute le variable auf X_i. $V_{\ \ i}$ e t alor e timée comme V_i, auf qu'au lieu de faire varier

toute le variable auf X_i, nou ne fai on varier uniquement X_i.
Ain i, pour e timer $V_{\ i} = \mathbf{E}[\mathbf{E}[Y|X_{\ i}]^2] - \mathbf{E}[\mathbf{E}[Y|X_{\ i}]]^2 = U_{\ i} - \mathbf{E}[Y]^2$, on e time $U_{\ i}$ par :

$$\hat{U}_{\ i} = \frac{1}{N}\sum_{k=1}^{N} f\left(x_{k1}^{(1)};\dots;x_{k(i\ \ 1)}^{(1)};x_{ki}^{(1)};x_{k(i+1)}^{(1)};\dots;x_{kp}^{(1)}\right) f\left(x_{k1}^{(1)};\dots;x_{k(i\ \ 1)}^{(1)};x_{ki}^{(2)};x_{k(i+1)}^{(1)}\dots;x_{kp}^{(1)};\right);$$

et on obtient

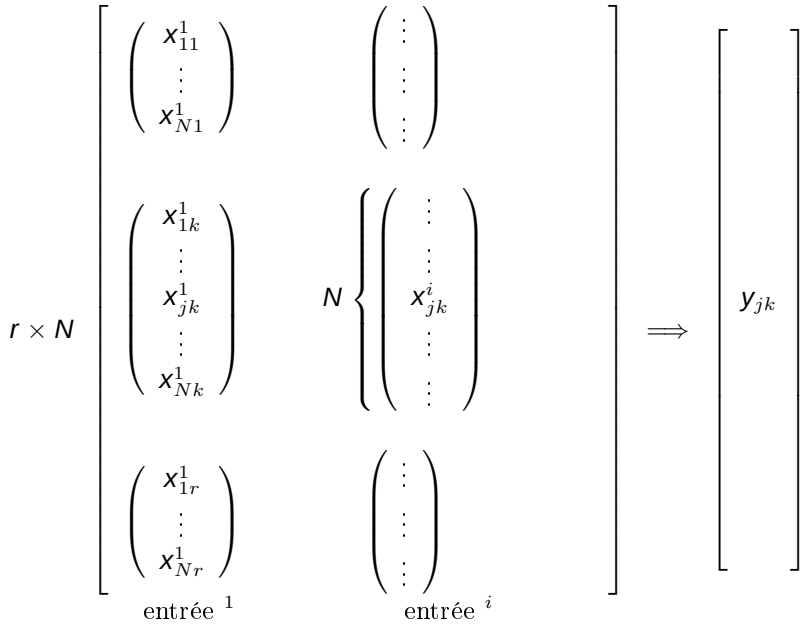
$$\hat{S}_{T_i} = 1 - \frac{\hat{U}_{\ i} - \hat{f}_0^{\ 2}}{\hat{V}}:$$

Quels indices estimer : stratégie à adopter En utili ant une taille d'échantillon de Monte Carlo de N, le nombre réel de imulation de variable d'entrée néce aire à l'e timation de indice de en ibilité e t $2N$, pui que cette e timation néce ite deux jeux de imulation . Le nombre d'appel à la fonction du modèle e t alor $N \times (k+1)$, où k e t le nombre d'indice e timé . Pour un modèle à p variable d'entrée, l'e timation de tou le indice de en ibilité néce ite $N \times (2^p)$ appel à la fonction. En revanche, n'e timer que le indice de premier ordre et le indice totaux ne demande que $N \times (2p+1)$ appel .
Il conviendra donc d'e timer dan un premier temp le indice de permier ordre et le indice totaux. S'il exi te de écart important entre ce deux indice , c'e t que la part de interaction e t non négligeable et il peut être utile d'e timer le indice d'ordre intermédiaire . Dan le ca contraire, l'effet de variable d'entrée era principalement de premier ordre et il ne era pa utile de 'intére er aux indice d'ordre intermédiaire .
En pratique, une taille d'échantillon de l'ordre de 10000 era uffi ante pour e - timer le indice de en ibilité d'un modèle comportant une dizaine de variable d'entrée. En outre, il era po ible d'e timer la variabilité de e timateur obtenu par boo trap. Lor que le modèle demande un temp d'exécution important, il e t illu oire de vouloir utili er de telle taille d'échantillon en un temp rai onnable. On a en général recourt à une approximation de la fonction f (urface de répon e), permettant de faire de imulation inten ive et donc d'e timer le indice de en ibilité. Le lecteur intere é par une revue de méthode de urface de répon e pour l'analy e de en ibilité pourra e référer à [3] par exemple.

