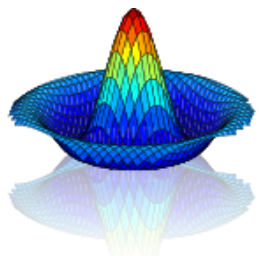
Analyse de sensibilité: mesure de l'importance des facteurs par décomposition de la variance

École Chercheur Mexico

Hervé Monod

INRA - Unité MIA de Jouy-en-Josas

Giens, le 9 juin 2010

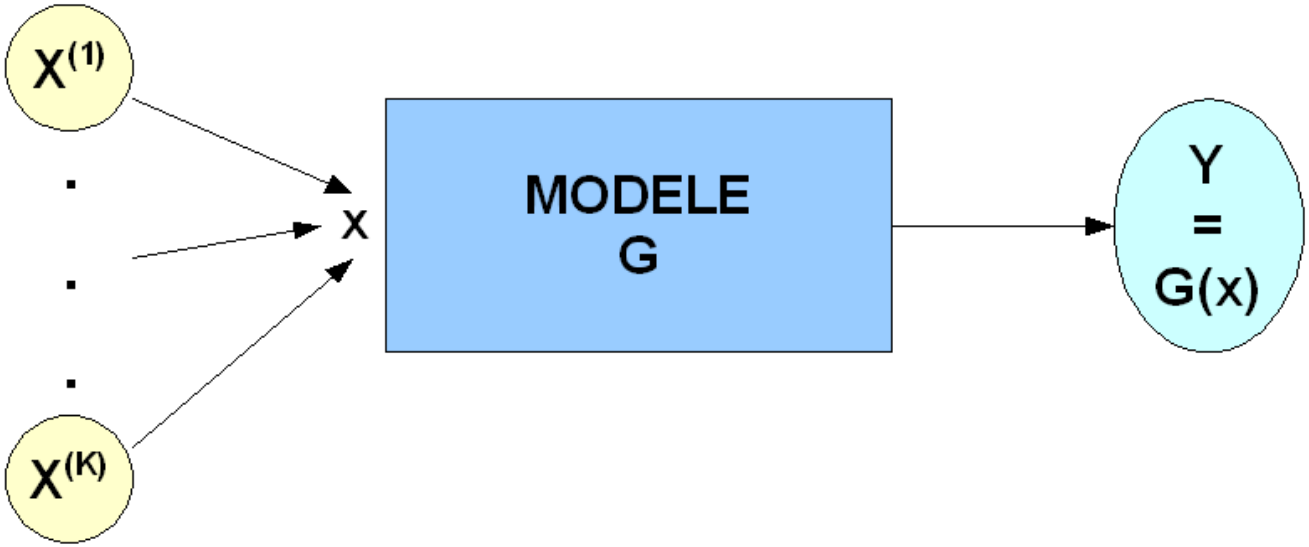


MEXICO

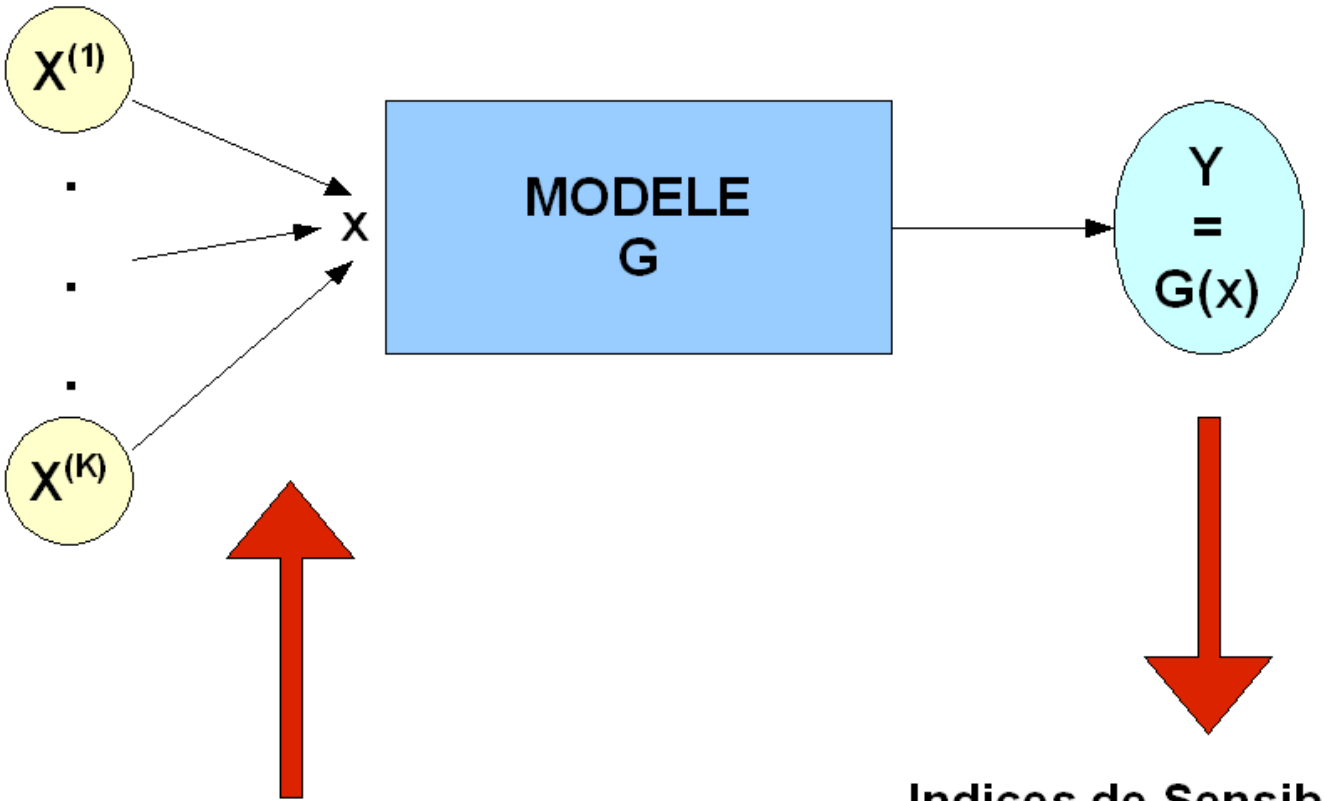
Plan

Plan

Position par rapport aux autres cours



Position par rapport aux autres cours



Simulations:

