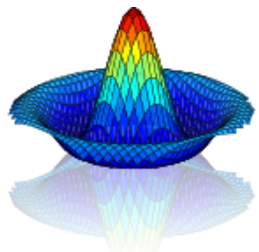
Analyse de sensibilité: mesure de l'importance des facteurs par décomposition de la variance

École Chercheur Mexico

Hervé Monod

INRA - Unité MIA de Jouy-en-Josas

Giens, le 12 mai 2009



MEXICO

Plan

1. Introduction
2. Définition des indices de sensibilité
   - Décomposition de Sobol
   - Types d'indices et leur interprétation
3. Méthodes de calcul
   - Méthode de Sobol, version Saltelli 2002
   - Méthode FAST étendue (Saltelli et al. 1999)
4. Comparaison
5. Discussion
6. Références

Plan

1. Introduction
2. Définition des indices de sensibilité
   - Décomposition de Sobol
   - Types d'indices et leur interprétation
3. Méthodes de calcul
   - Méthode de Sobol, version Saltelli 2002
   - Méthode FAST étendue (Saltelli et al. 1999)
4. Comparaison
5. Discussion
6. Références

Caractéristiques

Méthodes "basées sur la variance" (variance-based methods) :

- Modèle = "boîte noire" :

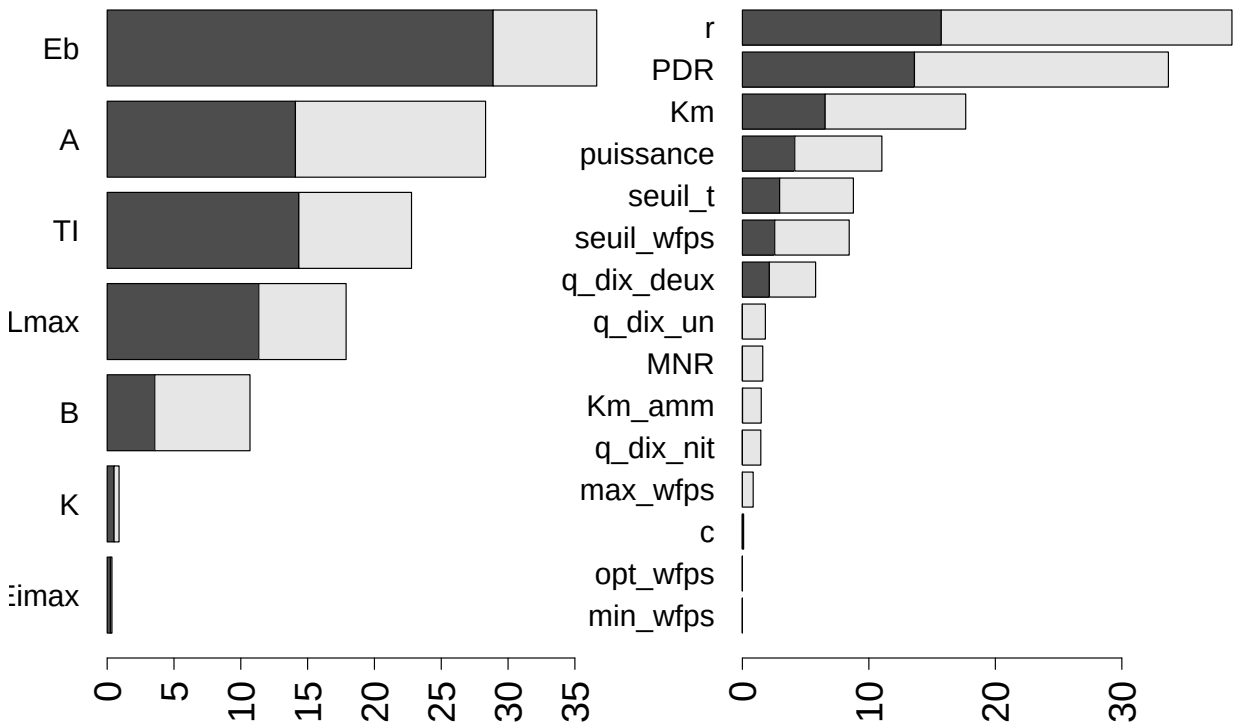
$$Y = \mathcal{G}(x^{(1)}, \ldots, x^{(K)}) \quad (1)$$

$$\text{ou } Y = G(x^{(1)}, \ldots, x^{(K)}) \quad (2)$$

- Objectif général :
  - mesurer l'importance des facteurs d'entrée (paramètres, variables d'entrée du modèle, etc.)
  - hypothèses minimales sur la forme du modèle ("model-free")
  - prise en compte de la nature continue des facteurs d'entrée
  - prise en compte des interactions
- Critère d'importance = contribution à la variabilité de la sortie
- Méthodes très coûteuses en nombre de simulations