2.3.1 La méthode de McKay

La méthode d'e timation de indice de en ibilité de premier ordre propo ée par McKay, [5], e ba e ur l'échantillonnage par hypercube latin répliqué (*r-LHSampling*). À partir d'un N-échantillon créé elon le plan d'échantillonnage par hypercube latin (N première ligne de la matrice ci-de ou), on crée r réplication (paquet de N ligne) en permutant indépendamment et aléatoirement le N valeur de chaque variable (i.e. colonne). La réunion de ce r réplication

donnera $N \times r$ échantillon pour chaque variable. Ce chéma d'échantillonnage par hypercube latin répliqué peut être repré enté par la figure 1.



$1 \leq j \leq N$: N valeur de entrée (pri e dan de intervalle équiprobable),
$1 \leq k \leq r$: r permutation de N-vecteur de imulation de entrée ,
$1 \leq i \leq p$: p paramètre .

FIGURE 1 – Échantillonnage par hypercube latin répliqué.

Le moyenne uivante ont alor définie :

$$\overline{y}_{j.} = \frac{1}{r}\sum_{k=1}^{r} y_{jk} \qquad \overline{y} = \frac{1}{N}\sum_{j=1}^{N} \overline{y}_{j.};$$

où $\overline{y_{j.}}$ e t la moyenne *inter* réplication et $\overline{y}$ e t la moyenne ur toute le valeur de y.

L'e timation de l'indice de en ibilité de premier ordre de la variable X_i, défini par (4) néce ite l'e timation de quantité $\mathrm{V}(\mathbf{E}[Y|X_i])$ et $\mathrm{V}(Y)$. La variance

totale $\mathrm{V}(Y)$ peut être e timée par :

$$\widehat{\mathrm{V}^{(\ )}(Y)} = \frac{1}{r}\sum_{k=1}^{r}\underbrace{\frac{1}{N}\sum_{j=1}^{N}(y_{jk}-\overline{y}_{.k})^2}_{\widehat{\mathrm{V}_k(Y)}}; \tag{10}$$

où $\overline{y}_{.k} = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} y_{jk}$ et $\widehat{\mathrm{V}_k(Y)}$ ont le e timation de la moyenne et de la variance de Y au ein de la réplication k (*intra* réplication). En utili ant la formule cla ique de l'analy e de la variance, pour une omme de carré *intra* et *inter* réplication , qui 'écrit :

$$\begin{aligned}\sum_{k=1}^{r}\sum_{j=1}^{N}(y_{jk}-\overline{y})^2 &= \underbrace{\sum_{k=1}^{r}\sum_{j=1}^{N}(\overline{y}_{.k}-\overline{y})^2}_{inter} + \underbrace{\sum_{k=1}^{r}\sum_{j=1}^{N}(y_{jk}-\overline{y}_{.k})^2}_{intra}\\ &= N\sum_{k=1}^{r}(\overline{y}_{.k}-\overline{y})^2 + \sum_{k=1}^{r}\sum_{j=1}^{N}(y_{jk}-\overline{y}_{.k})^2;\end{aligned}$$

on a :

$$\widehat{\mathrm{V}^{(\ )}(Y)} = \frac{1}{Nr}\sum_{k=1}^{r}\sum_{j=1}^{N}(y_{jk}-\overline{y})^2 - \frac{1}{r}\sum_{k=1}^{r}(\overline{y}_{.k}-\overline{y})^2.$$

Or, pour un échantillonnage *LHS*, comme $\mathbf{E}\left[(\overline{y}_{.k}-\overline{y})^2\right]$ e t en $\frac{1}{N}$, le dernier terme de cette égalité peut être con idéré comme négligeable pour une taille d'échantillon N uffi amment grande. McKay propo e alor l'e timation de la variance totale uivante :

$$\widehat{\mathrm{V}(Y)} = \frac{1}{Nr}\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk}-\overline{y})^2.$$

