

Méthodes d'Analyse de Sensibilité Globale

Robert Faivre

INRA - Mathématiques et Informatique Appliquées - Université Fédérale de Toulouse



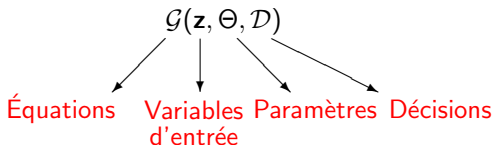
FormaSciences



La Rochelle, 26 mars 2018

Définition et Objectifs

► Sources d'incertitude dans un modèle



► Types d'incertitude

- Manque de connaissance
ex: équation identifiée empiriquement sur un jeu de données
- Variabilité de l'environnement du système
ex: " température moyenne journalière " variable entre années
- Imprécision
ex: paramètre de croissance estimé sur un jeu de données
- Règle de décision mal maîtrisée
ex: seuil de déclenchement de l'irrigation

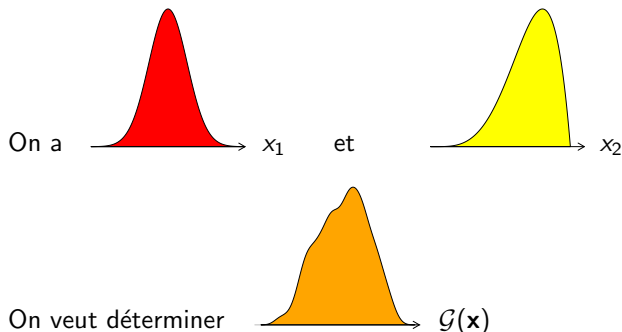
Notations

$$\begin{aligned}\mathbf{y} &= \mathcal{G}(\mathbf{x}) \\ &= \mathcal{G}(x_1, \dots, x_K)\end{aligned}$$

- ▶ $x_1, \dots, x_K$ = variables d'entrée et paramètres incertains
- ▶ $x_1, \dots, x_K$ les K facteurs incertains
- ▶ $\mathbf{x} = (x_1, \dots, x_K) \in \mathbb{R}^K$
- ▶ $\mathbf{y}$ peut être multidimensionnel

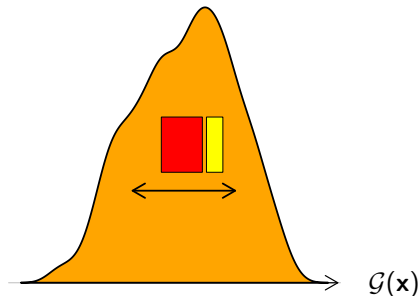
Objectif

- ▶ Répondre à la question suivante
 - ▶ Quel est le niveau d'incertitude de $\mathcal{G}(\mathbf{x})$ qui résulte de l'incertitude dans les x_k ?



Objectif

- ▶ Répondre à la question
 - ▶ Quelles sont les principales sources d'incertitude parmi $x_1, \dots, x_K$ qui influencent $\mathcal{G}(\mathbf{x})$?



Variance de $\mathcal{G}(\mathbf{x}) = \text{Effet de } x_1 + \text{Effet de } x_2 + \dots$

Analyse d'incertitude

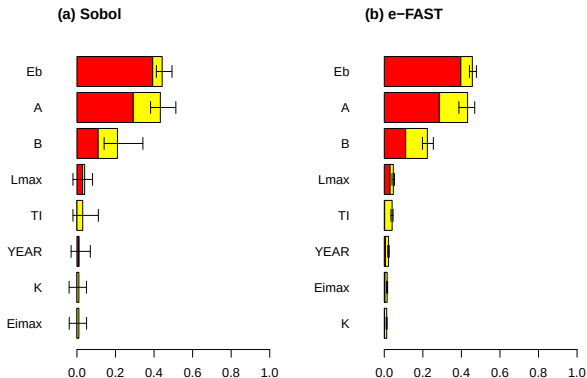
- ▶ donner des informations sur l'incertitude associée aux prédictions d'un modèle
- ▶ optimiser des variables décisionnelles pour réduire l'incertitude

Analyse de sensibilité

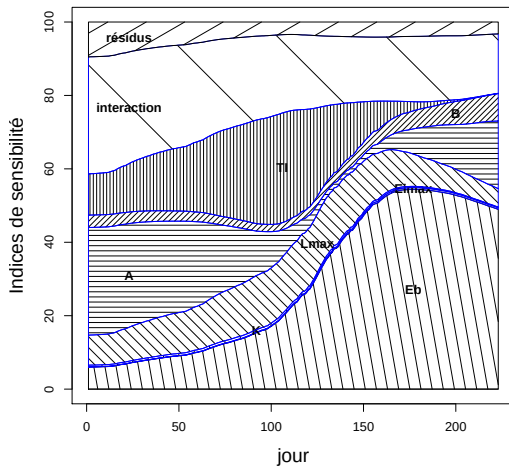
- ▶ identifier les paramètres et les variables d'entrée qui ont une forte influence sur les sorties d'un modèle
→ Important de les connaître avec précision
- ▶ identifier les paramètres et les variables d'entrée qui ont une influence moindre sur les sorties
→ Moins important de les connaître avec précision
- ▶ analyser le comportement du modèle sur toute une gamme de variation des x_k avec le minimum d'hypothèses sur le modèle, en intégrant les interactions

Analyse de sensibilité du rendement simulé

Indices de sensibilité principaux et totaux



Analyse de sensibilité de l'évolution de la biomasse simulée



1. Analyse d'incertitude

Un exemple très simple

- ▶ Équation : $\mathcal{G}(x_1, x_2) = x_1 + 2x_2$
- ▶ Incertitude sur x_1 et x_2 : $x_1 \sim \mathcal{N}(20, 16)$ et $x_2 \sim \mathcal{N}(60, 64)$

Si X_1 et X_2 sont deux variables indépendantes de distribution gaussienne alors $\alpha X_1 + \beta X_2$ suit une distribution gaussienne

$$\begin{aligned} E(\alpha X_1 + \beta X_2) &= \alpha E(X_1) + \beta E(X_2) \\ \text{Var}(\alpha X_1 + \beta X_2) &= \alpha^2 \text{Var}(X_1) + \beta^2 \text{Var}(X_2) \end{aligned}$$

$$\mathcal{G}(x_1, x_2) \sim \mathcal{N}(140, 272)$$

En général, c'est plus dur !