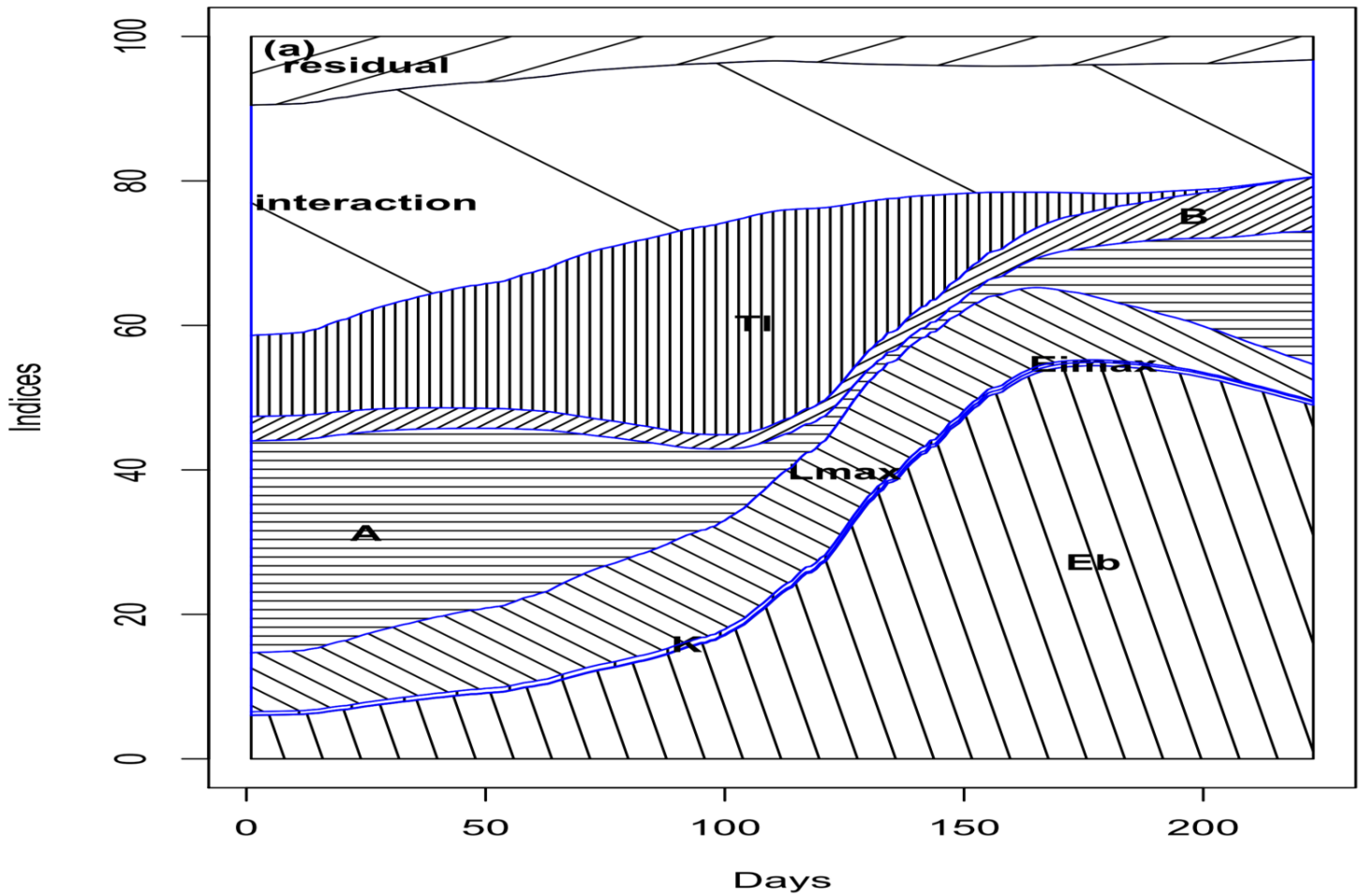
Résultat attendu : mesures globales d'importance

Exemple : modèle de culture wwdm (Y = biomasse), modèle biophysique Ceres-N20 (Y = émission de N_2O par une parcelle agricole)