- planification
- échantillonnage

Indices de Sensibilité:

- définition
- méthode de calcul

Contexte

Modèle = "boîte noire" :

$$Y = \mathcal{G}(x_1, \ldots, x_K)$$

avec

- x_i facteur d'entrée *continu* (paramètre, variable d'entrée du modèle, etc.)
- hypothèses minimales sur la forme du modèle ("model-free")

Objectif : mesurer l'importance globale sur Y des facteurs d'entrée x_k

- sur toute une gamme de variation des x_k
- avec le minimum d'hypothèses sur le modèle
- en intégrant les interactions

Bibliographie

Ilya M. Sobol

`http://mcm2001.sbg.ac.at/invtalk.html`

Andrea Saltelli

`http://www.oecd.org/speaker/`

Histoires parallèles mathématique, statistique, physique :

- Hoeffding W. 1948, *Ann. Math. Statist.* : décomposition de fonctions de variables aléatoires
- Cukier et al., 1973, *J. Chemical Physics* : méthode FAST pour évaluer l'influence de paramètres
- Hora et Iman, 1986, *Sandia Labs* : concept d'importance
- Sobol' I.M., 1990-1993, *Math. Model. Comput. Exp.* : analyse de sensibilité fonctionnelle
- depuis les années 90 : formalisation et nombreux développements (Sobol, Saltelli A. et al., Owen A. ; conférences SAMO, etc.). En particulier
  - Saltelli et al, 1999 : extended FAST (*Technometrics*)
  - Saltelli, 2002 : algo pour indices de Sobol (*Comput. Phys. Commun.*)
  - Saltelli et al, 2010 : étude comparative (*Comput. Phys. Commun.*)

Plan

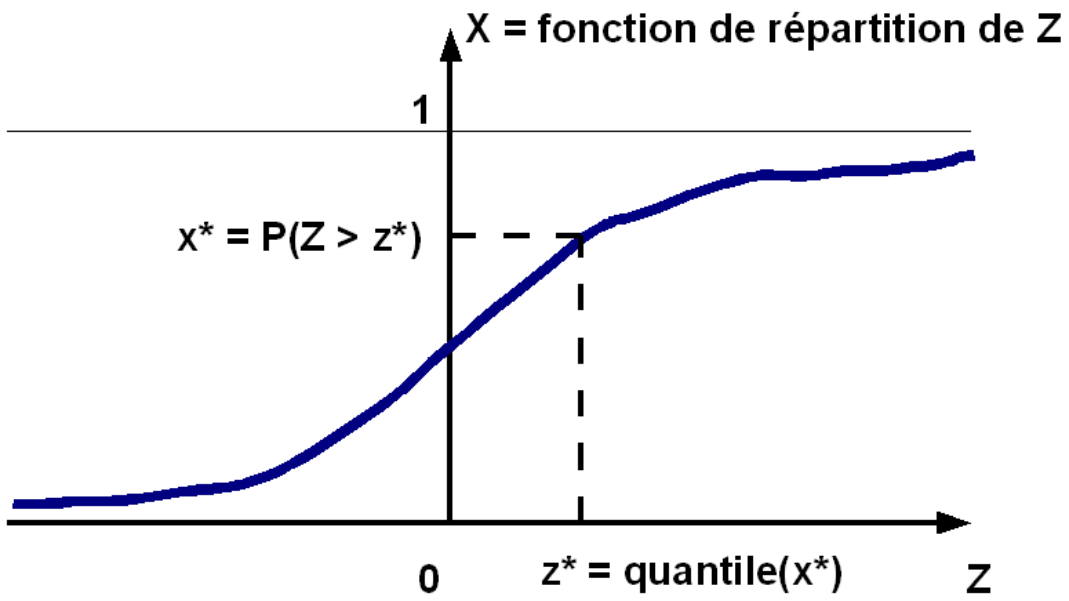
Hypothèses de base

- (1) $x_k \sim$ Loi uniforme sur $[0,1]$ (ss perte de généralité)
- (2) x_k indépendants (complètement croisés)
- (3) modèle = fonction de carré intégrable des entrées

$\Longrightarrow$ Variations des entrées : loi uniforme sur $\Omega = [0,1]^K$

Remarque

Distribution d'incertitude de $Z \Leftrightarrow$ Loi uniforme de X sur $[0, 1]$



Fonction de répartition de Z

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 :

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

Expression des composantes $\Rightarrow$ *intégrales*

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

$$f_1(x_1) = \int_{[0,1]} \mathcal{G}(x_1, x_2)\, dx_2 - f_0$$

$$f_2(x_2) = \text{idem}$$

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

Conséquences

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

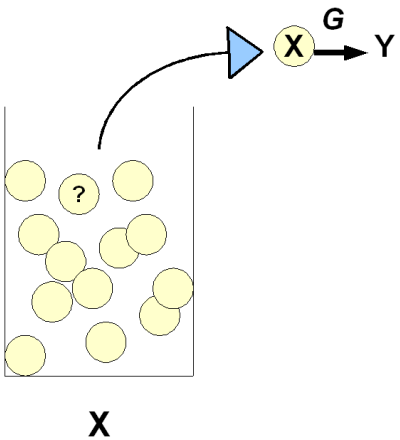
$$\begin{array}{rcccccc} \mathrm{Var}(\mathcal{G}) & = & \mathrm{Var}(f_1) & + & \mathrm{Var}(f_2) & + & \mathrm{Var}(f_{1,2}) \\ 1 & = & \frac{\mathrm{Var}(f_1)}{\mathrm{Var}(\mathcal{G})} & + & \frac{\mathrm{Var}(f_2)}{\mathrm{Var}(\mathcal{G})} & + & \frac{\mathrm{Var}(f_{1,2})}{\mathrm{Var}(\mathcal{G})} \\ & & \mathrm{SI}_1 & & \mathrm{SI}_2 & & \mathrm{SI}_{12} \\ & & \text{eff. princ. } x_1 & & \text{eff. princ. } x_2 & & \text{interaction} \end{array}$$

