Indices de Sobol

Résumé de prises de notes du livre *Analyse de sensibilite et exploration de modèles : application aux sciences de la nature et de l'environnement* [1] et d'articles, notamment *Sobol 2001*[2].

1 Cadre de l'analyse de sensibilité

On note $x = (z, \theta) = (x_1, ..., x_K)$ l'ensemble des variables d'entrées z et des paramètres du modèle θ sur lesquels on souhaite obtenir un indice de sensibilité. On parlera par la suite de variables ou de paramètres pour les composantes de x indépendamment du fait qu'elles soient dans z ou dans θ. On note le modèle $\mathcal{G}$ et y sa sortie, de sorte que $\mathcal{G}(x_1, ..., x_K) = y$. On ne dissociera pas ici $\mathcal{G}$ de la fonction de code *FC* qui est sa version physique réelle avec des erreurs d'approximation liées aux contraintes informatiques.

Là où l'analyse d'incertitude s'attache à donner une incertitude sur $\mathcal{G}(x)$ en prenant en considération l'incertitude sur x, l'analyse de sensibilité revient à déterminer lesquels des paramètres parmi $(x_1, ..., x_K)$ ont une forte influence sur les sorties du modèle. L'article de Sobol de 2001 *Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates*[2] note les objectifs possibles de l'analyse de senbilité :

- Classer les variables parmi $(x_1, ..., x_K)$ par ordre d'influence sur la sortie du modèle
- Déterminer les variables parmi $(x_1, ..., x_K)$ très peu influentes qu'on peut éventuellement enlever du modèle pour le simplifier tout en gardant sa légitimité.
- Supprimer les interactions dans $\mathcal{G}$ mettant en jeu trop de variables inutilement (pour peu d'influence sur la sortie au final). Cela rentre dans l'idée de simplifier le modèle au maximum.

Pour qualifier l'influence d'une variable X_i sur la sortie du modèle, on utilise les indices de Sobol, respectivement :

$$SI_i = \frac{\mathbb{V}\left(\mathbb{E}(Y|X_i)\right)}{\mathbb{V}(y)} = 1 - \frac{\mathbb{E}\left(\mathbb{V}(Y|X_i)\right)}{\mathbb{V}(Y)}$$

l'indice de sensibilité principal qui mesure l'effet de la variable X_i seule, c'est-à-dire seulement à travers son effet sans interaction avec les autres variables $(X_j)_{j\neq i}$, sur la sortie du modèle $\mathcal{G}(Y)$. Cela revient à mesurer la variabilité des simulations lorsque ces simulations sont moyennées sur $(X_j)_{j\neq i}$ (qu'on notera X_{-i}), et

$$TSI_i = \frac{\mathbb{E}\left(\mathbb{V}(Y|X_{-i})\right)}{\mathbb{V}(Y)} = 1 - \frac{\mathbb{E}\left(\mathbb{V}(Y|X_{-i})\right)}{\mathbb{V}(Y)}$$

l'indice de sensibilité total qui mesure l'effet de la variable X_i seule et en interaction avec les autres variables X_{-i}.

Pour obtenir ces indices, il faut échantillonner l'espace des paramètres qu'on appelle parfois *plan d'expérience*. Cette échantillonage peut prendre plusieurs formes.

Le *Latin Hypercube Sampling* consiste à répartir les points d'échantillons assez uniformément dans l'espace des paramètres en découpant chaque plage des paramètres $(X_i)_{i\in\{1,...,n\}}$ en un même nombre N de segments et à tirer un point dans chaque segment, puis ensuite à combiner aléatoirement les tirages de $(x_1^N, ..., x_K^N)$ de sorte à avoir une matrice de taille $N \times K$. Pour N suffisamment grand, la méthode de Monte-Carlo garantit la convergence des estimateurs des indices de sensibilité. Le problème de cette méthode est qu'elle ne contrôle absolument pas la répartition conjointe des échantillons.