- ▶ Equations plus complexes, relation non linéaire entre $\mathcal{G}(\mathbf{x})$ et $\mathbf{x}$
→ Impossible de déterminer l'expression analytique de la distribution de $\mathcal{G}(\mathbf{x})$
- ▶ La distribution de $\mathbf{x}$ n'est pas toujours connue
→ Choix subjectif
- ▶ Temps de calcul parfois long avec certains modèles
→ Le nombre de simulations peut être limitant

Procédure en quatre étapes

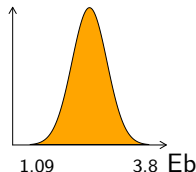
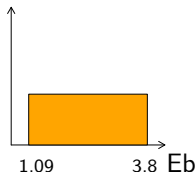
1. Définir les distributions de $x_1, \dots, x_K$
2. Générer des échantillons à partir des distributions définies à l'étape 1
3. Calculer $\mathcal{G}(\mathbf{x})$ pour chaque série de $x_1, \dots, x_K$ générée
4. Estimer la distribution de $\mathcal{G}(\mathbf{x})$

Étape 1 - Définition des distributions

Les distributions de probabilité des facteurs incertains (paramètres ou variables d'entrée) peuvent être définies en utilisant :

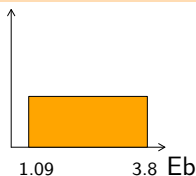
- ▶ La littérature scientifique et l'expertise
- ▶ Des séries de mesures (série climatique ...)
- ▶ Les valeurs des paramètres estimées

Exemple : d'après un article publié par Jeuffroy et Recous en 1999 dans EJA, l'efficacité d'utilisation de rayonnement intercepté varie entre 1.09 et 3.8 g.MJ-1 pour le blé

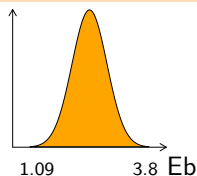


Étape 2 - Génération d'échantillons à partir des distributions de $x_1, x_2, \dots, x_K$

- ▶ Il faut générer suffisamment de valeurs de $x_1, x_2, \dots, x_K$
- ▶ Différentes méthodes d'échantillonnage peuvent être utilisées : échantillonnage aléatoire (Monte Carlo), échantillonnage en hypercube latin, ...
- ▶ En pratique, on utilise un logiciel pour générer N valeurs de $x_1, x_2, \dots, x_K$ (ex: $N = 20000$).



1.26 3.36 1.90 1.18 1.47
2.05 2.44 1.30 2.86 2.48



1.83 2.23 2.01 2.44 2.60
2.66 1.71 2.71 2.20 2.80

Étape 3 - Calcul de $\mathcal{G}(\mathbf{x})$ pour chaque série $x_1, x_2, \dots, x_K$ générée

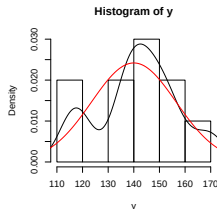
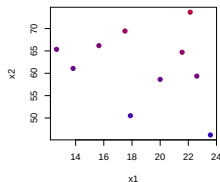
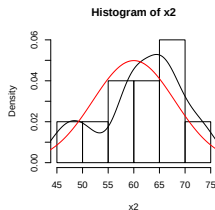
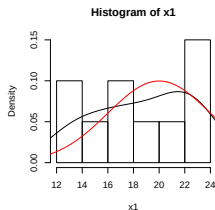
- ▶ Exécuter le modèle N fois, pour toutes les valeurs générées des facteurs d'entrée $x_1, x_2, \dots, x_K$

Étape 4 - Approximation de la distribution de $\mathcal{G}(\mathbf{x})$

- ▶ Décrire les N valeurs de $\mathcal{G}(\mathbf{x})$ calculées à l'étape 3
- ▶ Différentes approches possibles
 - ▶ calcul de la moyenne et de la variance,
 - ▶ calcul de quantiles (quartiles, déciles, ...),
 - ▶ histogramme,
 - ▶ fonction de distribution cumulée,
 - ▶ box plot...

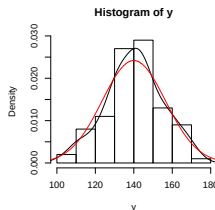
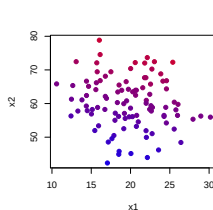
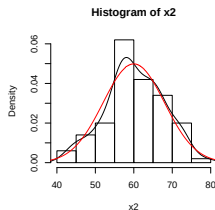
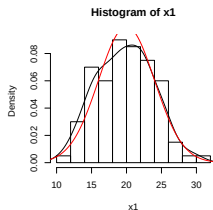
Application au modèle simple

► $N = 10$



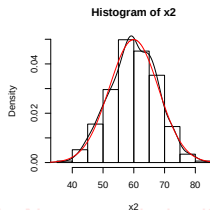
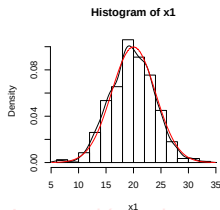
Application au modèle simple

► $N = 100$

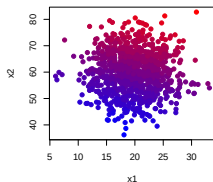
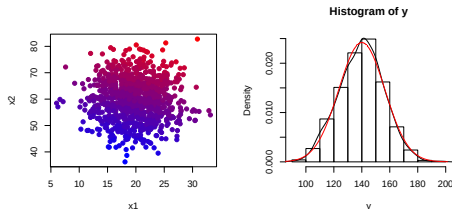


Application au modèle simple

► $N = 1000$

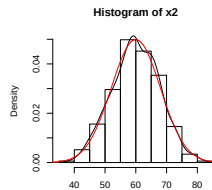
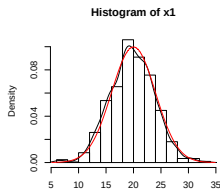


La précision ne dépend que de N et non de la dimension K

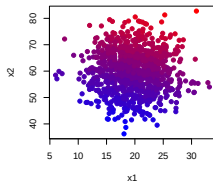
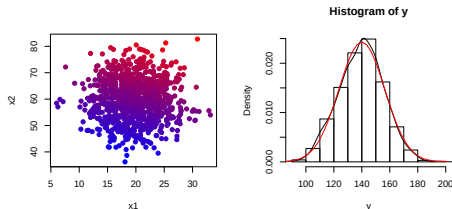


Application au modèle simple

► $N = 1000$



La précision ne dépend que de N et non de la dimension K



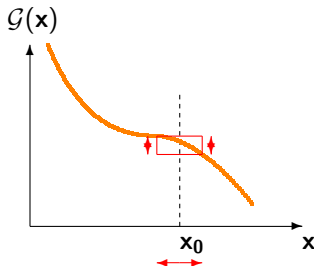
2. Analyse de sensibilité

Question

- Quelles sont les principales sources de variation qui influencent $\mathcal{G}(\mathbf{x})$?

Analyse de sensibilité locale

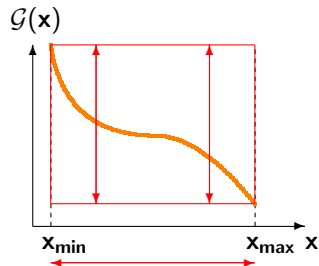
Variation de $\mathcal{G}(\mathbf{x})$ autour de $\mathbf{x}_0$



basée sur le calcul de dérivées

Analyse de sensibilité globale

Variation globale de $\mathcal{G}(\mathbf{x})$ quand $\mathbf{x}$ varie dans son domaine d'incertitude



basée sur le calcul d'indices

Elle consiste à

- ▶ Définir des indices de sensibilité
- ▶ Calculer ces indices en faisant varier les facteurs incertains $x_1, \dots, x_K$ sur leurs domaines

Procédure en quatre étapes

1. Définir les distributions de $x_1, \dots, x_K$
2. Générer des échantillons à partir des distributions définies à l'étape 1
3. Calculer $\mathcal{G}(\mathbf{x})$ pour chaque série de $x_1, \dots, x_K$ générée
4. Calculer les indices de sensibilité