Généralisation

- ≥ 2 facteurs

$$
\begin{array}{rcccccccccc}
\mathcal{G} - f_0 & = & f_1 & + & \ldots & + & f_K & + & f_{1,2} & + \ldots + & f_{1,\ldots,K} \\
\mathrm{Var}(\mathcal{G}) & = & \mathrm{Var}(f_1) & + & \ldots & + & \mathrm{Var}(f_K) & + & \mathrm{Var}(f_{1,2}) & + \ldots + & \mathrm{Var}(f_{1,\ldots,K}) \\
1 & = & \frac{\mathrm{Var}(f_1)}{\mathrm{Var}(\mathcal{G})} & + & \ldots & + & \frac{\mathrm{Var}(f_K)}{\mathrm{Var}(\mathcal{G})} & + & \frac{\mathrm{Var}(f_{1,2})}{\mathrm{Var}(\mathcal{G})} & + \ldots + & \frac{\mathrm{Var}(f_{1,\ldots,K})}{\mathrm{Var}(\mathcal{G})} \\
 & & \mathrm{SI}_1 & & \ldots & & \mathrm{SI}_K & & \mathrm{SI}_{12} & \ldots & \mathrm{SI}_{1,\ldots,K}
\end{array}
$$

Interprétation probabiliste



Formalisme probabiliste		Formalisme fonctionnel		
$Y\|x$	$=$	$\mathcal{G}(x)$		
$\mathrm{E}(Y)$	$=$	$\int_\Omega \mathcal{G}(x)\, d\, x$	$=$	f_0
$\mathrm{Var}(Y)$	$=$	$\int_\Omega (\mathcal{G}(x) - f_0)^2\, d\, x$		

Interprétation probabiliste (effets factoriels) :

$$
\begin{aligned}
f_0 &= \mathrm{E}(Y) \\
f_1(x_1) &= \mathrm{E}(Y|x_1) - \mathrm{E}(Y) \\
f_2(x_2) &= \mathrm{E}(Y|x_2) - \mathrm{E}(Y) \\
f_{1,2}(x_1, x_2) &= \mathrm{E}(Y|x_1, x_2) - f_1 - f_2 + f_0
\end{aligned}
$$

Interprétation probabiliste (effets principaux) :

$$f_i(x_i) = \mathrm{E}(Y|x_i) - \mathrm{E}(Y)$$

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

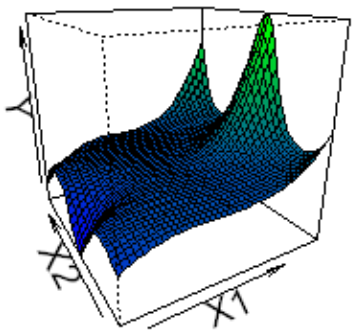
D'où les indices de sensibilité principaux (de 1er ordre)

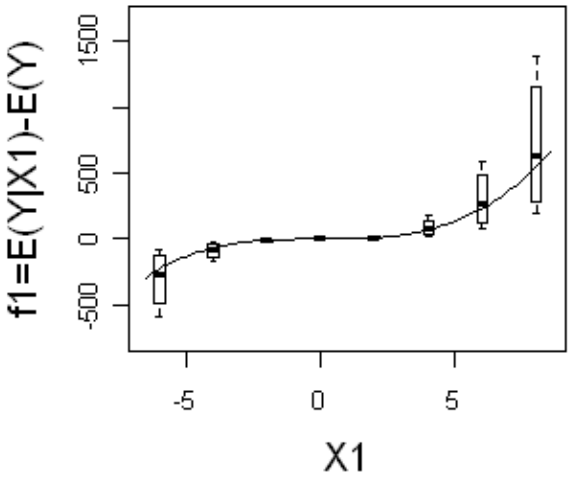
$$\mathrm{SI}_i = \frac{\mathrm{Var}\,\mathrm{E}(Y|x_i)}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{E}\,\mathrm{Var}(Y|x_i)}{\mathrm{Var}(Y)}$$

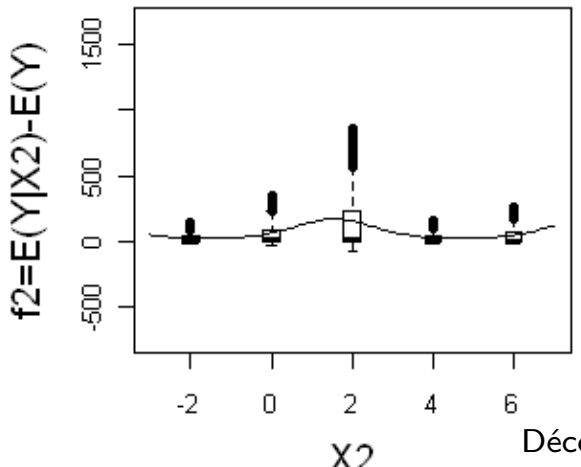
$$\mathrm{TSI}_i = \frac{\mathrm{E}\,\mathrm{Var}(Y|x_{\sim i})}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}\,\mathrm{E}(Y|x_{\sim i})}{\mathrm{Var}(Y)}$$

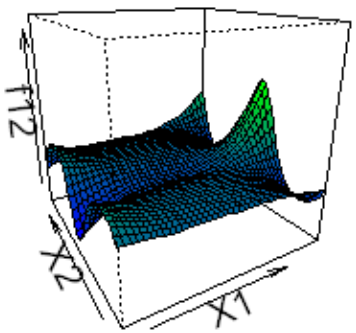
Décomposition de Sobol : exemple

Exemple : fonction $\mathcal{G}(x_1, x_2) = (x_1)^3 \times \exp(\sin(x_2))$









Comment évaluer les indices de sensibilité ?