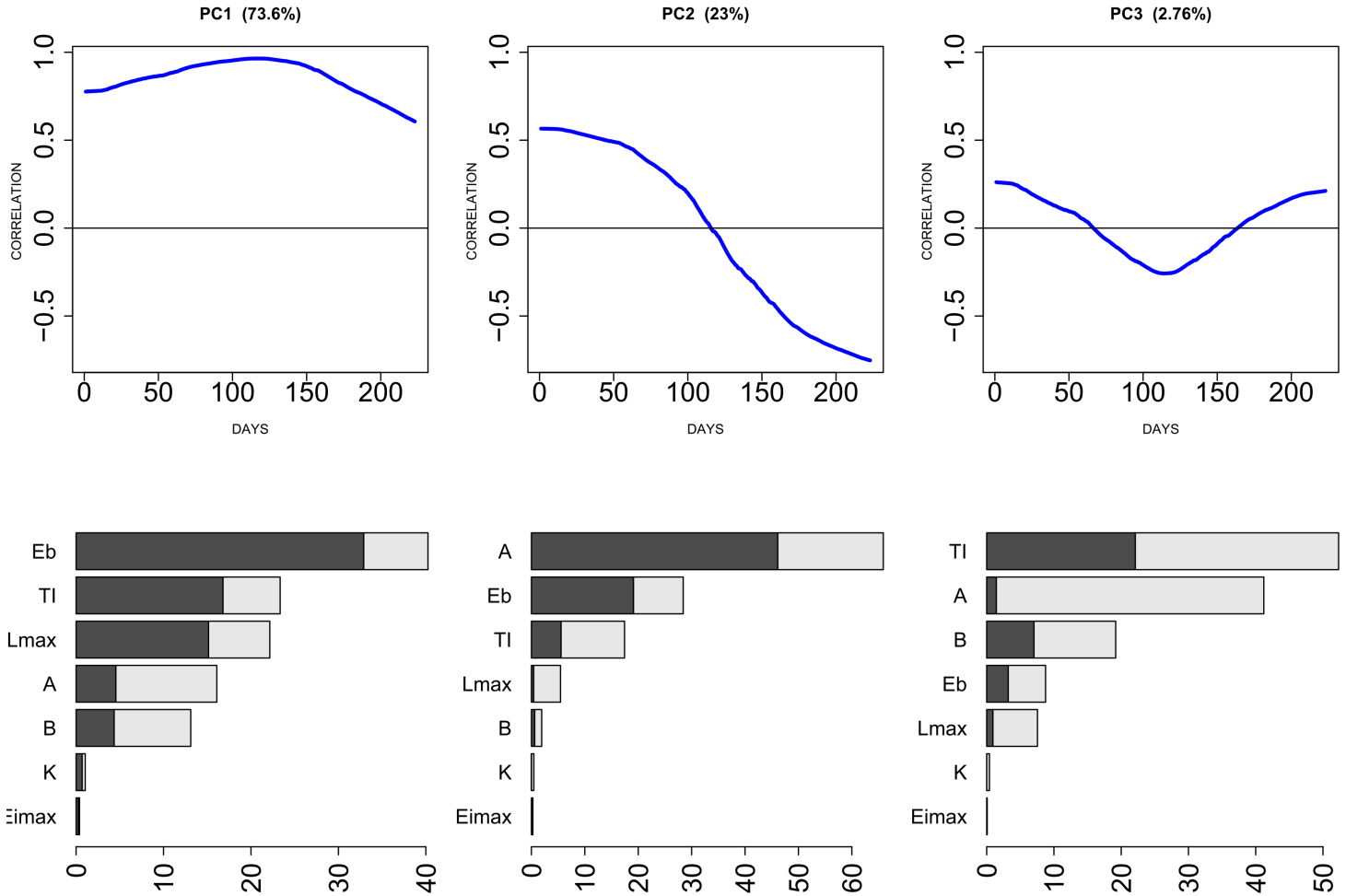
Résultat attendu : cas multivarié

Exemple : modéle de culture wwdm, matière sèche journalière



Voir aussi : Saltelli, Tarantola et Campolongo, Statist.Sci. 2000

Résultat attendu : cas multivarié (2)



Lamboni et al., 2009

Historique

Très brefs éléments d'histoire :

- méthode FAST apparue en chimie (Cukier et al., 1973, *J. Chemical Physics*)
- concept d'importance proposé par Hora et Iman (1986, *Sandia Labs*)
- travaux de Sobol, de l'intégration en grande dimension (1967) à l'analyse de sensibilité fonctionnelle (1990, 1993) - concept HDMR (High Dimensional Model Representation)
- formalisation et nombreux développements depuis les années 1990 (Sobol, Saltelli et al., conférences SAMO, etc.)

Principe

Principe général des méthodes basées sur la variance :

- incertitude sur chaque $x^{(k)}$ modélisée par une distribution de probabilité *continue*
- définition précise des indices de sensibilité qui dépendent :
  - du modèle
  - des lois de distribution des $x^{(k)}$
  - de la variable de sortie étudiée Y
- calcul analytique des indices (rarement possible) . . .
- . . . ou calcul approché par échantillonnage-simulation-estimation en utilisant la fonction-code G :
  - plan factoriel + analyse de variance
  - méthode de Sobol et amélioration de Saltelli 2002
  - méthodes FAST et FAST étendue

Plan

1 Introduction

2 Définition des indices de sensibilité
- Décomposition de Sobol
- Types d'indices et leur interprétation

3 Méthodes de calcul
- Méthode de Sobol, version Saltelli 2002
- Méthode FAST étendue (Saltelli et al. 1999)

4 Comparaison

5 Discussion

6 Références

Décomposition de Sobol d'une fonction déterministe

Cas à 2 facteurs d'entrée pour simplifier la présentation

Propriété : Pour 'toute' fonction $Y = \mathcal{G}(x^{(1)}, x^{(2)})$ il existe une décomposition unique :

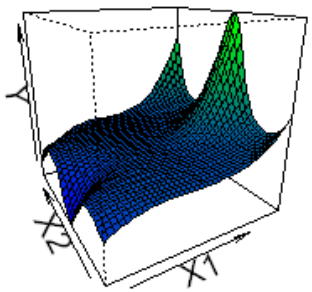
$$Y = f_0 + f_1(x^{(1)}) + f_2(x^{(2)}) + f_{1,2}(x^{(1)}, x^{(2)})$$

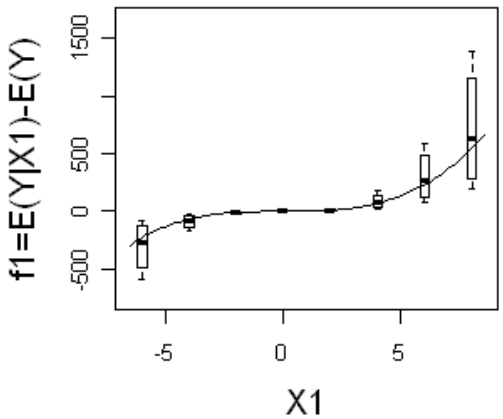
avec orthogonalité entre les composantes :