Discrétisation de l'espace des paramètres

- ▶ Plan factoriel - Analyse de variance
- ▶ Méthode de Morris

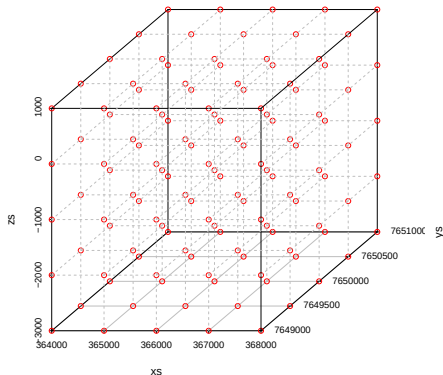
Espace des paramètres continu

1. Indices basés sur la régression
2. Indices de Sobol' estimés par échantillonnage intensif
3. Méthodes Pick and Freeze et extended FAST

On suppose les paramètres $x_1, \dots, x_K$ indépendants

On utilise des représentants des X - plan factoriel

- 5 niveaux par facteurs sur le modèle volcan



Rappels d'ANOVA à effets fixes (sur plan factoriel complet)

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijr}$$

- ▶ X_1 (resp. X_2) facteur à I (resp. J) niveaux ; $\sum_{i=1}^I \alpha_i = \sum_{j=1}^J \beta_j = 0$
- ▶ $\varepsilon_{ijr}, r = 1, \dots, R$ iid $\sim \mathcal{N}(0, \sigma^2)$

Estimateurs

- ▶ $\hat{\mu} = Y_{...}$
- ▶ $\hat{\alpha}_i = Y_{i..} - Y_{...}$
- ▶ $\hat{\beta}_j = Y_{.j.} - Y_{...}$
- ▶ $\hat{\gamma}_{ij} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$

$$Y_{...} = \frac{1}{I \times J \times R} \sum_{i,j,r} Y_{ijr}$$

$$Y_{i..} = \frac{1}{J \times R} \sum_{j,r} Y_{ijr}$$

$$Y_{.j.} = \frac{1}{I \times R} \sum_{i,r} Y_{ijr}$$

$$Y_{ij.} = \frac{1}{R} \sum_r Y_{ijr}$$

Indices basés sur la décomposition de la variance

Décomposition de la variance (Somme des Carrés des Écarts)

$$SCT = SCM + SCR$$

$$\text{variance totale} = \text{variance expl. par le modèle} + \text{variance résiduelle}$$

$$SCT = SCE_{\alpha} + SCE_{\beta} + SCE_{\gamma} + SCR_{\varepsilon}$$

$$SCT = \sum_{ijr} (Y_{ijr} - Y_{...})^2$$

$$\text{Coefficient de détermination } R^2 = \frac{SCM}{SCT}$$

Contributions des facteurs = indices de sensibilité

- ▶ Contribution principale du facteur X_1 : $\frac{SCE_{\alpha}}{SCT}$ ($\frac{SCE_{\beta}}{SCT}$ pour X_2)
- ▶ Contribution totale du facteur X_1 : $\frac{SCE_{\alpha} + SCE_{\gamma}}{SCT}$ ($\frac{SCE_{\beta} + SCE_{\gamma}}{SCT}$ pour X_2)

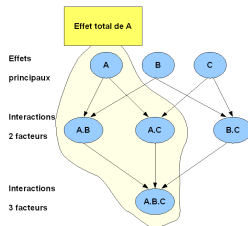
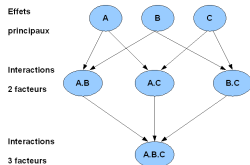
Statistiques de test associées

- ▶ Modèle additif vs Modèle complet :

$$T = \frac{SCE_{\gamma} / (IJ - I - J + 1)}{SCE_{\varepsilon} / (IJR - IJ)} \sim F_{H0}(IJ - I - J + 1, IJR - IJ)$$

Structuration des facteurs

- indices par facteur



Structuration des facteurs

- indices par groupe de facteurs

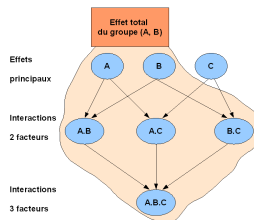
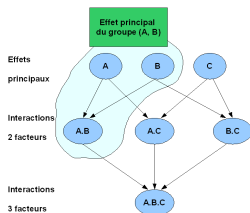


Tableau d'analyse de variance

Modèle volcan

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
xs	4	0.8823	0.2206	1.892	0.1090
a	4	2410.6622	602.6655	5169.770	0
p	4	1281.6185	320.4046	2748.4867	0
xs:a	16	0.3451	0.0216	0.1850	0.9998
xs:p	16	0.4868	0.0304	0.2610	0.9986
a:p	16	631.6311	39.4769	338.6401	0
xs:a:p	64	0.70335	0.0110	0.09423	1
Residuals	3000	349.72474	0.1166		

Pourcentages de variance expliquée

xs	0.02	xs:a	0.01	xs:a:p	0.02
a	51.55	xs:p	0.01		
p	27.41	a:p	13.51	Residuals	7.48

Indices totaux

xs	0.05
a	65.09
p	40.95

Modèle déterministe

- ▶ décomposition totale de la variance (pas d'erreur résiduelle)
- ▶ plan factoriel complet très grand
- ▶ plans fractionnaires utiles (voir PECNUM)

Méthode par plan factoriel complet

- ▶ chaque facteur est décomposé en un certain nombre de niveaux
 - ▶ plan factoriel complet : croisement de tous les niveaux
-
- ▶ Prenons un modèle avec 20 paramètres (facteurs) incertains.
 - ▶ Pour un plan factoriel complet de 20 facteurs à 3 niveaux, cela nécessite $3^{20} = 3\,486\,784\,401$ évaluations du modèle.
 - ▶ Si une évaluation prend environ 0.01 seconde, cela nécessite 581 130.7 heures de calcul.
 - ▶ Avec une ferme de calcul de 1000 processeurs, cela représente environ 581 heures, soit 24.2 jours.
 - ▶ Pour une puissance de 200 Watts par unité centrale, la consommation électrique correspond à 116 226.1 kWh, soit la consommation annuelle de 16 familles françaises environ et un coût de 11.6 K€ (10 c. / kWh).

Une planification raisonnée, un enjeu pour le Développement Durable

Méthode par plan factoriel complet

- ▶ chaque facteur est décomposé en un certain nombre de niveaux
 - ▶ plan factoriel complet : croisement de tous les niveaux
-
- ▶ Prenons un modèle avec 20 paramètres (facteurs) incertains.
 - ▶ Pour un plan factoriel complet de 20 facteurs à 3 niveaux, cela nécessite $3^{20} = 3\,486\,784\,401$ évaluations du modèle.
 - ▶ Si une évaluation prend environ 0.01 seconde, cela nécessite 581 130.7 heures de calcul.
 - ▶ Avec une ferme de calcul de 1000 processeurs, cela représente environ 581 heures, soit 24.2 jours.
 - ▶ Pour une puissance de 200 Watts par unité centrale, la consommation électrique correspond à 116 226.1 kWh, soit la consommation annuelle de 16 familles françaises environ et un coût de 11.6 K€ (10 c. / kWh).

Une planification raisonnée, un enjeu pour le Développement Durable

Méthode par plan factoriel complet

- ▶ chaque facteur est décomposé en un certain nombre de niveaux
 - ▶ plan factoriel complet : croisement de tous les niveaux
-
- ▶ Prenons un modèle avec 20 paramètres (facteurs) incertains.
 - ▶ Pour un plan factoriel complet de 20 facteurs à 3 niveaux, cela nécessite $3^{20} = 3\,486\,784\,401$ évaluations du modèle.
 - ▶ Si une évaluation prend environ 0.01 seconde, cela nécessite 581 130.7 heures de calcul.
 - ▶ Avec une ferme de calcul de 1000 processeurs, cela représente environ 581 heures, soit 24.2 jours.
 - ▶ Pour une puissance de 200 Watts par unité centrale, la consommation électrique correspond à 116 226.1 kWh, soit la consommation annuelle de 16 familles françaises environ et un coût de 11.6 K€ (10 c. / kWh).