Pour évaluer les indices de sensibilité :

- impossibilité en général d'un calcul exact
- différence avec l'ANOVA : définition par rapport à des intervalles de variation continus
- nécessité d'un échantillonnage et d'une procédure d'estimation

Comment évaluer les indices de sensibilité ?

Les quatre étapes vues hier :

- Définir les distributions de $x_1, \ldots, x_K$
- Générer des échantillons
- Calculer les réponses du modèle (simulations)
- Estimer les indices et leur précision

Trois approches principales :

- "quasi-répétitions" $\Rightarrow$ méthodes de Sobol-Saltelli (Sobol 1993 ; Saltelli 2002 ; Saltelli et al. 2010)
- projections sur une base $\Rightarrow$ méthodes FAST (Saltelli et al 1999 ; Tarantola et al 2006)
- passage par un méta-modèle $\Rightarrow$ régression, processus gaussiens (Storlie 2009 ; Sudret 2008 ; Petropoulos et al 2009)

Méthode de Sobol

Méthode de Sobol, version Saltelli (2002) : *quasi-répétitions*

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

Remarque : étude par simulations dans Saltelli et al., 2010

Méthode *Sobol-Saltelli2002* : plan d'expériences

$$A = \overbrace{\begin{pmatrix} x_{A,1;1} & \cdots & x_{A,1;K} \\ \cdots & \cdots & \cdots \\ x_{A,N;1} & \cdots & x_{A,N;K} \end{pmatrix}}^{\text{Monte-Carlo}}$$

$$B = \overbrace{\begin{pmatrix} x_{B,1;1} & \cdots & x_{B,1;K} \\ \cdots & \cdots & \cdots \\ x_{B,N;1} & \cdots & x_{B,N;K} \end{pmatrix}}^{\text{Monte-Carlo}}$$

$$\rightarrow C_i = \overbrace{\begin{pmatrix} x_{A,1;1} & \cdots & x_{B,1;i} & \cdots & x_{A,1;K} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{A,N;1} & \cdots & x_{B,N;i} & \cdots & x_{A,N;K} \end{pmatrix}}^{\text{Combinaisons}}$$

pour $i = 1, \ldots, K$

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

Méthode *sobol2002* : estimation

$\Longrightarrow$ estimation de SI_i à partir de A et C_i

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

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

$$\Rightarrow \widehat{\mathrm{TSI}}_i = \frac{\|y_A - y_{C_i}\|^2}{2\widehat{\mathrm{Var}}(y_A, y_B)}$$

$\Longrightarrow$ estimation de SI_i à partir de B et C_i

Méthode *sobol2002* : précautions

Les indices obtenus par calcul numérique sont des estimations :

- propriétés de non-biais et de convergence
- mais 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
- $\rightarrow$ par bootstrap

Sous R (librairie sensitivity)

1. Générer les matrices A et B

```
# taille des deux échantillons de base
    n <- 1000
# tirages uniformes entre -1 et +1
    set.seed(76378)
    X1 <- matrix(runif(7 * n), nrow=n)
    X2 <- matrix(runif(7 * n), nrow=n)
# adaptation aux domaines d'incertitude des facteurs
    wwdm.X1 <- lhs2intervalle(X1, wwdm.factors[1:7,])
    wwdm.X2 <- lhs2intervalle(X2, wwdm.factors[1:7,])
```

Sous R (librairie sensitivity)

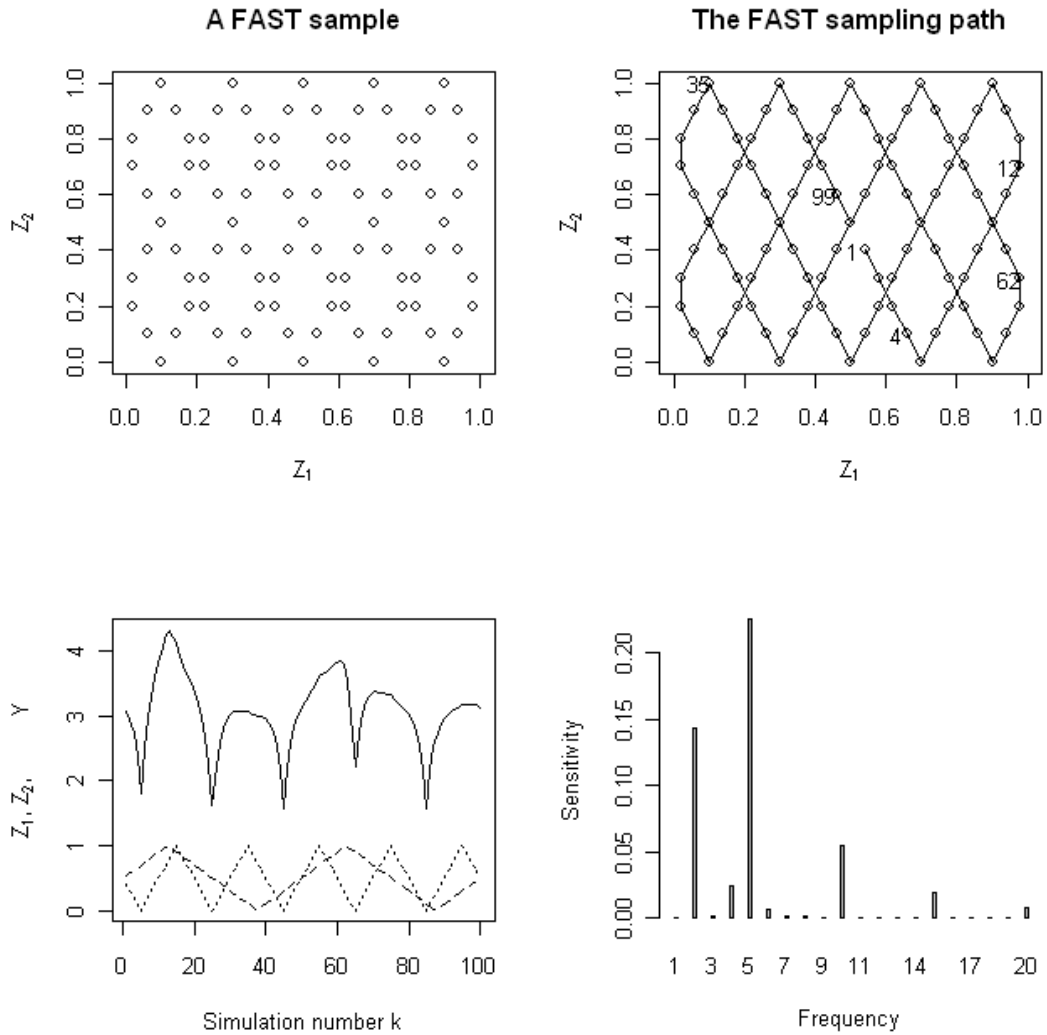
2. Appliquer la méthode de Sobol

```
# utilisation de la fonction sobol2002
    wwdm.sobol2002 <- sobol2002(model=wwdm.simule,
                                X1=wwdm.X1, X2=wwdm.X2,
                                nboot=20,
                                year=3)
# interprétation des résultats
    plot(wwdm.sobol2002)
```