Pour bien recouvrir un espace K-dimensionnel, des méthodes alternatives ont été développées dès les années 1950, dans un but premier d'analyse numérique de calcul d'intégrales. Les travaux de Sobol sur les méthodes de quasi-Monte-Carlo font partie de ces travaux. Le nom de Sobol a plus tard était associé à l'analyse de sensibilité car ces méthodes d'estimation ont été prouvées utiles dans le cadre du calcul des indices de sensibilité. *L'idée des méthodes de quasi-Monte-Carlo est de substituer aux séquences générées aléatoirement dans la méthode de Monte-Carlo des suites déterministes ayant de meilleures propriétés de convergence dans certains cas.*[1] Une fonction mathématique nommé discrépance permet en quelque sorte de quantifier la non répartition des points à la façon de la loi uniforme. Pour optimiser l'uniformité de remplissage de l'échantillon sur un espace, le critère de discrépance doit être minimal. Les suites de Sobol sont des suites à discrépance faible qui peuvent éventuellement être utilisées à la place de Monte-Carlo pour calculer des indices de sensibilité.

Il existe aussi des critères géométriques tels que le critère *maximin*, basé sur des distances, pour juger de la qualité d'un plan d'expérience. Cette méthode était toutefois jugée très coûteuse au moment de la parution du livre[1].

2 Cadre conceptuel des indices de Sobol

Soient $(X_1, ..., X_K)$ les variables du modèle sélectionnées pour l'analyse de sensibilité et $(D_1, ..., D_K)$ leurs domaines respectifs. Si on ne sait pas donner de domaine à une variable a priori, on peut éventuellement prendre à partir d'un échantillon empirique les quantiles à 0.05 et 0.95.

Chacune des lois X_i a une loi de probabilité π_i. On peut ramener les $(x_1, ..., x_K)$ à un domaine de valeurs dans $I^K = [0, 1]^K$ en prenant $p_i = F_i(x_i) = \mathbb{P}(X_i \leq x_i)$ grâce à F_i la fonction de répartition de X_i qui assure la bijection entre D_i et $[0, 1]$[1].

Sauf indication explicite, les entrées sont supposées indépendantes les unes des autres. [1] La loi de probabilité de $X = (X_1, ..., X_K)$ est donc $\pi = \pi_1 \times ... \times \pi_K$. [1] [ça m'interroge]

L'article de Sobol[2] présente la décomposition dite ANOVA, d'une fonction intégrable f définie sur un hypercube I^K. Cette fonction peut se décomposer sous la forme suivante, qui fait intervenir 2^K termes correspondant aux 2^K sous-ensembles possibles des variables d'entrées :

$$f(x) = f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + ... + f_{1,2,...,K}(x_1, ..., x_K) \tag{1}$$

Cette décomposition est dite ANOVA de $f(x)$ si :

$$\int_0^1 f_{i_1...i_s}(x_{i_1}, ..., x_{i_s})dx_k = 0$$

pour tout $k = i_1, ..., i_s$. Il suit de cette contrainte que les membres de la définition 1 sont orthogonaux et peuvent être exprimés comme intégrales de $f(x)$:

$$\int f(x)dx = f_0 \tag{2}$$

$$\int f(x) \prod_{k \neq i} dx_k = f_0 + f_i(x_i) \tag{3}$$

$$\int f(x) \prod_{k \neq i,j} dx_k = f_0 + f_i(x_i) + f_{ij}(x_i, x_j) \tag{4}$$

et caetera. La dénomination ANOVA vient de *Analysis of Variances* car sous l'hypothèse que f est de carré intégrable, si la loi de X, que nous avions nommée π, est uniformément distribuée sur I^K, alors on peut interpréter les quantités

$$D_{i_1...i_s} = \int_0^1 f^2_{i_1...i_s}(x_{i_1}, ..., x_{i_s})dx_{i_1}...dx_{i_s} \tag{5}$$

comme les variances respectives de $f_{i_1...i_s}(X_{i_1}, ..., X_{i_s})$.