Soient $\overline{Y}_{j.}$ et $\overline{Y}$ le variable aléatoire dont $\overline{y}_{j.}$ et $\overline{y}$ ont le réali ation ur notre matrice d'échantillonnage. Comme :

$$\begin{aligned}\mathbf{E}\left[(\overline{Y}_{j.}-\overline{Y})^2\right] \simeq \mathrm{V}(\overline{Y}_{j.}) &= \mathrm{V}(\mathbf{E}\left[\overline{Y}_{j.}|X_i\right]) + \mathbf{E}\left[\mathrm{V}(\overline{Y}_{j.}|X_i)\right]\\ &= \mathrm{V}(\mathbf{E}\left[Y|X_i\right]) + \frac{1}{r}\mathbf{E}\left[\mathrm{V}(Y|X_i)\right];\end{aligned}$$

le terme $\mathrm{V}(E[Y|X_i])$ e t e timé par :

$$\frac{1}{N}\sum_{j=1}^{N}(\overline{y}_{j.}^{(i)}-\overline{y})^2 - \frac{1}{r}\frac{1}{Nr}\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk}^{(i)}-\overline{y}_{j.})^2;$$

où $\frac{1}{Nr}\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk}^{(i)} - \bar{y}_{j.})^2$ e t l'e timateur de $\mathbf{E}\left[\mathrm{V}(Y|X_i)\right]$, avec $y_{jk}^{(i)}$ et $\bar{y}_{j.}^{(i)}$ obtenu en fixant, pour la variable X_i, le r réplication , (x_{jk}^i con tant ur k, c'e t-à-dire $x_{j1}^i = x_{j2}^i = \cdots = x_{jr}^i$ pour tout $1 \leq j \leq N$).
L'indice de en ibilité de premier ordre de la variable X_i, défini par (4) e t alor e timé par :

$$S_i = \frac{r\sum_{j=1}^{N}(\bar{y}_{j.}^{(i)} - \bar{y})^2 - \frac{1}{r}\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk}^{(i)} - \bar{y}_{j.})^2}{\sum_{j=1}^{N}\sum_{k=1}^{r}(y_{jk} - \bar{y})^2}.$$

3 Modèles à entrées dépendantes

L'hypothè e de l'indépendance de facteur d'entrée faite précédemment e t néce aire pour garantir l'interprétabilité de indice (un indice d'ordre un n'exprime plu la en ibilité à une unique variable i cette dernière e t corrélée avec d'autre) et la validité de leur méthode d'e timation par Monte-Carlo (le intégrale multidimen ionnelle ont évaluée comme de produit d'intégrale unidimen ionnelle).
Nou pré enton dan cette ection le tratégie po ible pour réali er une analy e de en ibilité ur un modèle à variable d'entrée non indépendante .

3.1 Indices multidimensionnels

Lor que toute le variable d'entrée ne ont pa dépendante , mai qu'elle peuvent être regroupée en clu ter de variable dépendante (le variable au ein d'un clu ter ont dépendante mai le variable de différent clu ter ont indépendante), il e t po ible de con idérer de indice de en ibilité mutlidimen ionnel [4] qui expriment la en ibilité de la variance de Y à un clu ter de facteur .
Si par exemple le deux variable X_i et X_j ont dépendante , mai indépendante du re te de autre variable , la en ibilité à la variable bidimen ionnelle $(X_i; X_j)$ era exprimé par l'indice multidimen ionnel

$$S_{fi,jg} = \frac{V\left(E[Y|X_i; X_j]\right)}{V(Y)}.$$

Il e t po ible de définir de indice d'ordre upérieur exprimant la en ibilité de Y à l'interaction entre cette variable bidimen ionnelle $(X_i; X_j)$ et d'autre variable uni ou multidimen ionnelle . Le clu ter de variable étant indépendant entre eux, l'interprétabilité (et en particulier la ommation de indice de tout ordre à 1) e t con ervée.
L'e timation de ce indice peut être faite par Monte Carlo avec une approche

imilaire à celle utili ée pour e timer le indice de en ibilité de Sobol cla ique (cf. [4] pour plu de détail).