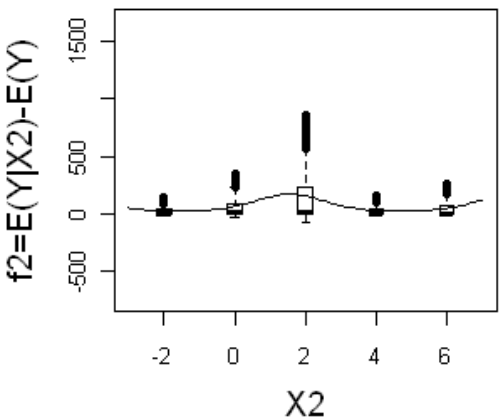
$$\int_{D_2} \int_{D_1} f_U(x^{(1)}, x^{(2)})\, f_V(x^{(1)}, x^{(2)})\, dx^{(1)}\, dx^{(2)} = 0 \text{ si } U \neq V$$

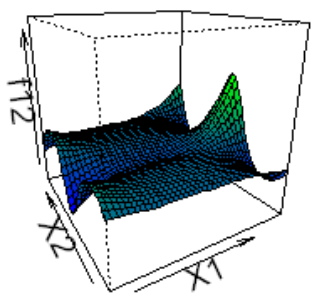
Décomposition de Sobol : exemple

Exemple : fonction $\mathcal{G}(x^{(1)}, x^{(2)}) = (x^{(1)})^3 \times exp(sin(x^{(2)}))$









Conséquence de l'orthogonalité :

- décomposition de la variance
- définition des indices de sensibilité

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(f_1) + \mathrm{Var}(f_2) + \mathrm{Var}(f_{1,2}) \\
1 &= \frac{\mathrm{Var}(f_1)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(f_2)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(f_{1,2})}{\mathrm{Var}(Y)} \\
1 &= \mathrm{SI}_1 + \mathrm{SI}_2 + \mathrm{SI}_{12}
\end{aligned}$$

eff. princ. $x^{(1)}$ | eff. princ. $x^{(2)}$ | interaction

Formulation mathématique (effets principaux) :

$$f_i(x^{(i)}) = \text{ intégrale multiple de } f \text{ sur les } x^{(j)}, j \neq i$$

Interprétation probabiliste (effets principaux) :

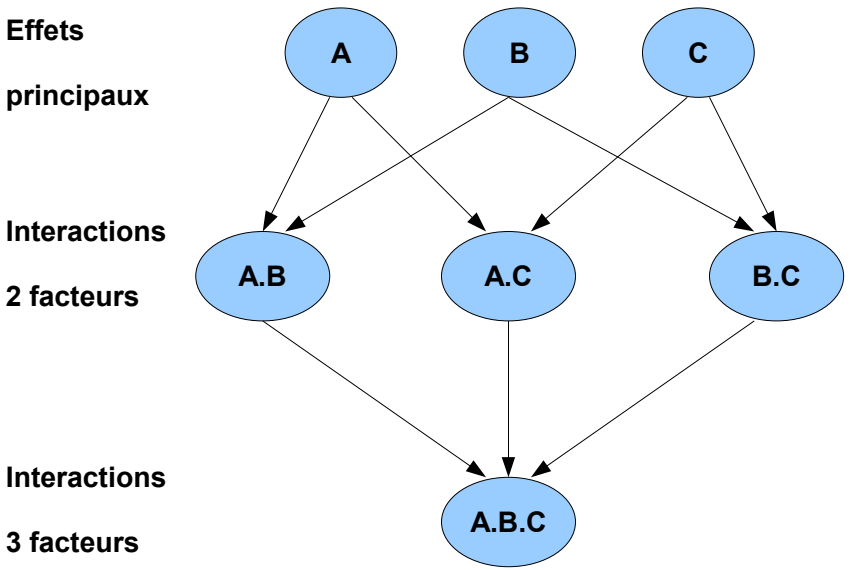
$$\begin{aligned} f_i(x^{(i)}) &= \mathrm{E}(Y|x^{(i)}) - \mathrm{E}(Y) \\ \mathrm{Var}(f_i) &= \mathrm{Var}\ \mathrm{E}(Y|x^{(i)}) \\ \mathrm{SI}_i &= \frac{\mathrm{Var}\ \mathrm{E}(Y|x^{(i)})}{\mathrm{Var}(Y)} \\ &= 1 - \frac{\mathrm{E}\ \mathrm{Var}(Y|x^{(i)})}{\mathrm{Var}(Y)} \end{aligned}$$

Interprétation probabiliste (interactions) :

$$
\begin{aligned}
f_0 &= \mathrm{E}(Y) \\
f_1(x^{(1)}) &= \mathrm{E}(Y|x^{(1)}) - \mathrm{E}(Y) \\
f_2(x^{(2)}) &= \mathrm{E}(Y|x^{(1)}) - \mathrm{E}(Y) \\
f_{1,2}(x^{(1)}, x^{(12)}) &= \mathrm{E}(Y|x^{(1)}, x^{(2)}) - f_1 - f_2 + f_0
\end{aligned}
$$