$$D = \left(\int f^2(x)dx \right) - f_0^2 = \sum_{s=1}^{K} \sum_{i_1<...<i_s}^{K} \int f_{i_1...i_s}^2(x_{i_1}, ..., x_{i_s})dx_{i_1}...dx_{i_s} = \sum_{s=1}^{K} \sum_{i_1<...<i_s}^{K} D_{i_1...i_s} \quad (6)$$

traduit la variance globale de $f(X)$. Les ratios

$$S_{i_1...i_s} = \frac{D_{i_1...i_s}}{D} \quad (7)$$

sont appelés indices de sensibilité globaux. Ils traduisent la fraction de la variance totale de $f(X)$ influencée par les interactions mettant en jeu tout $(X_{i_1}, ..., X_{i_s})$. $\sum_{s=1}^{K} \sum_{i_1<...<i_s}^{K} S_{i_1...i_s} = 1$. $\sum_{i=1}^{K} S_i = 1$ singifie que le modèle est purement la somme de K fonctions unidimensionnelles sans interactions entre elles. Si f est régulière par morceaux, $S_{i_1...i_s} = 0$ signifie que $f_{i_1...i_s}(x_{i_1}, ..., x_{i_s})$ est nulle partout. L'indice $S_{i_1...i_s}$ diffère de $S_{(i_1...i_s)}$ indice de sensitivité global pour tous le sous-groupe $(i_1...i_s)$ introduit plus bas, qui contient la mesure de la variance de toutes les interactions parmi $(X_{i_1}, ..., X_{i_s})$ d'ordres inférieurs ou égaux à s.

Les termes

$$S_i = \frac{D_i}{D} = \frac{\sum_{i=1}^{K} \int f_i^2(x_i)dx_i}{D} \quad (8)$$

sont les indices de sensibilité globaux des variables X_i.

Pour calculer l'indice de sensibilité global d'un sous-groupe de variables $x_U = (x_{k_1}, ..., x_{k_m})$ avec $1 \leq m \leq K-1$ et $1 \leq k_1 \leq ... \leq k_m \leq K$, on considère :

$$D_U = \sum_{s=1}^{m} \left(\sum_{(i_1<...<i_s)\in\{(k_1,...,k_m)\}} D_{i_1...i_s} \right) \quad (9)$$

auquel on associe l'indice de sensibilité global du sous-groupe :

$$S_U = \frac{D_U}{D} \quad (10)$$

Concrètement, cela signifie que si on s'intéresse à $x_U = (x_1, x_3, x_5)$ parmi 6 variables, on aura :

$$D_U = \left(\sum_{(i_1)\in\{(1,3,5)\}} D_{i_1} \right) + \left(\sum_{(i_1<i_2)\in\{(1,3,5)\}} D_{i_1,i_2} \right) + \left(\sum_{(i_1<i_2<i_3)\in\{(1,3,5)\}} D_{i_1,i_2,i_3} \right) \quad (11)$$

$$= (D_1 + D_3 + D_5) + (D_{1,3} + D_{1,5}) + (D_{1,3,5}) \quad (12)$$

De plus, $x = (x_U, x_{-U})$ et on peut donc définir $D_U^{tot} = D - D_{-U}$ et de même :

$$S_{(U)}^{tot} = \frac{D_U^{tot}}{D} = 1 - \frac{D_{-U}}{D} \quad (13)$$

D_U^{tot} correspond à la somme de $D_{i_1...i_s}$ mais étendue à tous les groupes $(i_1, ..., i_s)$ où au moins un i_l appartient à U. Pour $x_U = (x_1, x_2)$ parmi 3 variables, on aura donc :

$$S_{(U)} = S_1 + S_2 + S_{1,2} \quad (14)$$

$$S_{(U)}^{tot} = S_1 + S_2 + S_{1,2} + S_{1,3} + S_{2,3} + S_{1,2,3} = 1 - S_{(3)} = 1 - S_3 \quad (15)$$

3 Calcul des indices de Sobol

The main breakthrough in Sobol (1990) is the computation algorithm that allows a direct estimation of global sensitivity indices using values of f(x) only. And this is a Monte Carlo algorithm.[1]