3.2 Utilisation des indices de Sobol d'ordre 1

Lor que l'analy e de en ibilité e t menée dan le but de avoir quelle variable ou quel groupe de variable qui, une foi fixé, conduit à la plu grande réduction de la variance de Y, Saltelli et Tarantola [14] expliquent que le indice de en ibilité d'ordre un ont toujour le indicateur à utili er en pré ence de corrélation. En effet, i en pré ence de corrélation l'indice d'ordre un S_i n'exprime plu uniquement la en ibilité à une variable X_i mai également une partie de en ibilité aux variable avec le quelle elle e t corrélée, fixer X_i conduit à également jouer ur la di tribution de variable avec le quelle elle e t corrélée, et donc conduit à réduire d'autant plu la variance de la répon e du modèle.
Si l'e timation de Monte-Carlo de indice de premier ordre pré entée précédemment (ection 2.3) n'e t plu valable en l'ab ence d'indépendance entre le variable d'entrée, la méthode de McKay (ection 2.3.1) e t toujour valable. Néanmoin , cette méthode d'e timation e t trè gourmande en nombre d'évaluation de la fonction f du modèle, ce qui peut être problématique lor que l'évaluation de f e t coûteu e en temp de calcul. Nou pré enton ci-aprè une méthode d'e timation par polynôme locaux rédui ant con idérablement ce nombre d'évaluation [1].

Estimation par polynômes locaux La méthode d'e timation de indice de en ibilité d'ordre 1 de Da Veiga [1] con i te à e timer dan un premier temp l'e pérance de Y conditionnellement à chaque variable d'entrée X_i, pui dan un econd temp à e timer la variance de cette e pérance conditionnelle pour obtenir l'e timateur de l'indice de en ibilité. L'avantage principal de cette méthode e t qu'elle ne fait appel qu'à un nombre réduit d'appel à la fonction, contrairement à la méthode de McKay précédente.
Noton $m_i(x) = E[Y|X_i = x]$. On approche $m(x)$ localement par un polynôme

$$m_i(z) \simeq \sum_{j=0}^{p} {}_j(z-x)^j \qquad \forall z \in \mathcal{V}(x)$$

où $\mathcal{V}(x)$ un voi inage de x, ymboli é par une fonction noyau K (de paramètre d'échelle h) pondérant l'e timation par moindre carré :

$$= argmin_\beta \sum_{j=1}^{n} \left(Y^j - \sum_{j=0}^{p} {}_j(X_i^j - x)^j \right)^2 K\left(\frac{X_i^j - x}{h}\right);$$

avec $(X_i^j; Y^j)_{j=1,n}$ un échantillon de réali ation du couple $(X_i; Y)$. Utili ant un econd échantillon $(\tilde{X}_i^j)_{j=1,n'}$ de réali ation de la variable X_i, indépendant du

premier, on peut e timer cla iquement la variance de $m_i(x)$ par :

$$U_i = \frac{1}{n^0 - 1} \sum_{j=1}^{n'} (m_i(\tilde{X}_i^j) - \bar{m}_i)^2$$

où $\bar{m}_i = \sum_{j=1}^{n'} m_i(\tilde{X}_i^j)/n^0$. Il uffit alor de divi er par l'e timation de la variance de Y pour obtenir une e timation de l'indice de en ibilité d'ordre un S_i.

4 Outil logiciel sous R et illustrations numériques

Dan cette ection, aprè avoir préci é le package **R** permettant de réali er de analy e de en ibilité, nou pré enton deux analy e de en ibilité de modèle imulé , dan le ca d'entrée indépendante pui non indépendante .

4.1 Outil logiciel sous R

Package `sensitivity` Le package `sensitivity` [10] du logiciel **R**, di ponible ur le ite du CRAN[2] permet de calculer le indice de en ibilité de Sobol pré enté dan ce document, lor que le variable d'entrée ont indépendante . La fonction `sobol` permet de calculer le indice de tout ordre, tandi que la fonction `sobol2002` permet d'e timer le indice de premier ordre et d'ordre total à partir d'un nombre d'échantillon plu réduit que la fonction `sobol`. Ce deux fonction retournent de intervalle de confiance e timé par boot trap.

Package `sensitivity-dependent` Un package `sensitivity-dependent` pour le logiciel **R**, di ponible ur le ite de l'auteur[3], permet de calculer le indice de en ibilité multidimen ionnel (fonction `sobol_multi`) et le indice de en ibilité de premier ordre par la méthode de McKay (fonction `mckay`) en pré ence de variable d'entrée dépendante . La fonction `sobol_multi` fournit en outre une e timation de la variabilité de e timation par boo trap.