Une planification raisonnée, un enjeu pour le Développement Durable

Méthode par plan factoriel complet

- ▶ chaque facteur est décomposé en un certain nombre de niveaux
 - ▶ plan factoriel complet : croisement de tous les niveaux
-
- ▶ Prenons un modèle avec 20 paramètres (facteurs) incertains.
 - ▶ Pour un plan factoriel complet de 20 facteurs à 3 niveaux, cela nécessite $3^{20} = 3\,486\,784\,401$ évaluations du modèle.
 - ▶ Si une évaluation prend environ 0.01 seconde, cela nécessite 581 130.7 heures de calcul.
 - ▶ Avec une ferme de calcul de 1000 processeurs, cela représente environ 581 heures, soit 24.2 jours.
 - ▶ Pour une puissance de 200 Watts par unité centrale, la consommation électrique correspond à 116 226.1 kWh, soit la consommation annuelle de 16 familles françaises environ et un coût de 11.6 K€ (10 c. / kWh).

Une planification raisonnée, un enjeu pour le Développement Durable

Méthode par plan factoriel complet

- ▶ chaque facteur est décomposé en un certain nombre de niveaux
 - ▶ plan factoriel complet : croisement de tous les niveaux
-
- ▶ Prenons un modèle avec 20 paramètres (facteurs) incertains.
 - ▶ Pour un plan factoriel complet de 20 facteurs à 3 niveaux, cela nécessite $3^{20} = 3\,486\,784\,401$ évaluations du modèle.
 - ▶ Si une évaluation prend environ 0.01 seconde, cela nécessite 581 130.7 heures de calcul.
 - ▶ Avec une ferme de calcul de 1000 processeurs, cela représente environ 581 heures, soit 24.2 jours.
 - ▶ Pour une puissance de 200 Watts par unité centrale, la consommation électrique correspond à 116 226.1 kWh, soit la consommation annuelle de 16 familles françaises environ et un coût de 11.6 K€ (10 c. / kWh).

Une planification raisonnée, un enjeu pour le Développement Durable

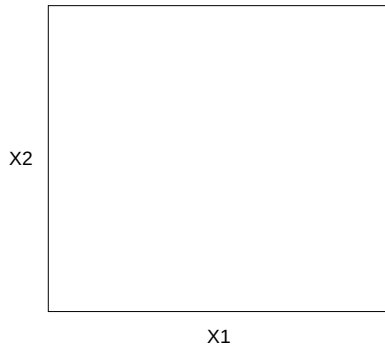
Méthode par plan factoriel complet

- ▶ chaque facteur est décomposé en un certain nombre de niveaux
 - ▶ plan factoriel complet : croisement de tous les niveaux
-
- ▶ Prenons un modèle avec 20 paramètres (facteurs) incertains.
 - ▶ Pour un plan factoriel complet de 20 facteurs à 3 niveaux, cela nécessite $3^{20} = 3\,486\,784\,401$ évaluations du modèle.
 - ▶ Si une évaluation prend environ 0.01 seconde, cela nécessite 581 130.7 heures de calcul.
 - ▶ Avec une ferme de calcul de 1000 processeurs, cela représente environ 581 heures, soit 24.2 jours.
 - ▶ Pour une puissance de 200 Watts par unité centrale, la consommation électrique correspond à 116 226.1 kWh, soit la consommation annuelle de 16 familles françaises environ et un coût de 11.6 K€ (10 c. / kWh).

Une planification raisonnée, un enjeu pour le Développement Durable

Méthode de Morris

Une exploration astucieuse de l'espace

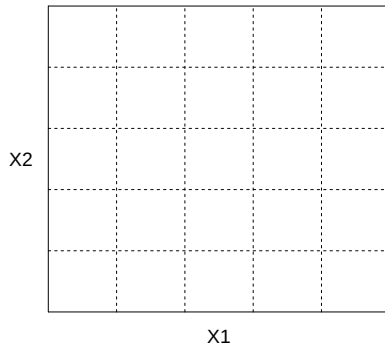


Grands principes



Méthode de Morris

Une exploration astucieuse de l'espace



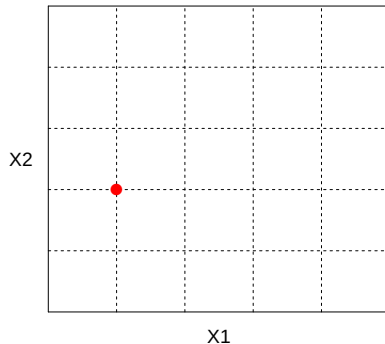
Grands principes

► Discrétisation de l'espace



Méthode de Morris

Une exploration astucieuse de l'espace

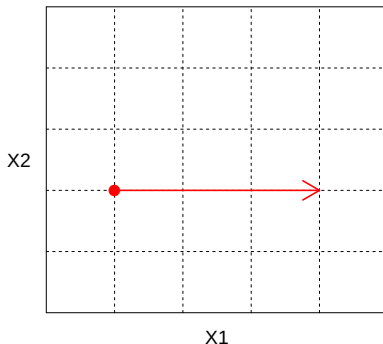


Grands principes

- ▶ Discrétisation de l'espace
- ▶ Un point de départ aléatoire
- ▶

Méthode de Morris

Une exploration astucieuse de l'espace

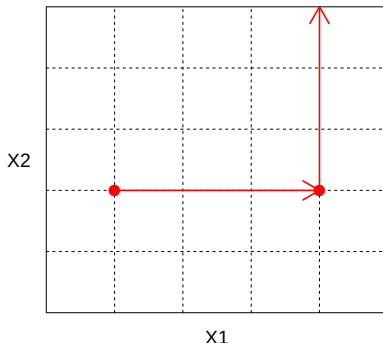


Grands principes

- ▶ Discrétisation de l'espace
- ▶ Un déplacement sur un seul facteur à la fois (méthode OAT One At a Time)
- ▶ Un point de départ aléatoire
- ▶

Méthode de Morris

Une exploration astucieuse de l'espace

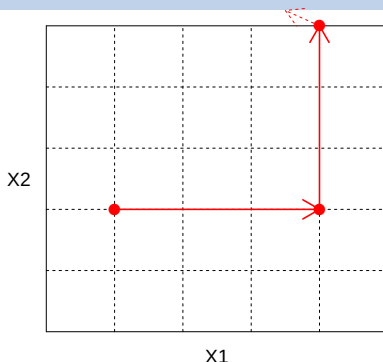


Grands principes

- ▶ Discrétisation de l'espace
- ▶ Un déplacement sur un seul facteur à la fois (méthode OAT One At a Time)
- ▶ Un point de départ aléatoire
- ▶

Méthode de Morris

Une exploration astucieuse de l'espace

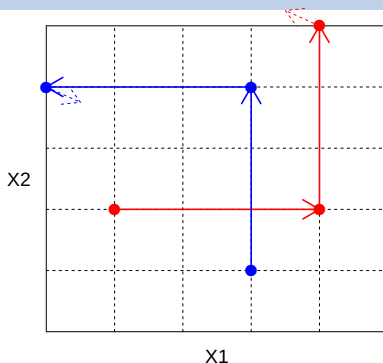


Grands principes

- ▶ Discrétisation de l'espace
- ▶ Un déplacement sur un seul facteur à la fois (méthode OAT One At a Time)
- ▶ Un point de départ aléatoire
- ▶