Méthode FAST

Fourier Amplitude Sensitivity Test (FAST) :

- 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 $s_j = -\pi, \ldots, \pi$, par pas de π/N
  - choix raisonné du jeu de fréquences ω_i (but : harmoniques distinctes entre facteurs jusqu'à l'ordre M (4 à 6))
- simulations
- estimation par analyse fréquentielle



(in Monod et al. 2006)

FAST étendu (eFAST) :

- 1 trajectoire FAST par facteur x_k
- ω_k élevé
- $\omega_l, l \neq k$ faibles
- $\Rightarrow$ indices de sensibilité principal et total de x_k

Méthode FAST étendue : précautions

Les indices obtenus par calcul numérique sont des estimations :

- risque d'imprécision suivant la qualité et la taille de l'échantillonnage
- pas de méthode bien définie pour estimer la précision
- possibilité (lourde) : répéter après permutation des fréquences et tirage aléatoire du point de départ

Sous R (librairie sensitivity)

1. Définir les paramètres des lois de distribution

```
# Bornes des intervalles d'incertitude (lois uniformes)
  wwdm.bounds <- apply( cbind(wwdm.factors$binf,
                              wwdm.factors$bsup),
                        1,
                        function(x){list(min=x[1],max=x[2])} )
```

Sous R (librairie sensitivity)

2. Appliquer la méthode FAST étendue

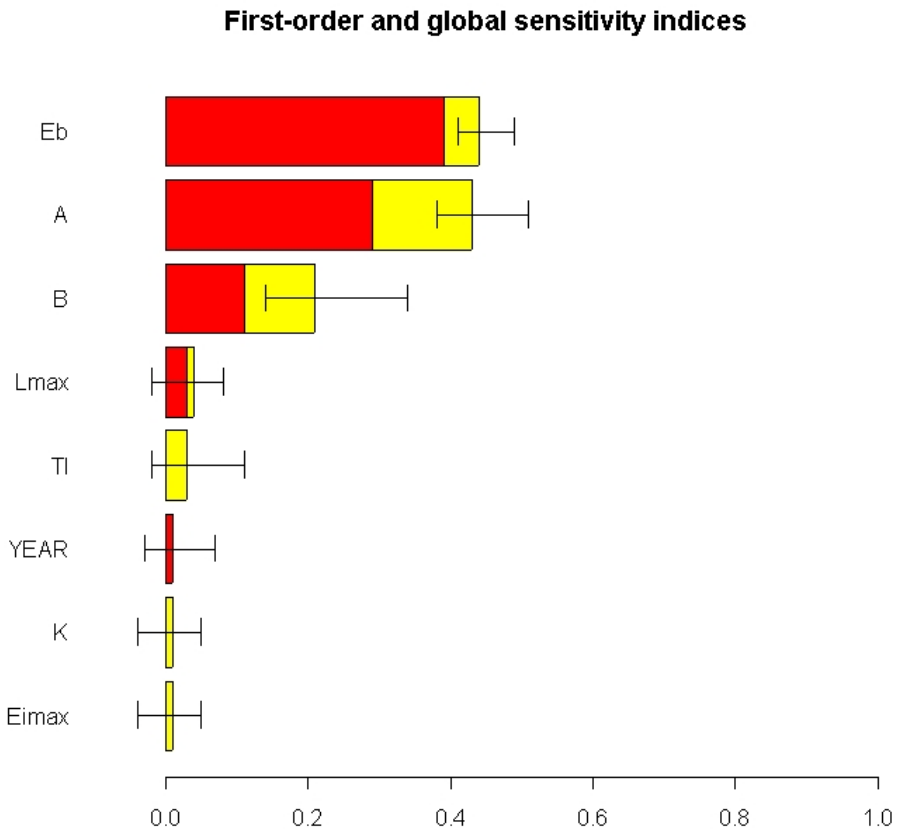
```
# lancement de la commande fast99
    wwdm.fast99 <- fast99(model=wwdm.simule,
                          factors=7,
                          n=1000,
                          q=rep("qunif",7),
                          q.arg=wwdm.bounds)
```