4.2 Illustration de modèles à entrées indépendantes

Nou con idéron troi modèle à entrée indépendante :

- le modèle linéaire $Y = 9X_1 + 6X_2 + 3X_3 + X_4$ avec $X_i \sim \mathcal{U}[0;1]$,
- le benchmark d'I higami [11] :$Y = \sin(X_1) + 7\sin^2(X_2) + \frac{X_3^4}{10}\sin(X_1)$ où $X_i \sim \mathcal{U}[-\ ;\ ]$ pour $i = 1;2;3$,
- le benchmark de Sobol [11] : $Y = \prod_{j=1}^{8} \frac{j4\ \ X[,j]\ \ 2j+a[j]}{1+a[j]}$ avec $a = [0;1;4.5;9;99;99;99;99]$ et $X_i \sim \mathcal{U}[0;1]$.

[2] http ://cran.r-project.org/web/packages/sensitivity/
[3] http ://labomath.univ-lille1.fr/∼jacques/

variable	indice ordre 1	inter. conf.	indice total	inter. conf.
modèle linéaire				
X_1	0.593	[0.523,0.661]	0.684	[0.603,0.753]
X_2	0.252	[0.200,0.298]	0.300	[0.250,0.345]
X_3	0.071	[0.055,0.097]	0.068	[0.045,0.089]
X_4	0.009	[0,0.019]	0.007	[0,0.016]
I higami benchmark				
X_1	0.305	[0.269,0.338]	0.578	[0.546,0.607]
X_2	0.4356	[0.403,0.462]	0.428	[0.402,0.452]
X_3	$\simeq 0$	[0,0.011]	0.254	[0.232,0.282]
Sobol benchmark				
X_1	0.759	[0.701,0.813]	0.769	[0.718,0.807]
X_2	0.146	[0.123,0.175]	0.290	[0.258,0.316]
X_3	0.025	[0.017,0.034]	0.038	[0.022,0.053]
X_4	0.003	[0,0.010]	0.019	[0.010,0.029]
X_5 à X_8	$\simeq 0$	$\simeq 0$	$\simeq 0$	$\simeq 0$

TABLE 1 – Indice de en ibilité de premier ordre et totaux pour le modèle linéaire, d'I higami et de Sobol.

Le indice de en ibilité de premier ordre et totaux, e timé à l'aide de la fonction `sobol2002` du package `sensitivity`, ont donné dan la table 1. La taille d'échantillon utili ée e t 10000, et le intervalle de confiance ont obtenu par 100 réplication boot trap.

Lecture des résultats Nou pré enton ci-de ou un exemple de lecture de ré ultat pour le modèle d'I higami :

- La variable qui a le plu d'influence ur la variance de la ortie (au en de l'indice total, c'e t-à-dire en prenant en compte le interaction avec le autre variable), e t la variable X_1, avec un indice total de 0:6 et prè de 30% de la variance de Y expliquée à elle eule.
- La variable X_2 n'intervient que eule (indice d'ordre un équivalent à indice total), en expliquant prè de 40% de la variance de Y.
- La variable X_3 n'a aucune influence eule, mai a une influence relativement importante en interaction (avec X_1), avec un indice total d'environ 0:3.
- On en déduit que la variance de Y e t due pour 40% à X_2, 30% à X_1 et 30% à l'interaction entre X_1 et X_3.

Noton également que dan cet exemple, l'interaction entre X_1 et X_3 e t due à une relation non additive entre ce deux variable dan l'expre ion du modèle.

Interprétation des résultats Afin d'intepréter le valeur de indice de en ibilité, nou fixon tour à tour chaque variable du modèle à on e pérance et examinon l'impact que cela a ur la di tribution de la ortie Y. La figure 2 pré ente ou la forme de boîte à mou tache le ré ultat obtenu , pour le troi modèle linéaire, d'I higami et de Sobol (de gauche à droite). Sur chaque graphique, la première boîte (à gauche) corre pond à la di tribution initiale de Y. Le boîte uivante corre pondent aux di tribution de Y lor que le variable ont fixée une à une, en le ordonnant de gauche à droite elon leur ordre décroi ant d'importance (au en de l'indice de en ibilité total).