Méthode de Morris

Une exploration astucieuse de l'espace



Grands principes

- ▶ Discrétisation de l'espace
- ▶ Un déplacement sur un seul facteur à la fois (méthode OAT One At a Time)
- ▶ Un point de départ aléatoire
- ▶ Une trajectoire correspond à K déplacements ($K+1$ points)

Méthode de Morris

Analyse

Mesures d'importance du facteur x_i évaluées sur R trajectoires T^r avec le segment $[\mathcal{G}(T_{(i-1)}^r), \mathcal{G}(T_{(i)}^r)]$ correspondant au saut du facteur x_i

► effet élémentaire : $\Delta_i^r = \frac{\mathcal{G}(T_{(i-1)}^r) - \mathcal{G}(T_{(i)}^r)}{\delta}$, δ la taille du saut

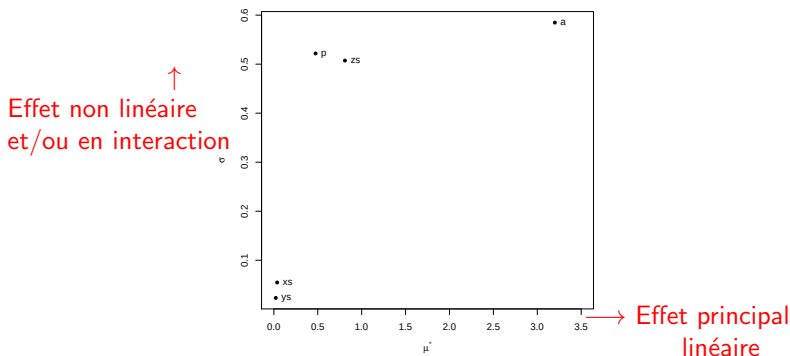
► effect moyen : $\mu_i = \frac{1}{R} \sum_{r=1}^R \Delta_i^r$

► effect absolu moyen : $\mu_i^* = \frac{1}{R} \sum_{r=1}^R |\Delta_i^r|$

► écart-type des effets élémentaires : $\sigma_i = \sqrt{\frac{1}{R-1} \sum_{r=1}^R (\Delta_i^r - \mu_i)^2}$

Méthode de Morris

Résultat du criblage sur la fonction morris, package sensitivity de R



```
morris(compute_wls, factors=names(valfac), r=30, design=list(type="oat",  
levels=5, grid.jump=3))
```

On tire les entrées X sur leur intervalle

- ▶ tirages HyperCube Latin (LHS) ou Monte Carlo ou suite à faible discrédance

plan succinct

1. src
2. décomposition d'Hoeffding - Sobol'
3. indices de solbol
4. méthode de calcul
5. eFast

Indices basés sur la régression

On dispose d'un échantillon $X_1, \dots, X_K$ et des $Y = \mathcal{G}(\mathbf{X})$ correspondants

Coefficient de corrélation linéaire (src de sensitivity)

$$\rho_i = \rho(X_i, Y) = \frac{\text{Cov}(X_i, Y)}{\sqrt{\text{Var}(X_i)}\sqrt{\text{Var}(Y)}}$$

Si $Y = \sum_{i=1}^K \beta_i X_i$ et si les entrées X_i sont indépendantes,
alors $\sum_{i=1}^K \rho^2(X_i, Y) = 1$.

Coefficient de corrélation partielle (pcc de sensitivity)

$$PCC_i = PCC(X_i, Y) = \rho(X_i - \hat{X}_i^{-(i)}, Y - \hat{Y}^{-(i)})$$

Si les entrées sont corrélées, les PCC sont plus appropriés.

Si la relation est monotone mais non linéaire, on travaille sur les rangs (même hypothèse d'additivité).

Tout dépend du pouvoir explicatif du modèle linéaire additif.

Indices basés sur la régression

On dispose d'un échantillon $X_1, \dots, X_K$ et des $Y = \mathcal{G}(\mathbf{X})$ correspondants

Coefficient de corrélation linéaire (src de sensitivity)

$$\rho_i = \rho(X_i, Y) = \frac{\text{Cov}(X_i, Y)}{\sqrt{\text{Var}(X_i)}\sqrt{\text{Var}(Y)}}$$

Si $Y = \sum_{i=1}^K \beta_i X_i$ et si les entrées X_i sont indépendantes,
alors $\sum_{i=1}^K \rho^2(X_i, Y) = 1$.

Coefficient de corrélation partielle (pcc de sensitivity)

$$PCC_i = PCC(X_i, Y) = \rho(X_i - \hat{X}_i^{-(i)}, Y - \hat{Y}^{-(i)})$$

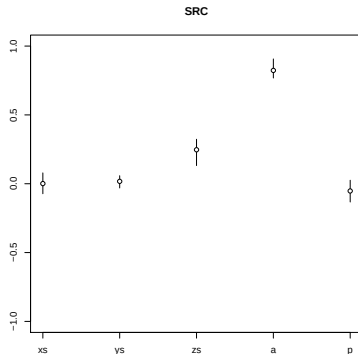
Si les entrées sont corrélées, les PCC sont plus appropriés.

Si la relation est monotone mais non linéaire, on travaille sur les rangs (même hypothèse d'additivité).

Tout dépend du pouvoir explicatif du modèle linéaire additif.

Standardized Regression Coefficients

Application sur le modèle volcan



```
X.lhs = improvedLHS(100, 5, dup=5)  
Y.lhs = apply(X.lhs,1,compute_wls)  
src(X.lhs, Y.lhs, rank = FALSE, nboot = 30, conf = 0.95)
```

Décomposition de Sobol d'une fonction déterministe

Présentation dans le cas de 2 facteurs d'entrée :

Propriété : Pour 'toute' fonction $\mathcal{G}(x_1, x_2)$ sur Ω , il existe une décomposition unique :

$$\mathcal{G}(x_1, x_2) = f_0 + f_1(x_1) + f_2(x_2) + f_{1,2}(x_1, x_2)$$

avec orthogonalité entre les composantes :

$$\int_{[0,1]} \int_{[0,1]} f_U \cdot f_V \, dx_1 \, dx_2 = 0 \text{ si } U \neq V$$

Décomposition d'Hoeffding - Sobol'

Expression des composantes $\Rightarrow$ intégrales

$$f_0 = \int_{[0,1]} \int_{[0,1]} \mathcal{G}(x_1, x_2) dx_1 dx_2$$

$$f_1(\mathbf{x}_1) = \int_{[0,1]} \mathcal{G}(\mathbf{x}_1, x_2) dx_2 - f_0$$

$$f_2(\mathbf{x}_2) = \text{idem}$$

$$f_{1,2}(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{G}(\mathbf{x}_1, \mathbf{x}_2) - f_1(\mathbf{x}_1) - f_2(\mathbf{x}_2) + f_0$$

Conséquences

- décomposition analogue de la variance
- $\Rightarrow$ définition des indices de sensibilité globaux

$$\text{Var}(\mathcal{G}) = \text{Var}(f_1) + \text{Var}(f_2) + \text{Var}(f_{1,2})$$

$$1 = \frac{\text{Var}(f_1)}{\text{Var}(\mathcal{G})} + \frac{\text{Var}(f_2)}{\text{Var}(\mathcal{G})} + \frac{\text{Var}(f_{1,2})}{\text{Var}(\mathcal{G})}$$