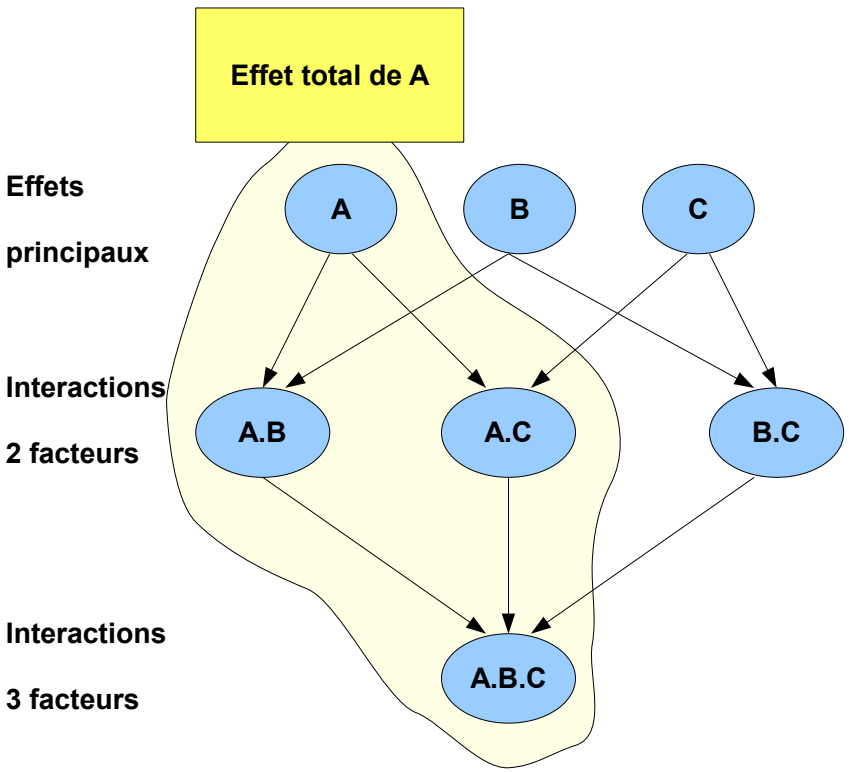
Indices de sensibilité

Indices de Sensibilité factoriels (SI)



Indices de sensibilité

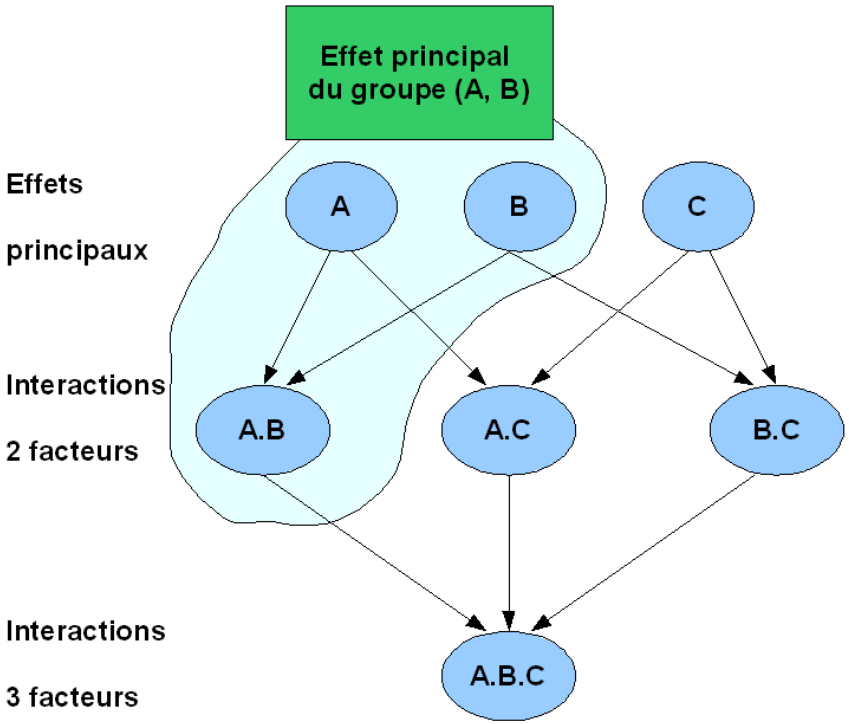
Indice de Sensibilité totale à A (TS_A)