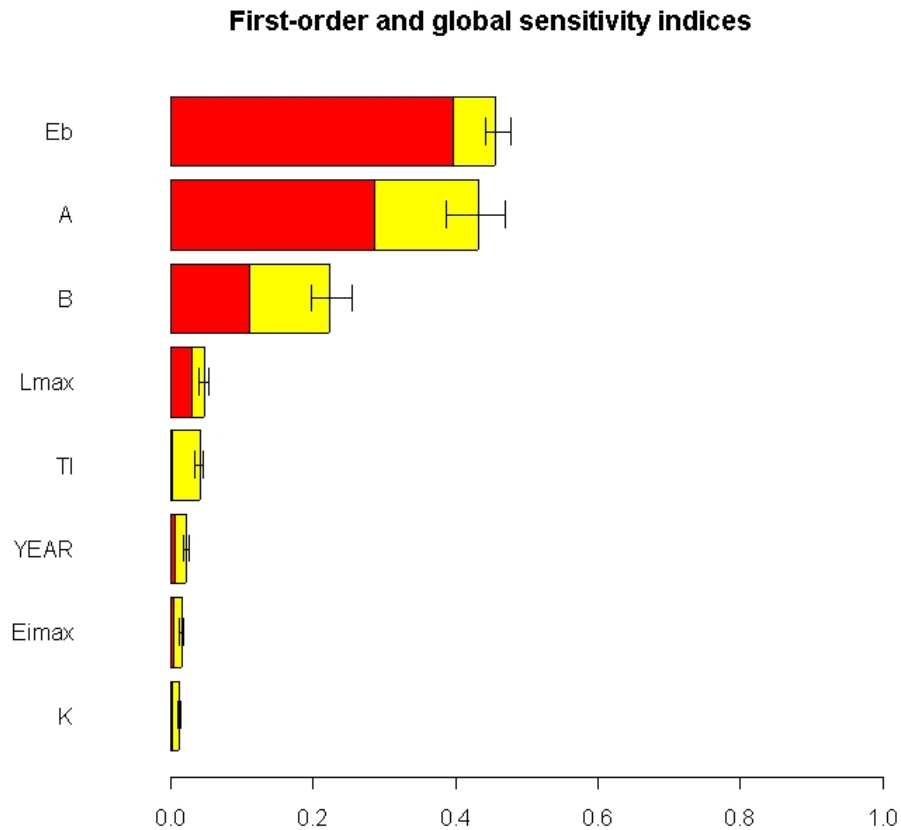
3. Interpréter les résultats

```
# interprétation des résultats
    plot(wwdm.fast99)
```

Comparaison Sobol et FAST étendue

Modèle wwdm :
Sobol (1000×9 répété 20 fois) *vs* eFAST (1000×8 répété 20 fois)





Source : Makowski et al. 2006

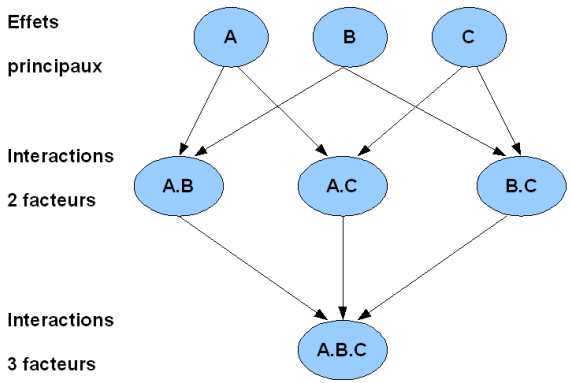
Plan

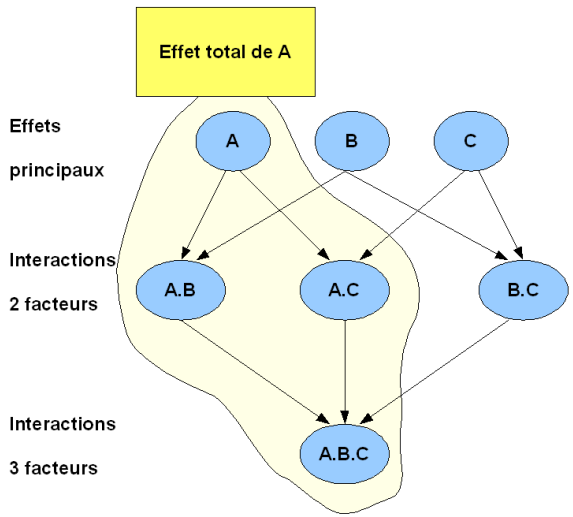
Utilisations concrètes des indices

- priorisation des facteurs (Factor Prioritization) : SI_i
- fixation de facteurs (Factor Fixing) : TSI_i
- diminution de la variance (Variance Cutting) : SI_i
- ?cartographie de l'effet des facteurs (Factor Mapping) ?