Pour calculer les $S_{(U)}$ et $S^{tot}_{(U)}$, il suffit d'estimer[2] les intégrales :

$$\int f(x)dx \tag{16}$$

$$\int f^2(x)dx \tag{17}$$

$$\int f(x)f(x_U, x'_{-U})dxdx'_{-U} \tag{18}$$

$$\int f(x)f(x'_U, x_{-U})dxdx'_U \tag{19}$$

Considérons deux échantillons indépendants ξ et ξ' uniformément distribués sur I^K, qu'on divise en $\xi = (\eta, \zeta)$ et $\xi' = (\eta', \zeta')$. Chaque itération de Monte-Carlo nécessite trois calculs du modèle : $f(\xi) = f(\eta, \zeta)$, $f(\eta, \zeta')$ et $f(\eta', \zeta)$. On obtient les estimateurs respectifs des intégrales précédentes, et donc de $S_{(U)}$ et $S^{tot}_{(U)}$.

On peut générer aléatoirement les ξ et ξ' qu'on stocke dans deux matrices, et avec quelques permutations de colonnes entre ξ et ξ' on peut ainsi caculer par Monte-Carlo les indices souhaités. Les permutations de colonnes entre matrices A et B (analogues de ξ et ξ') présentées dans le livre de Faivre [1] ou dans le notice du package *sensobol*[3] semblent être l'analogue de cette partie de la méthode.

Sobol 2001 propose également une autre méthode de Monte-Carlo pour l'estimation des indices de sensitivité globaux, basés sur une variante des intégrales à estimer. Faivre[1] explicite deux estimateurs (semble-t-il Sobol 1993 pour l'ordre 1 et Jansen 1999 pour l'ordre total, peut-être) dans le cadre de la méthode de Sobol. Il existe aussi des méthodes d'estimation des indices de sensitivité par la méthode *FAST* (*Fourier Amplitude Sensitivity Test*) et ses variantes.

4 Package *sensobol*

Le package *sensobol*, qui se veut oecuménique, offre 4 estimateurs des indices de sensitivité de premiers ordre et 7 estimateurs des indices de sensitivité d'ordre totaux.

Le *sampling design*, c'est-à-dire le tirage des points du plan d'expérience afin d'obtenir une approximation des indices de Sobol, peut se faire avec la méthode *Quasi-Random Numbers*, basée sur les travaux de Sobol, et qui propose donc du quasi-Monte-Carlo, sur la méthode présentée plus haut dite *Latin Hypercube Sampling*, ou sur du tirage aléatoire (selon une loi uniforme j'imagine).

5 Précision des estimations des indices de Sobol

L'évaluation de la précision des estimateurs peut-être effectuée par bootstrap[1].

6 Que faire de ce paragraphe ?

Les méthodes de criblage consistant à simplement simuler le modèle selon un criblage de l'espace des paramètres peuvent se révéler insuffisantes. En effet, si on a K paramètres pour lesquels on prend seulement 2 valeurs chacune (on appelle de nombre de valeurs testées *niveaux*), on a 2^K évaluations de $\mathcal{G}(x_1, ...x_K)$ à réaliser, et ce sans supposer qu'il y a des interactions entre les entrées (qui nécessiterait par exemple un criblage plus fin pour être décelées) et que la varation de chaque $x_i \mapsto \mathcal{G}(x_1, ...x_K)$ est monotone.

Références

[1] Mahévas Makowski Monod Faivre Iooss. *Analyse de sensibilité et exploration de modèles : application aux sciences de la nature et de l'environnement*. fre. Collection Savoir-faire. Versailles : Éd. Quae, 2013. ISBN : 978-2-7592-1906-3.

[2] I.M Sobol . "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates". en. In : *Mathematics and Computers in Simulation* 55.1-3 (fév. 2001), p. 271-280. ISSN : 03784754. DOI : 10.1016/S0378-4754(00)00270-6. URL : https://linkinghub.elsevier.com/retrieve/pii/S0378475400002706 (visité le 19/02/2024).

[3] Arnald Puy et al. "**sensobol** : An *R* Package to Compute Variance-Based Sensitivity Indices". en. In : *Journal of Statistical Software* 102.5 (2022). ISSN : 1548-7660. DOI : 10.18637/jss.v102.i05. URL : https://www.jstatsoft.org/v102/i05/ (visité le 19/02/2024).