 SI_1 SI_2 SI_{12} eff. princ. x_1 eff. princ. x_2

interaction

Décomposition d'Hoeffding - Sobol'

Généralisation

- ≥ 2 facteurs

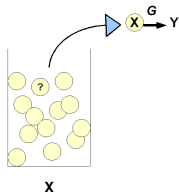
$$\mathcal{G} - f_0 = f_1 + \dots + f_K + f_{1,2} + \dots + f_{1,\dots,K}$$

$$\text{Var}(\mathcal{G}) = \text{Var}(f_1) + \dots + \text{Var}(f_K) + \text{Var}(f_{1,2}) + \dots + \text{Var}(f_{1,\dots,K})$$

$$1 = \frac{\text{Var}(f_1)}{\text{Var}(\mathcal{G})} + \dots + \frac{\text{Var}(f_K)}{\text{Var}(\mathcal{G})} + \frac{\text{Var}(f_{1,2})}{\text{Var}(\mathcal{G})} + \dots + \frac{\text{Var}(f_{1,\dots,K})}{\text{Var}(\mathcal{G})}$$

$$\text{SI}_1 \quad \dots \quad \text{SI}_K \quad \text{SI}_{12} \quad \dots \quad \text{SI}_{1,\dots,K}$$

Interprétation probabiliste



Formalisme		
probabiliste		fonctionnel
$Y x$	$=$	$\mathcal{G}(x)$
$E(Y)$	$=$	$\int_{\Omega} \mathcal{G}(x) dx = f_0$
$\text{Var}(Y)$	$=$	$\int_{\Omega} (\mathcal{G}(x) - f_0)^2 dx$

Interprétation probabiliste (effets factoriels) :

$$f_0 = E(Y)$$

$$f_1(x_1) = E(Y|x_1) - E(Y)$$

$$f_2(x_2) = E(Y|x_2) - E(Y)$$

$$f_{1,2}(x_1, x_2) = E(Y|x_1, x_2) - f_1 - f_2 + f_0$$

Interprétation probabiliste (effets principaux) :

$$f_i(x_i) = E(Y|x_i) - E(Y)$$

$$\text{Var}(f_i) = \text{Var} E(Y|x_i)$$

D'où les indices de sensibilité principaux (de 1er ordre)

$$\text{SI}_i = \frac{\text{Var} E(Y|x_i)}{\text{Var}(Y)} = 1 - \frac{E \text{Var}(Y|x_i)}{\text{Var}(Y)}$$

$$\text{TSI}_i = \frac{E \text{Var}(Y|x_{\sim i})}{\text{Var}(Y)} = 1 - \frac{\text{Var} E(Y|x_{\sim i})}{\text{Var}(Y)}$$

Théorème de la variance globale

Indices basés sur une décomposition de la variance

$$\begin{array}{ccccc} \text{Var}[\mathcal{G}(\mathbf{x})] & = & \underbrace{\text{Var}_{x_1} + \text{Var}_{x_2} + \text{Var}_{x_3} + \dots}_{\downarrow} & + & \underbrace{\text{Var}_{x_1.x_2} + \text{Var}_{x_1.x_3} + \dots}_{\downarrow} \\ \downarrow & & & & \\ \text{Variance totale} & & \text{Effets principaux des} & & \text{Termes d'interactions} \\ \text{de la variable} & & \text{facteurs incertains} & & \\ \text{de sortie} & & & & \end{array}$$

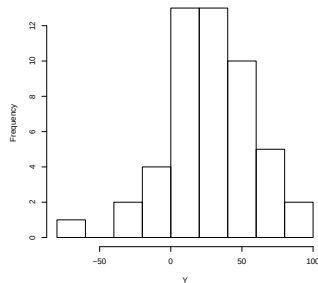
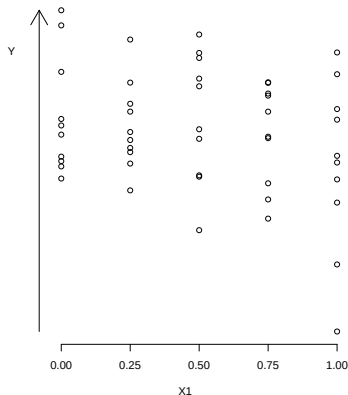
$$\text{Indice de premier ordre de } x_1 \text{ (SI}_1\text{)} = \frac{\text{Var}_{x_1}}{\text{Var}[\mathcal{G}(\mathbf{x})]}$$

$$\text{Indice de sensibilité totale de } x_i \text{ (TSI}_1\text{)} = \frac{\text{Var}_{x_1} + \text{Var}_{x_1.x_2} + \text{Var}_{x_1.x_3} + \dots}{\text{Var}[\mathcal{G}(\mathbf{x})]}$$

Théorème de la variance globale

Principe de la décomposition de la variance

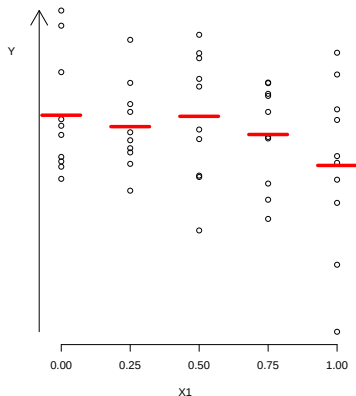
On dispose de mesures répétées sur 5 valeurs de x_1



Théorème de la variance globale

Visualisation des indices

On calcule la variance des moyennes estimées sur les 5 points.

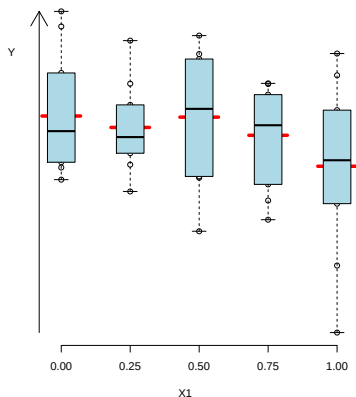


$$SI_1 = Var[Esp(\mathcal{G}(\mathbf{x})|x_1)]$$

Théorème de la variance globale

Visualisation des indices

On calcule la moyenne des variances.



$Esp(Var(\mathcal{G}|x_1))$

Théorème de la variance totale

$$Var[\mathcal{G}(\mathbf{x})] = \text{Var}[Esp(\mathcal{G}|x_1)] + Esp(Var(\mathcal{G}|x_1))$$

Théorème de la variance globale

Autre formulation des indices

- ▶ $SI_i = \frac{Var[Esp(\mathcal{G}(\mathbf{x})|x_i)]}{Var[\mathcal{G}(\mathbf{x})]}$
- ▶ $TSI_i = \frac{Esp(Var[\mathcal{G}|\mathbf{x}^{(-i)}])}{Var[\mathcal{G}(\mathbf{x})]} = \frac{Esp(Var[\mathcal{G}|\mathbf{x}_{\sim i}])}{Var[\mathcal{G}(\mathbf{x})]}$

Interprétation

- ▶ SI_i : pour chaque valeur du facteur x_i , on fait bouger les autres facteurs et on calcule les moyennes en chaque valeur. Ensuite, on regarde si la moyenne (espérance) bouge.
- ▶ TSI_i : pour tous les n-uplets de valeurs de tous les facteurs ($\mathbf{x}^{(-i)}$) sauf x_i , on fait bouger le facteur x_i et on calcule la variabilité. On en fait ensuite la moyenne.