Structuration des facteurs

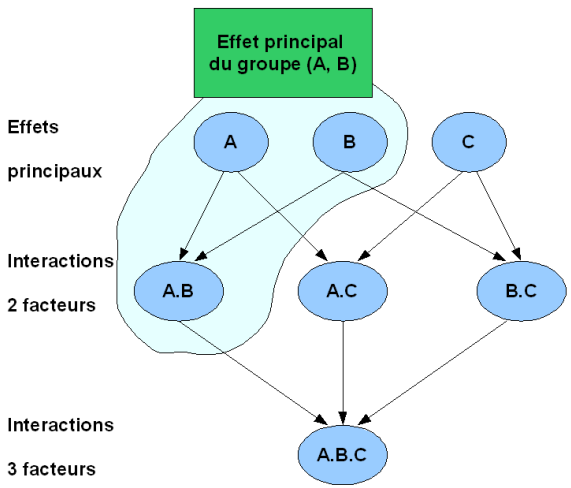
- indices par facteur

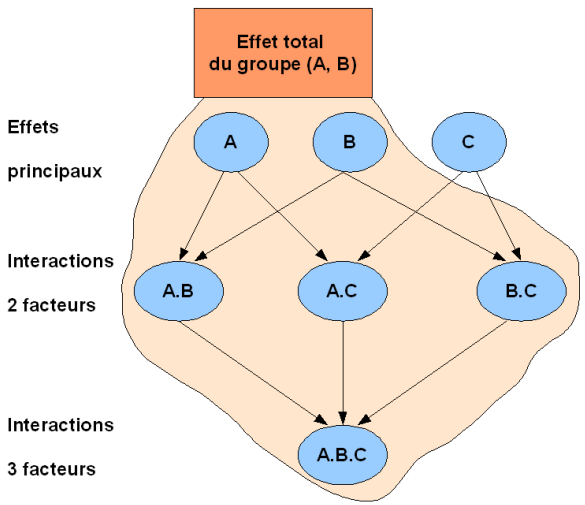




Structuration des facteurs

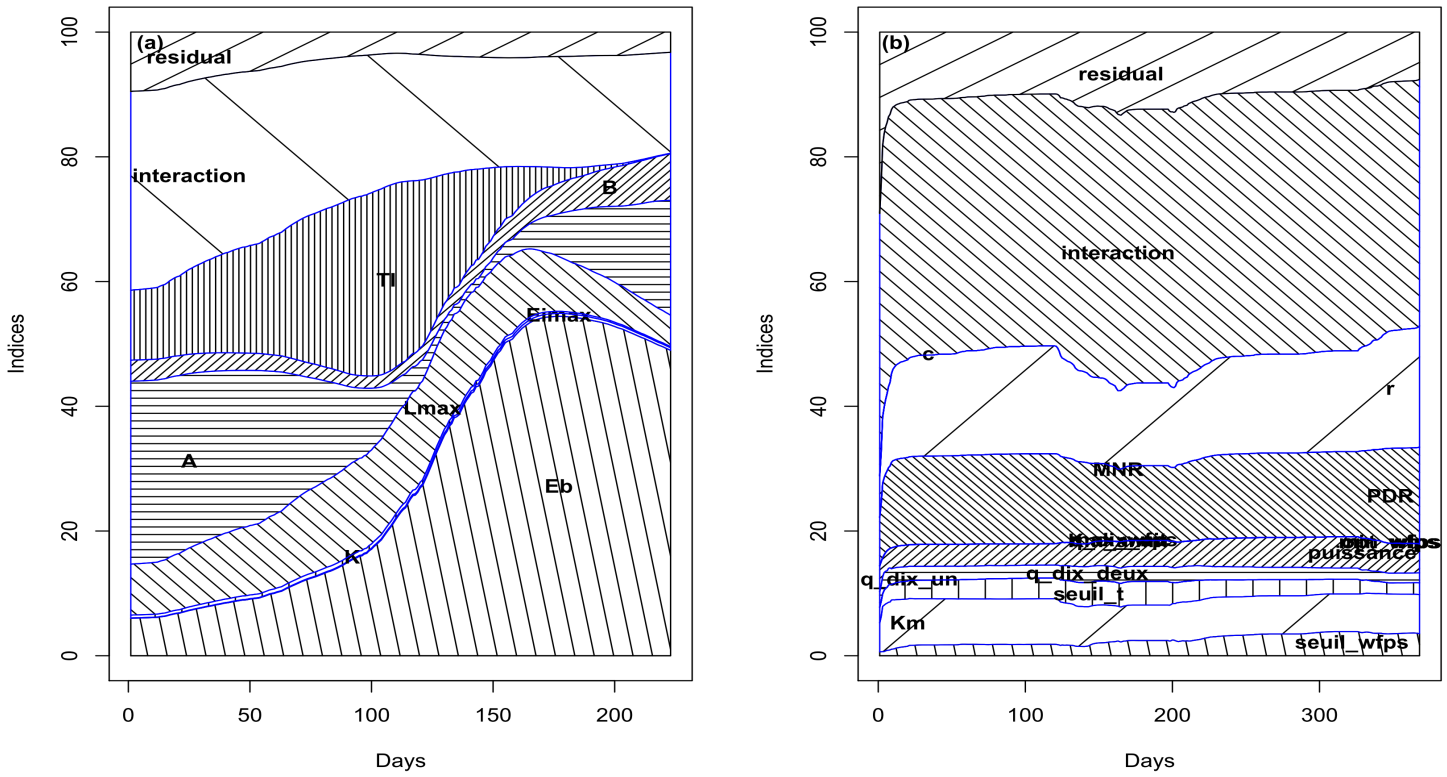
- indices par groupe de facteurs





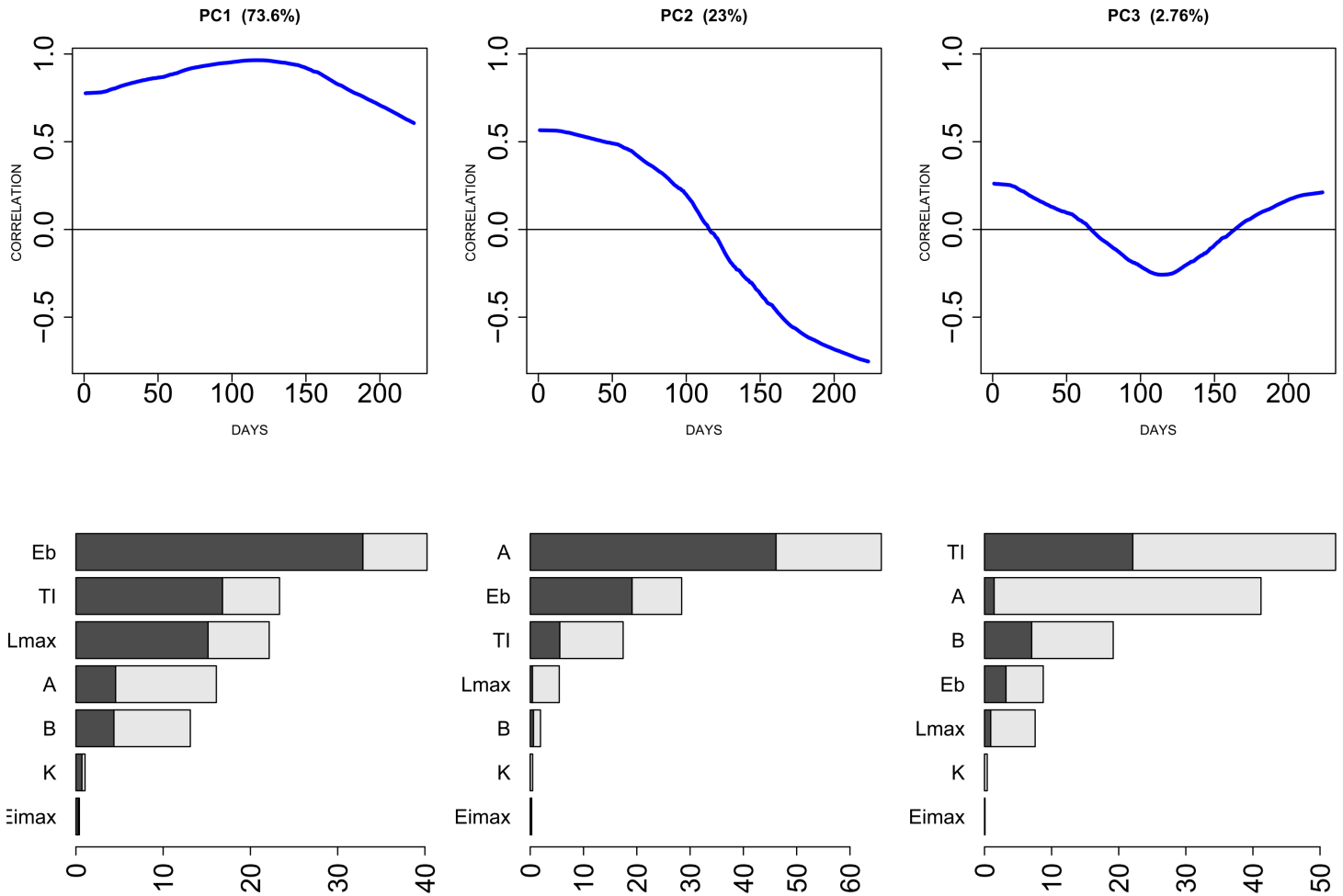
Sorties multivariées (1)