1 indice de sensibilité totale par facteur

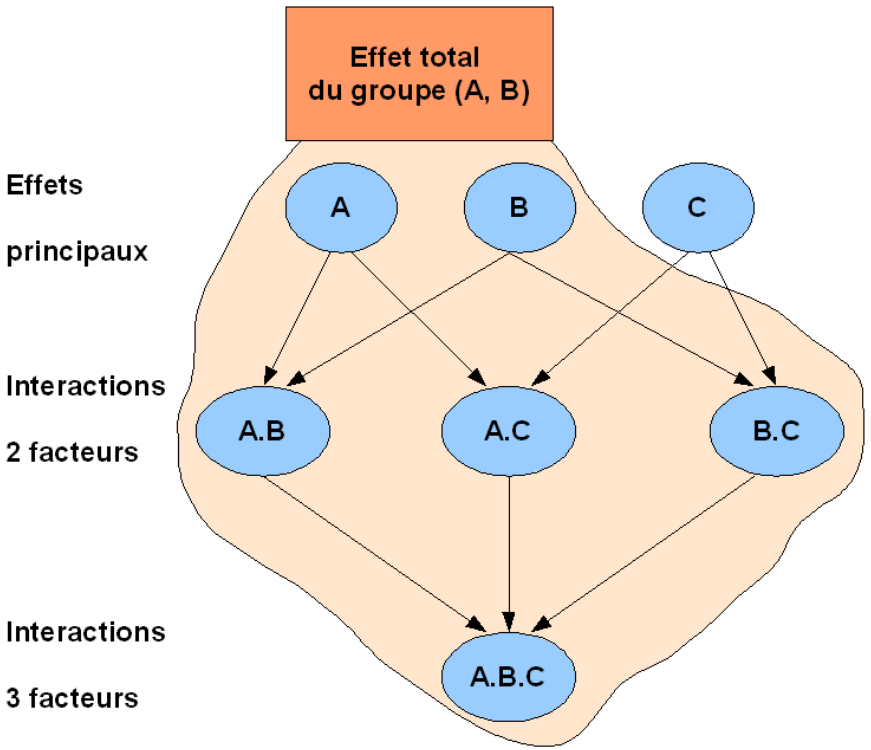
Indices de sensibilité

Sensibilité principale au groupe $\{A, B\}$ $(\mathrm{SI}_{\{A,B\}})$



Indices de sensibilité

Sensibilité totale au groupe $\{A, B\}$ *(*$\mathrm{TS}_{\{A,B\}}$*)*



Remarque : $\mathrm{TS}_{\{A,B\}} + \mathrm{SI}_C = 1$

Objectifs

Objectifs ciblés

- Priorisation des facteurs (Factor Prioritization)
  - -> quels facteurs faut-il déterminer en priorité pour réduire $\mathrm{Var}(Y)$?
  - -> ceux à effets principaux SI_i les plus élevés
- Fixation de facteurs (Factor Fixing)
  - -> quels facteurs peut-on fixer sans conséquence sur Y ?
  - -> ceux à effets totaux TSI_i les plus élevés
- diminution de la variance (Variance Cutting)
  - -> comment baisser $\mathrm{Var}(Y)$ sous un un seuil de tolérance en jouant sur un minimum de facteurs ?
  - -> rechercher un groupe de facteurs à effets $\mathrm{SI}_{\{i_1,\ldots,i_n\}}$ le plus élevé possible
- (cartographie de l'effet des facteurs) (Factor Mapping)

Retenir

- les indices de sensibilité sont des mesures très synthétiques de la sensibilité du modèle aux facteurs
- ils ont une définition mathématique précise
  - basée sur la variance
  - prenant en compte la totalité du domaine d'incertitude
  - indépendante de toute simulation
- mais il faut en général des simulations pour les estimer
- l'interprétation dépend des objectifs

Plan

1 Introduction

2 Définition des indices de sensibilité
- Décomposition de Sobol
- Types d'indices et leur interprétation

3 Méthodes de calcul
- Méthode de Sobol, version Saltelli 2002
- Méthode FAST étendue (Saltelli et al. 1999)

4 Comparaison

5 Discussion

6 Références

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
- d'où précaution "de statisticiens" : évaluer la précision, etc.

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

Méthode de Sobol

Principe de la méthode de Sobol, version Saltelli (2002) :

- 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

Méthode *sobol2002* : plan d'expériences

Monte-Carlo

$$A = \begin{pmatrix} x_{A,1}^{(1)} & \cdots & x_{A,1}^{(K)} \\ \cdots & \cdots & \cdots \\ x_{A,N}^{(1)} & \cdots & x_{A,N}^{(K)} \end{pmatrix}$$

Monte-Carlo

$$B = \begin{pmatrix} x_{B,1}^{(1)} & \cdots & x_{B,1}^{(K)} \\ \cdots & \cdots & \cdots \\ x_{B,N}^{(1)} & \cdots & x_{B,N}^{(K)} \end{pmatrix}$$

Combinaisons

$$\rightarrow C_i = \begin{pmatrix} x_{B,1}^{(1)} & \cdots & x_{A,1}^{(i)} & \cdots & x_{B,1}^{(K)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{B,N}^{(1)} & \cdots & x_{A,N}^{(i)} & \cdots & x_{B,N}^{(K)} \end{pmatrix}$$

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

$$\begin{pmatrix} x_{A,1}^{(1)} & \cdots & x_{A,1}^{(i)} & \cdots & x_{A,1}^{(K)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{A,N}^{(1)} & \cdots & x_{A,N}^{(i)} & \cdots & x_{A,N}^{(K)} \end{pmatrix} \quad \begin{pmatrix} y_{A,1} \\ \vdots \\ y_{A,N} \end{pmatrix}$$

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

$$\Rightarrow \quad \widehat{\mathrm{SI}}_i = \frac{y_A' \cdot y_{C_i} - \widehat{f_0}^2}{y_A' \cdot y_A - \widehat{f_0}^2}$$