→ Principe des méthodes d'analyse de variance sur plan factoriel ou des méthodes dites de de Pick and Freeze ou de Sobol' par échantillonnage alternatif, méthodes nécessitant beaucoup d'évaluation du modèle.

La précision dépend de N mais aussi de la dimension K

Théorème de la variance globale

Autre formulation des indices

- ▶ $SI_i = \frac{Var[Esp(\mathcal{G}(\mathbf{x})|x_i)]}{Var[\mathcal{G}(\mathbf{x})]}$
- ▶ $TSI_i = \frac{Esp(Var[\mathcal{G}|\mathbf{x}^{(-i)}])}{Var[\mathcal{G}(\mathbf{x})]} = \frac{Esp(Var[\mathcal{G}|\mathbf{x}_{\sim i}])}{Var[\mathcal{G}(\mathbf{x})]}$

Interprétation

- ▶ SI_i : pour chaque valeur du facteur x_i , on fait bouger les autres facteurs et on calcule les moyennes en chaque valeur. Ensuite, on regarde si la moyenne (espérance) bouge.
 - ▶ TSI_i : pour tous les n-uplets de valeurs de tous les facteurs ($\mathbf{x}^{(-i)}$) sauf x_i , on fait bouger le facteur x_i et on calcule la variabilité. On en fait ensuite la moyenne.
- > Principe des méthodes d'analyse de variance sur plan factoriel ou des méthodes dites de de Pick and Freeze ou de Sobol' par échantillonnage alternatif, méthodes nécessitant beaucoup d'évaluation du modèle.

La précision dépend de N mais aussi de la dimension K

Théorème de la variance globale

Autre formulation des indices

- ▶ $SI_i = \frac{Var[Esp(\mathcal{G}(\mathbf{x})|x_i)]}{Var[\mathcal{G}(\mathbf{x})]}$
- ▶ $TSI_i = \frac{Esp(Var[\mathcal{G}|\mathbf{x}^{(-i)}])}{Var[\mathcal{G}(\mathbf{x})]} = \frac{Esp(Var[\mathcal{G}|\mathbf{x}_{\sim i}])}{Var[\mathcal{G}(\mathbf{x})]}$

Interprétation

- ▶ SI_i : pour chaque valeur du facteur x_i , on fait bouger les autres facteurs et on calcule les moyennes en chaque valeur. Ensuite, on regarde si la moyenne (espérance) bouge.
 - ▶ TSI_i : pour tous les n-uplets de valeurs de tous les facteurs ($\mathbf{x}^{(-i)}$) sauf x_i , on fait bouger le facteur x_i et on calcule la variabilité. On en fait ensuite la moyenne.
- Principe des méthodes d'analyse de variance sur plan factoriel ou des méthodes dites de de Pick and Freeze ou de Sobol' par échantillonnage alternatif, méthodes nécessitant beaucoup d'évaluation du modèle.

La précision dépend de N mais aussi de la dimension K

Approches

- ▶ “quasi-répétitions” : méthodes de Sobol'-Saltelli
- ▶ projections sur une base : méthode FAST
- ▶ passage par un métamodèle : régression, processus gaussiens

Principe de la méthode de Sobol' ou Pick and Freeze

- ▶ base : 2 échantillonnages Monte Carlo, LHS ou suite à faible discrédance, A et B, de taille $N \times K$
- ▶ simulations (évaluations) : sur $N(K + 2)$ combinaisons des facteurs calculées à partir de A et B
- ▶ estimation des indices
- ▶ évaluation de la précision par répétition ou bootstrap

Méthode *Sobol-Saltelli2002* : plan d'expériences

$$A = \begin{matrix} \text{Monte-Carlo} \\ \left(\begin{array}{ccc} x_{A,1;1} & \dots & x_{A,1;K} \\ \dots & \dots & \dots \\ x_{A,N;1} & \dots & x_{A,N;K} \end{array} \right) \end{matrix}$$

$$B = \begin{matrix} \text{Monte-Carlo} \\ \left(\begin{array}{ccc} x_{B,1;1} & \dots & x_{B,1;K} \\ \dots & \dots & \dots \\ x_{B,N;1} & \dots & x_{B,N;K} \end{array} \right) \end{matrix}$$

Combinaisons

$\rightarrow C_i =$

$$\left(\begin{array}{ccccc} x_{A,1;1} & \dots & x_{B,1;i} & \dots & x_{A,1;K} \\ \dots & \dots & \dots & \dots & \dots \\ x_{A,N;1} & \dots & x_{B,N;i} & \dots & x_{A,N;K} \end{array} \right)$$

pour $i = 1, \dots, K$

$\Rightarrow N \times (2 + K)$ scénarios à simuler

Méthode *sobol2002* : estimation

⇒ estimation de SI_i à partir de A et C_i

$$\begin{pmatrix} x_{A,1;1} & \dots & x_{A,1;i} & \dots & x_{A,1;K} \\ \dots & \dots & \dots & \dots & \dots \\ x_{A,N;1} & \dots & x_{A,N;i} & \dots & x_{A,N;K} \end{pmatrix} \begin{pmatrix} y_{A,1} \\ \vdots \\ y_{A,N} \end{pmatrix} \Rightarrow \widehat{TSI}_i = \frac{\|y_A - y_{C_i}\|^2}{2\widehat{\text{Var}}(y_A, y_B)}$$

$$\begin{pmatrix} x_{A,1;1} & \dots & x_{B,1;i} & \dots & x_{A,1;K} \\ \dots & \dots & \dots & \dots & \dots \\ x_{A,N;1} & \dots & x_{B,N;i} & \dots & x_{A,N;K} \end{pmatrix} \begin{pmatrix} y_{C,1} \\ \vdots \\ y_{C,N} \end{pmatrix}$$

⇒ estimation de SI_i à partir de B et C_i

Généralisation

Exemple avec $K = 5$, $u = 1, 2$, $\sim u = 3, 4, 5$.

On dispose de deux tirages indépendants X et X' .

On gèle (Freeze) X_u , puis on régénère d'autres $X_{\sim u}$

- ▶ Tirage de $X = (X_1, X_2, X_3, X_4, X_5)$
- ▶ Construction de $X^u = (X_1, X_2, X'_3, X'_4, X'_5)$
- ▶ On évalue $Y = \mathcal{G}(X)$ et $Y^u = \mathcal{G}(X^u)$

$$\Rightarrow \text{Var}(\text{Esp}(Y|X^u)) = \text{Cov}(Y, Y^u). \text{ On a donc } \text{SI}_u = \frac{\text{Cov}(Y, Y^u)}{\text{Var}(Y)}$$

Méthode d'estimation des indices

Différentes manières de faire les calculs

$V_{X_i}(E_{X_{\sim i}}(Y X_i))$ pour calculer S_i	Calcul sensitivity
$\frac{1}{N} \sum_{j=1}^N f(A)_j f(B_A^{(i)})_j - f_0^2$	Sobol' 1993 (sobol2002)
$\frac{1}{N} \sum_{j=1}^N f(B)_j (f(A_B^{(i)})_j - f(A)_j)$	sobol2007
$V(Y) - \frac{1}{2N} \sum_{j=1}^N \left(f(B_j) - f(A_B^{(i)})_j \right)^2$	soboljansen
$E_{X_{\sim i}}(V_{X_i}(Y X_{\sim i}))$ pour calculer TS_i	
$V(Y) - \frac{1}{N} \sum_{j=1}^N f(A)_j f(A_B^{(i)})_j + f_0^2$	Homma 1996 (sobol2002)
$\frac{1}{N} \sum_{j=1}^N f(A)_j \left(f(A)_j - f(A_B^{(i)})_j \right)$	sobol2007
$V(Y) - \frac{1}{2N} \sum_{j=1}^N \left(f(A_j) - f(A_B^{(i)})_j \right)^2$	soboljansen

$A_B^{(i)}$ Matrice A pour laquelle la colonne i vient de B (C_i)

$B_A^{(i)}$ Matrice B pour laquelle la colonne i vient de A

Table 2 de Saltelli A. *et al.*, 2010.