Analyse date par date :



Sorties multivariées (2)

Analyse date par date :



Lamboni et al 2009 ; Lurette et al 2009

Autres variantes et extensions

- analyses de sensibilité sur sorties multivariées : alternatives
- entrées fonctionnelles complexes (Iooss et Ribatet 2009)
- prise en compte de corrélation des entrées (Da Veiga et al 2009)
- indices basés sur l'entropie

Plan

Intérêts et limites

Intérêts :

- Mesure de qualité de l'importance des facteurs
- Prise en compte explicite du continu
- "model-free"
- Mesures synthétiques
- Applications directes à des objectifs ciblés

Limites

- Coût en nombre de simulations
- (Pas complètement satisfaisant sur les interactions)
- A compléter pour comprendre l'effet des facteurs influents

Plan

- Cukier R.I., Schaibly J.H. and Shuler K.E. (1975). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III. Analysis of the approximations. *The Journal of Chemical Physics*, **63**, 1140-1149. *référence historique sur FAST*
- Da Veiga S., Wahl F., Gamboa F. (2009). Local polynomial estimation for sensitivity analysis for models with correlated inputs. *Technometrics*. *méthode pour entrées corrélées*
- Ginot V., Gaba S., Beaudouin R., Aries F., Monod H. (2006). Combined use of local and ANOVA-based global sensitivity analyses for the investigation of a stochastic dynamic model : application to the case study of an individual-based model of a fish population. *Ecological Modelling* **193**, 479-491. *application de méthodes d'AS locales et globales*
- 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*

- Iooss B., Ribatet M. (2009). Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering and System Safety*, **94**, 1194-1204. *analyse de sensibilité pour des facteurs d'entrée complexes*
- Ishigami T. and Homma T. (1989). An importance quantification technique in uncertainty analysis for computer models. *Japan Atomic Energy Research Institute Report JAERI-M* 89-111. *article de référence sur les indices de sensibilité*
- Lamboni M., Makowski D., Lehuger S., Gabrielle B., Monod H. (2009). Multivariate global sensitivity analysis for dynamic crop models. *Fields Crop Research*, **113**, 312-320. *analyse de sensibilité pour sorties multivariées*
- Lurette A., Touzeau S., Lamboni M. Monod H. (2009). Sensitivity analysis to identify key parameters influencing Salmonella infection in a pig batch. *Journal of Theoretical Biology*, **258**, 43-52. *analyse de sensibilité pour sorties multivariées*
- Makowski D., Naud C., Monod H., Jeuffroy M.-H., Barbottin A. (2006). Global sensitivity analysis for calculating the contribution of

genetic parameters to the variance of crop model prediction. *Reliability Engineering and System Safety* **91**, 1142–1147. *comparaison de méthodes par simulations sur un modèle*

- Monod h., Naud, C., Makowski, D. (2006). Uncertainty and sensitivity analysis for crop models. In : *Working with Dynamic Crop Models, Evaluation, Analysis, Parameterization, and Applications (Wallach, D., Makowski, D., Jones, J.W. Eds)*, Chapter 4, pp. 55-100. Elsevier : Amsterdam. *revue sur les méthodes d'analyse de sensibilité et d'incertitude*
- Petropoulos G., Wooster M.J., Carlson T.N. Kennedy M.C., Scholze M. (2009). A global Bayesian sensitivity analysis of the 1d SimSphere soil-vegetation-atmospheric transfer (SVAT) model using Gaussian model emulation. *Ecological Modelling*, **x**. *analyse de sensibilité via un méta-modèle de processus gaussien et une approche bayésienne*
- Pujol G., Iooss B. (2008). The sensitivity package, v1.4-0. Documentation d'un package R accessible sur le Web.
- 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. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, **145**, 280-297. *amélioration de la méthode de Sobol*
- Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M., Tarantola S. (2008). *Global Sensitivity Analysis. The Primer.* Wiley. *ouvrage de synthèse le plus récent*
- 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**, 259-270. *comparaison par simulations de variantes de la méthode de Sobol-Saltelli*
- 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*
- Storlie C.B., Swiler L.P., Helton J.C., Sallaberry C.J. (2009). Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding

models. *Reliability Engineering and System Safety*, **94**, 1735-1763. *analyse de sensibilité via un méta-modèle non paramétrique*

- Sudret B. (2008). Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering and System Safety*, **93**, 964-979. *analyse de sensibilité via un méta-modèle*
- Tarantola S., Gatelli D., Mara T. (2006). Random balance designs for the estimation of first order indices. *Reliability Engineering and System Safety*, **91**, 717-727. *variante récente de la méthode FAST*