$\Longrightarrow$ estimation de TSI_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
- $\longrightarrow$ 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)
```

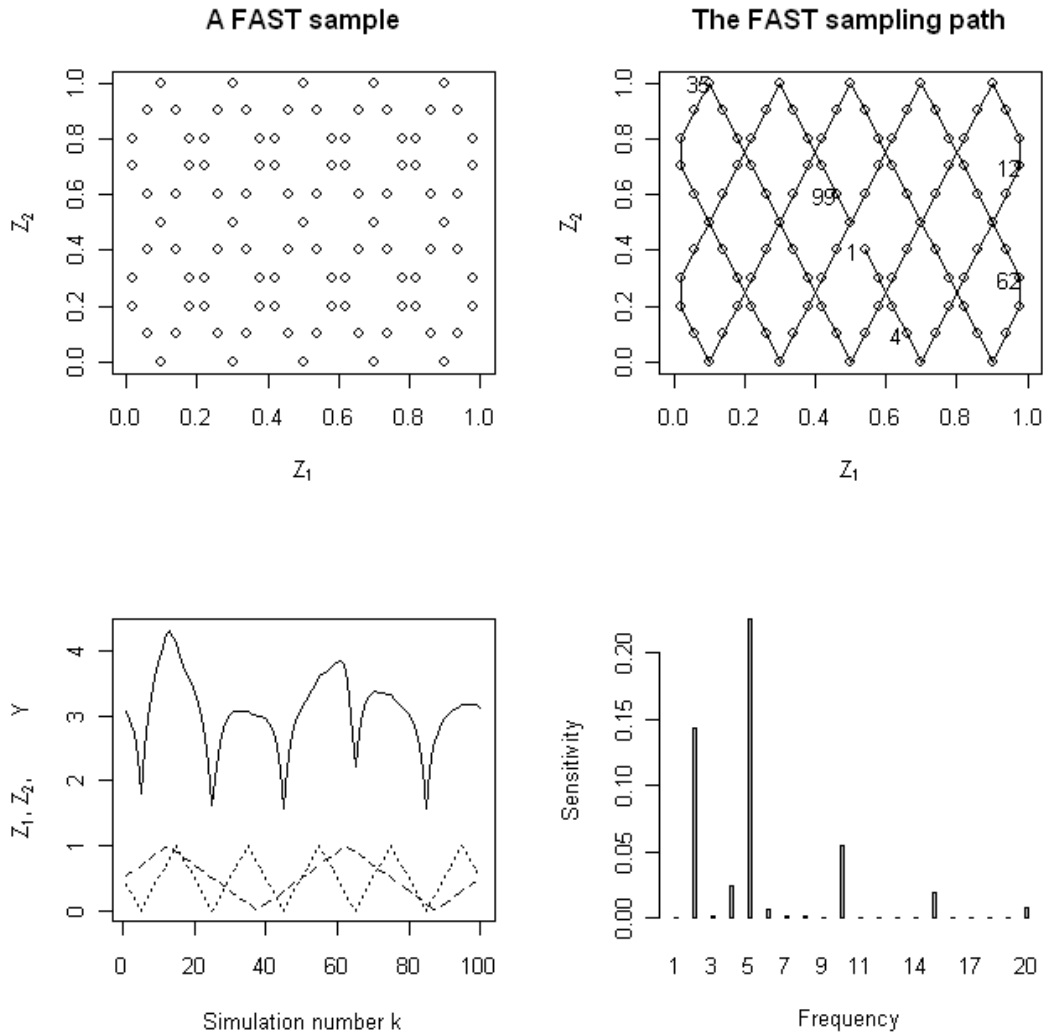
Méthode FAST

Fourier Amplitude Sensitivity Test (FAST) :

- échantillonnage :
  - trajectoire déterministe remplissant l'espace ("space-filling path") :
    $x_j^{(i)} = g_i(\sin \omega_i s_j)$ 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

Variantes :

- FAST étendu (eFAST, Saltelli et al. 1999)
- Random Balance Designs (RBD, Tarantola et al. 2006)



(in Monod et al. 2006)

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 aléatoir 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

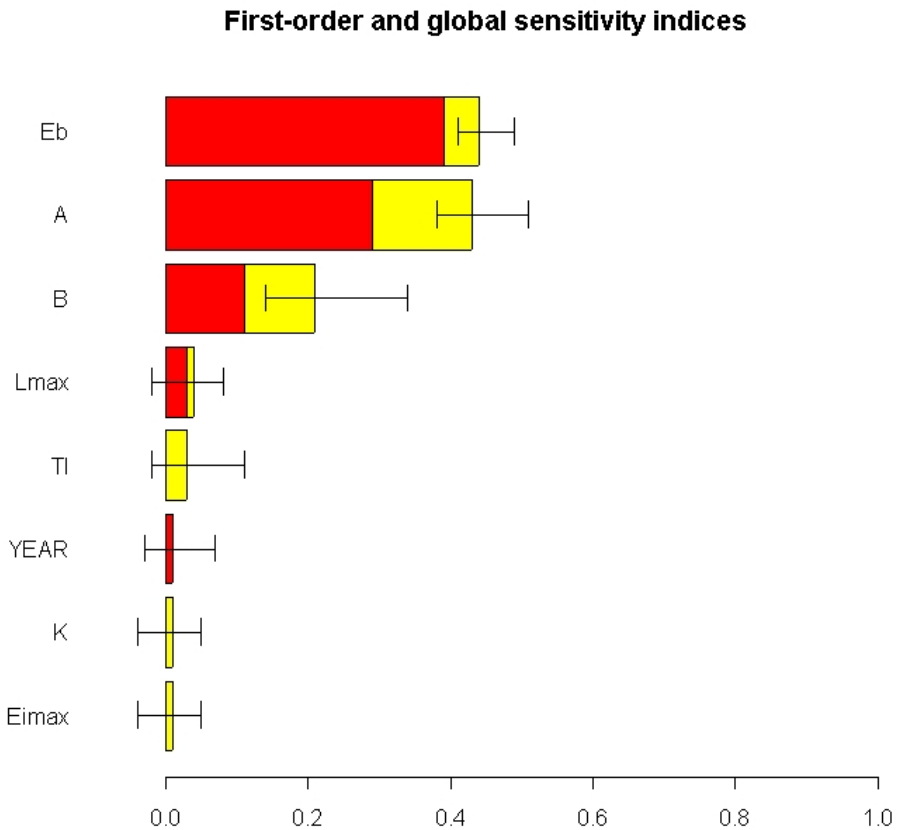
```
# interprétation des résultats
    plot(wwdm.fast99)
```

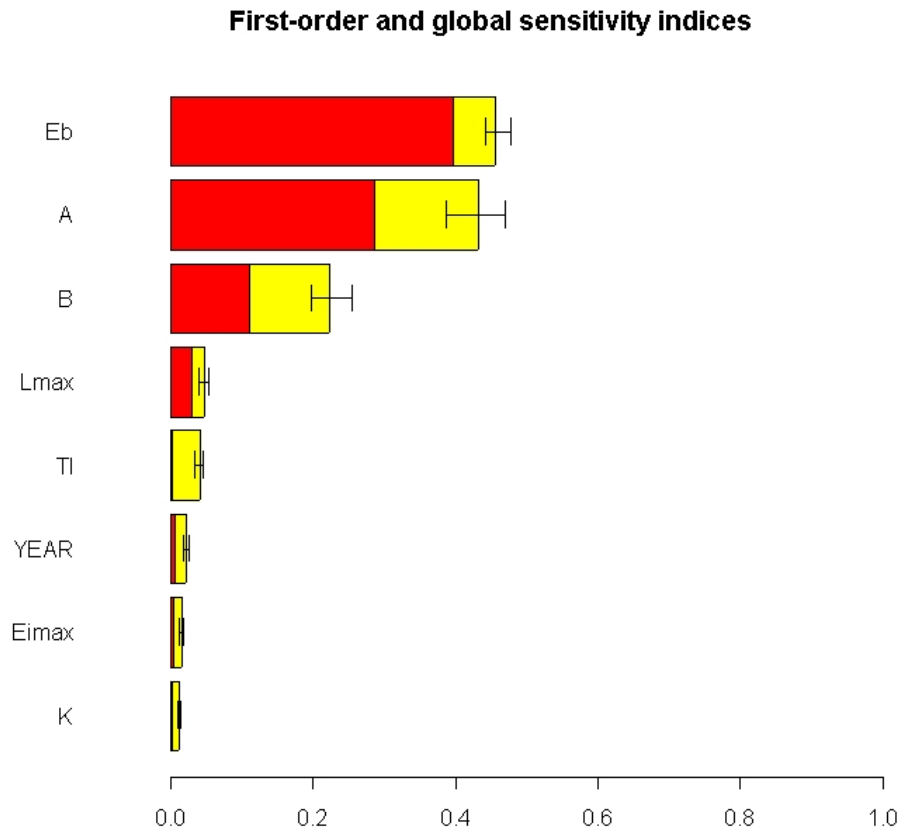
Plan

1. Introduction
2. Définition des indices de sensibilité
   - Décomposition de Sobol
   - Types d'indices et leur interprétation
3. Méthodes de calcul
   - Méthode de Sobol, version Saltelli 2002
   - Méthode FAST étendue (Saltelli et al. 1999)
4. Comparaison
5. Discussion
6. Références

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

1. Introduction
2. Définition des indices de sensibilité
   - Décomposition de Sobol
   - Types d'indices et leur interprétation
3. Méthodes de calcul
   - Méthode de Sobol, version Saltelli 2002
   - Méthode FAST étendue (Saltelli et al. 1999)
4. Comparaison
5. Discussion
6. Références

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

1. Introduction
2. Définition des indices de sensibilité
   - Décomposition de Sobol
   - Types d'indices et leur interprétation
3. Méthodes de calcul
   - Méthode de Sobol, version Saltelli 2002
   - Méthode FAST étendue (Saltelli et al. 1999)
4. Comparaison
5. Discussion
6. Références

- 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.
- 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.
- 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.
- 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.
- 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.

- 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.
- Saltelli A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, **145**, 280-297.
- Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M., Tarantola S. (2008). *Global Sensitivity Analysis. The Primer.* Wiley.
- Sobol, I.M. (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling and Computer Experiments* **1**, 407–414.
- 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.

Licence

Copyrights MEXICO 2009 ©

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation ; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".
see http ://www.gnu.org/licenses/fdl.html