Précautions

Les indices calculés sont des estimations :

- ▶ propriétés de non-biais et de convergence
- ▶ risque d'imprécision suivant la qualité et la taille de l'échantillonnage
 - ▶ l'estimation de l'indice d'un facteur peut être négative (surtout pour des facteurs peu influents)
 - ▶ l'estimation de l'indice total d'un facteur peut être inférieure à celle de son indice principal
- ▶ nécessité d'estimer la précision → bootstrap

Principe du bootstrap

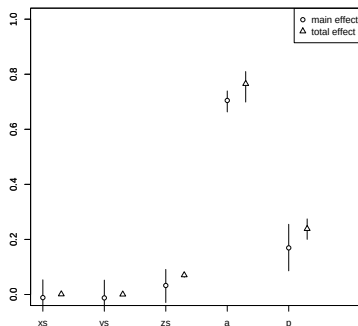
On dispose d'un échantillon de taille N d'une même loi : $\mathbf{y} = (y_1, \dots, y_N)$

- ▶ On calcule sa moyenne et variance
- ▶ On rééchantillonne B fois $\mathbf{y}$ (tirages des y_i avec remise) → $\mathbf{y}^{*1}, \dots, \mathbf{y}^{*B}$
- ▶ on calcule la moyenne sur chaque $\mathbf{y}^{*b}$ → une distribution d'estimateurs de la moyenne

Exemple d'estimation des indices

soboljansen de sensitivity

Application sur le modèle volcan



```
X1 = improvedLHS(1000,5,dup=3) ; X2 = improvedLHS(1000,5,dup=3)
soboljansen(compute_wls, X1,X2, nboot=30, conf = 0.95)
```

Fourier Amplitude Sensitivity Test (FAST)

Principe

- ▶ échantillonnage : trajectoire déterministe remplissant l'espace ("space-filling path") :
 - ▶ $x_{j,i} = \frac{1}{2} + \frac{1}{\pi} \arcsin(\sin(\omega_i s_j + \phi_j))$ avec $s_j = \pi, \dots, \pi$, par pas de π/N
 - ▶ choix raisonné du jeu de fréquences ω_i (harmoniques distinctes entre facteurs jusqu'à l'ordre M (4 à 6))
- ▶ estimation : par analyse fréquentielle

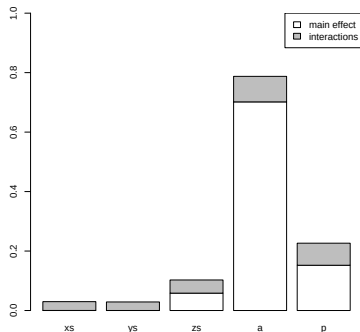
FAST étendu (eFAST)

- ▶ 1 trajectoire FAST par facteur x_i
 - ▶ ω_i élevé
 - ▶ $\omega_k, k \neq i$ faibles
 - ▶ $\Rightarrow$ indices de sensibilité principal et total de x_i

Méthode extended Fast

fast99 de sensitivity

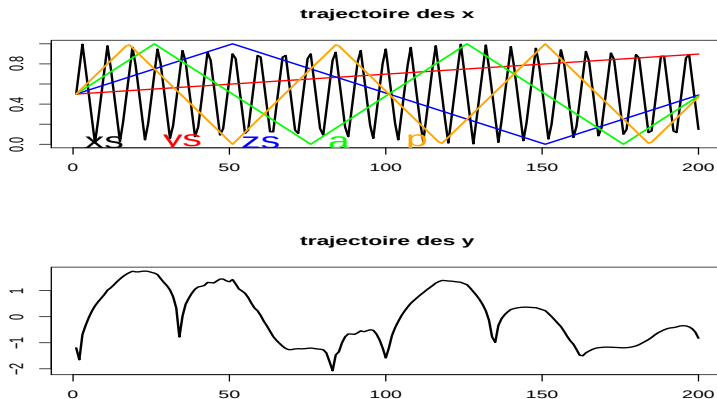
Application sur le modèle volcan



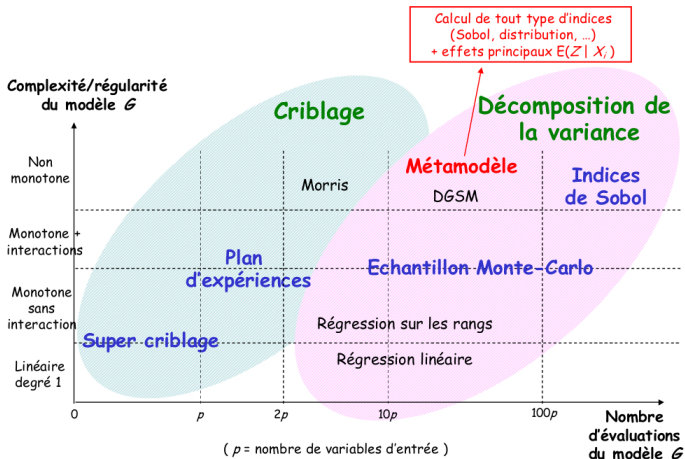
```
fast99(model = compute_wls, names(valfac), 1000, q = rep("qunif",5), q.arg =  
list(min=0,max=1))
```

Principe d'échantillonnage et trajectoires

Application sur le modèle volcan



Classification des méthodes d'analyses de sensibilité



Références

- ▶ Réseau MEXICO <https://reseau-mexico.fr/> Voir les supports des cours des différentes écoles-chercheurs
- ▶ GdR Mascot-Num <http://www.gdr-mascotnum.fr/>
- ▶ Faivre R., Iooss B., Mahévas S., Makowski D., Monod H., editors, 2013. Analyse de sensibilité et exploration de modèles. Applications aux modèles environnementaux. Collection "Savoir Faire", Quae, 2013, 352p.
- ▶ Hoeffding W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19, 293-325. référence historique sur la décomposition d'une fonction de variables par espérances conditionnelles
- ▶ Lamboni M., Monod, Makowski D., 2011. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *RESS*, 96(4), 450-459.
- ▶ Morris M. D., 1991, Factorial sampling plans for preliminary computational experiments, *Technometrics*, 33, 161-174.
- ▶ Pujol G., Iooss B., Janon A., 2017. Package 'sensitivity'
- ▶ Saltelli A., Chan K., Scott E.M., editors., 2000. *Sensitivity Analysis*. Wiley Series in Probability and Statistics, Wiley, New York.
- ▶ Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M., Tarantola S., 2008. *Global Sensitivity Analysis : The Primer*. Wiley, New York.
- ▶ Saltelli A., Tarantola S., and Chan K. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41, 39-56. introduction de FAST étendue
- ▶ Saltelli A., Annoni P., Azzini I., Campolongo F., Ratto M., Tarantola S. , 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications* 181 (2010) 259-270.
- ▶ Sobol', I.M. (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling and Computer Experiments* 1, 407-414. référence sur le calcul des indices de sensibilité