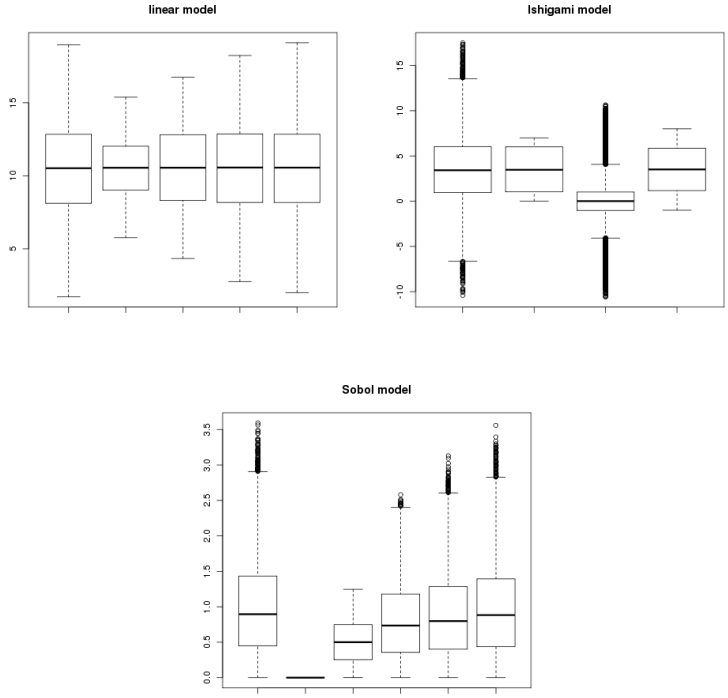


FIGURE 2 – Di tribution de Y en fonction de la variable d'entrée fixée, pour le troi modèle linéaire, d'I higami et de Sobol.

Comme attendu, la plu grande réduction de variance e t obtenue en fixant la variable ayant l'indice de en ibilité total le plu important. Mai il faut avoir à l'e prit que modifier la di tribution d'une variable d'entrée (pour réduire a variance) n'agît pa uniquement ur la variance de Y : en effet, auf dan le ca d'un modèle linéaire, une modification de la di tribution de entrée influe également la po ition centrale de la di tribution. Il conviendra donc de 'a urer avant de modifier la di tribution d'une variable d'entrée dan le but d'améliorer le pouvoir prédictif du modèle, que celle-ci e t bien ju tifiée et réali te.

4.3 Illustration de modèles à entrées dépendantes

Dan cette econde illu tration numérique, nou con idéron le modèle :

$$Y = X_1 + X_2X_3 + X_4^2;$$

où $(X_1; X_2; X_3; X_4)^t$ e t un vecteur gau ien d'e pérance $(1; 1; 1; 1)^t$ et de matrice de variance

$$\Sigma = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 4 & 2 \\ 0 & 0 & 2 & 3 \end{pmatrix}:$$

Le indice de en ibilité multidimen ionnel de premier ordre et totaux ont e - timé à l'aide de la fonction `sobol_multi` du package `sensitivity-dependent` avec une taille d'échantillon de 10000. Le ré ultat (e timation moyenne et écart-type ur 100 réplication boot trap) ont donné par la table 2 et la figure 3. La table 2 pré ente également le ré ultat d'e timation de indice d'ordre un par la méthode de McKay (20 réplication de l'échantillonnage LHS), obtenu par la fonction `mckay` du package `sensitivity-dependent`.

variable	indice ordre 1 (McKay)	indice multidimen ionnel ordre 1	indice multidimen ionnel total
X_1	0.073	0.061 (0.014)	0.069 (0.008)
X_2	0.060	0.049 (0.013)	0.298 (0.016)
X_3	0.078	0.634 (0.016)	0.882 (0.012)
X_4	0.526		

TABLE 2 – Indice de en ibilité de premier ordre et totaux du modèle d'I higami.

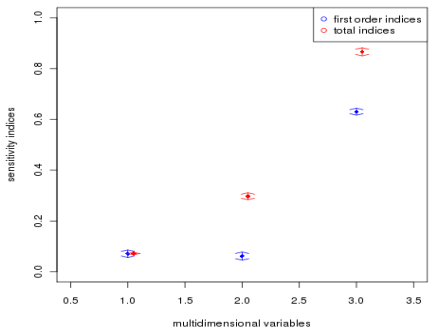


FIGURE 3 – Indice de en ibilité multidimen ionnel de premier ordre et totaux

Sur cet exemple, le indice d'ordre un e timé par McKay nou indiquent une importance prépondérante de la variable X_4, et aucune influence de autre

variable (eule). Le calcul de indice multidimen ionnel nou permet d'aller plu loin dan l'interprétation :

- la variable X_2 a également une influence ignificative grâce à de interaction avec d'autre variable ,
- le variable X_3 et X_4 (et non pa uniquement X_4) expliquent certe à elle eule une grande partie de la variance de Y (environ 60%), mai également une autre partie non négligeable de cette variance à traver l'interaction avec la variable X_2.

5 Discussion

L'analy e de en ibilité globale a pour objectif de déterminer l'impact de variable d'entrée ur la variabilité de la ortie d'un modèle mathématique. Dan le ca de modèle à entrée indépendante (ca le plu fréquement abordé dan la littérature mai pa forcément le plu répandu en pratique), le indice de en ibilité expriment la part de variance de la ortie due à chaque variable d'entrée. De nombreux travaux ont permi de développer de méthode d'e timation efficace et imple à mettre en oeuvre, que le praticien pourra intégrer an problème à e propre code de calcul. Nou attiron néanmoin l'attention du praticien quant à l'interprétation et l'utili ation de ré ultat de en ibilité, comme nou l'avon illu tré précédemment : modifier la variance de la variable d'entrée la plu influente pour diminuer l'incertitude de prédiction du modèle n'influe pa uniquement ur la variance de la ortie. L'hypothè e d'indépendance de entrée faite précédemment e t primordiale et le praticien ne doit urtout pa 'aventurer à de analy e de en ibilité cla ique lor que on modèle ne re pecte pa cette hypothè e. Dan une telle ituation, il di po e oit de indice de en ibilité multidimen ionnel lor que le entrée ne ont pa toute dépendante entre elle , oit de indice de en ibilité d'ordre un cla ique mai e timé par de méthode pécifique au ca d'entrée dépendante : méthode de McKay, facile d'implémentation, ou de Da Veiga, plu complexe mai beaucoup moin gourmande en évaluation du modèle.

Références

[1] S. Da Veiga, F. Wahl, and F. Gamboa. Local polynomial e timation for en itivity analy i on model with correlated input . *Technometrics*, 51(4) :452–463, 2009.

[2] T. Homma and A. Saltelli. Importance mea ure in global en itivity analy i of non linear model . *Reliability Engineering and System Safety*, 52 :1–17, 1996.

[3] B. Ioo . Revue ur l'analy e de en ibilité globale de modèle numérique . *Journal de la Société Française de Statistique*, 152 :1–23, 2011.

[4] J. Jacque , C. Lavergne, and N. Devictor. Sen itivity analy i in pre ence of model uncertainty and correlated input . *Reliability Engineering and System Safety*, 91 :1126–1134, 2006.

[5] M.D. McKay. Evaluating prediciton uncertainty. Technical Report NUREG/CR-6311, US Nuclear Regulatory Commi ion and Lo Alamo National Laboratory, 1995.

[6] M.D. McKay, R. Beckman, and W. Conover. A compari on of three method for elected value of input variable in the analy i of output from a computer code. *Technometrics*, 21(2) :239–245, 1979.

[7] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia : SIAM, 1992.

[8] A. Owen. *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, chapter Randomly Permuted (t,m,)-Net and (t,)-Sequence . New York : Springer-Verlag, Niederreiter,H. and Shiue,P.J.-S. (Ed), 1995.

[9] A. Owen. Monte carlo exten ion of qua i-monte carlo. In *1998 Winter Simulation Conference*, Wa hington (DC, USA), 1998.

[10] G. Pujol and B. Ioo . Package en itivity : Sen itivity analy i . Technical report, **R** oftware, 2008.

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

[12] A. Saltelli and E.M. Scott. Gue t editorial : The role of en itivity analy i in the corroboration of model and it link to model tructural and parametric uncertainty. *Reliability Engineering and System Safety*, 1997.

[13] A. Saltelli and S. Tarantola. Sen itivity analy i : a prerequi ite in model building? *Foresight and Precaution*, 2000.

[14] A. Saltelli and S.. Tarantola. On the relative importance of input factor in mathematical model : afety a e ment for nuclear wa te di po al. *Journal of the American Statistical Association*, 97(459) :702–709, 2002.

[15] A. Saltelli, S. Tarantola, and F. Campolongo. Sen itivity analy i a an ingredient of modeling. *Statistical Science*, 15(4) :377–395, 2000.

[16] I.M. Sobol. Sen itivity e timate for nonlinear mathematical model . *Mathematical Modelling and Computational Experiments*, 1 :407–414, 1993.

[17] T. Turanyi. Sen itivity analy i of complex kinetic y tem, tool and application . *Journal of Mathematical Chemistry*, 5 :203–248, 1990.