

Synthèse et enjeux de l'exploration numérique

Mai 2015

Ecole Pec'num

Bertrand Iooss (EDF R&D/MRI)

Les incertitudes : problématiques génériques



Étude exploratoire : comprendre un **phénomène**, un processus expérim/indust.

Étude de sûreté : calculer un **risque** de défaillance (marges, événements rares)

Étude de conception : optimiser et maîtriser **les performances** d'un système

Incertitudes

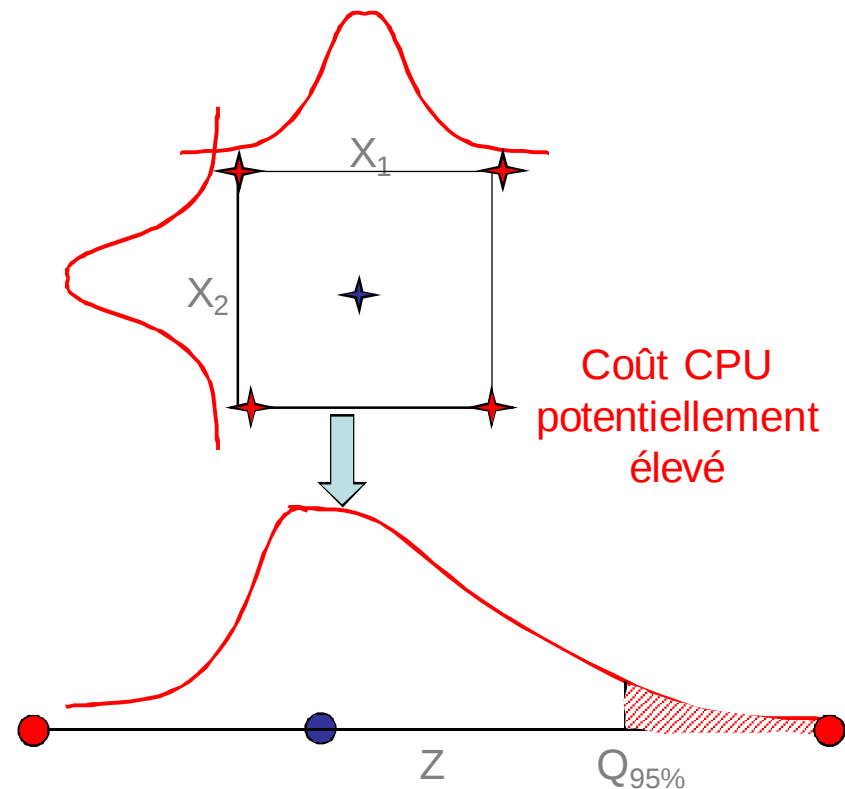
- Variables environnementales
- Paramètres physiques
- Paramètres du processus

Plan d'expériences

Code de calcul
ou
Expérience

Potentiell. métamodèle

- Distributions des réponses
- Probabilité de défaillance
- « Entrées » les plus influentes



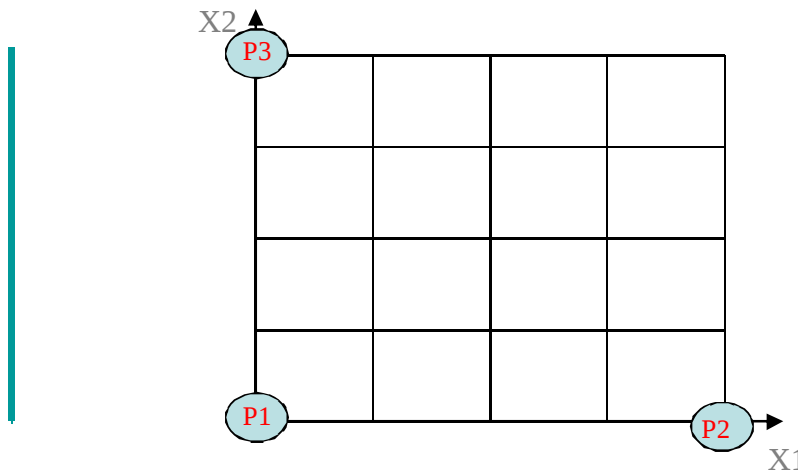
Points clés : - enrichir les démonstrations de sûreté, hiérarchiser les risques
- optimiser coûts / performance (procédés) & disponibilité (install.)

1^{er} outil : les plans d'expériences



Objectif : Placer des points « intelligemment » dans le domaine des entrées $\mathbf{X} \in \mathbb{R}^p$

« Mauvais » plan pour l'exploration :



L'exploration d'un plan "One-at-A-Time" est pauvre
(zones vides, interactions ?,
non monotonie ?)

⇒ **Plans factoriels fractionnaires à 2 niveaux** ($N = 2^{p-r}$ simulations) permet de réduire drastiquement le nombre de simulations

[cf Walter]

Analyse des résultats : estimation des coefficients d'un modèle polynomial

⇒ **Plans dits à bon remplissage de l'espace** (« space filling designs ») ont pour objectif d'explorer « au mieux » l'espace des possibles sur la sortie du modèle $G(\mathbf{X})$

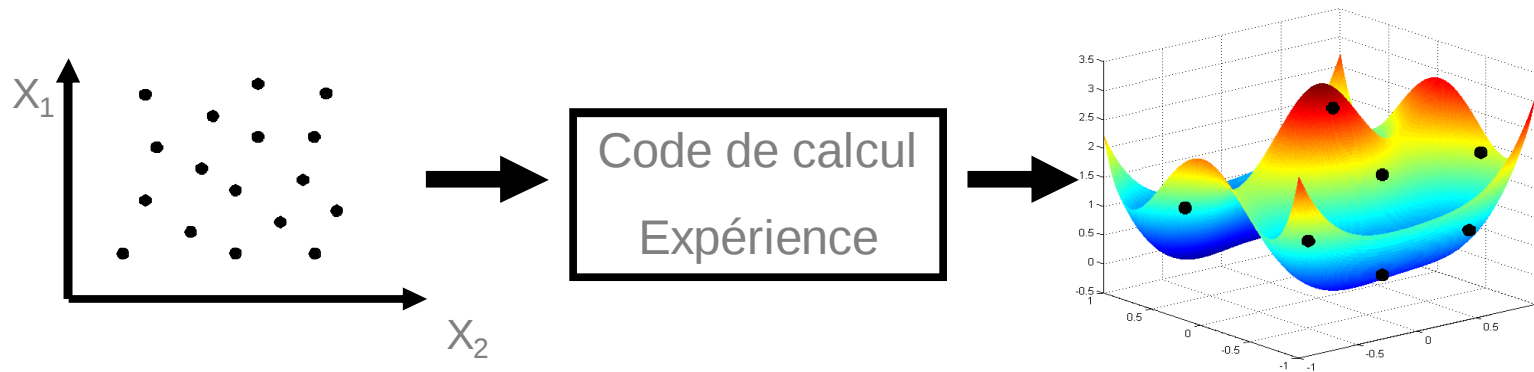
[cf Hervé & Luc]

⇒ Utiles pour l'analyse de sensibilité et la construction de métamodèles, ...

2^{ème} outil : les métamodèles [cf Walter & Sebamandine]

Impossible d'afficher l'image.

- Fonction statistique représentative du code de calcul (polynôme, splines, chaos polynomial, réseau de neurones, krigeage, ...) avec un tps de calcul négligeable
- Approximation à partir de n simulations du code ($n \sim 10 p$)
- Important de contrôler les **qualités d'approximation et de prédiction**

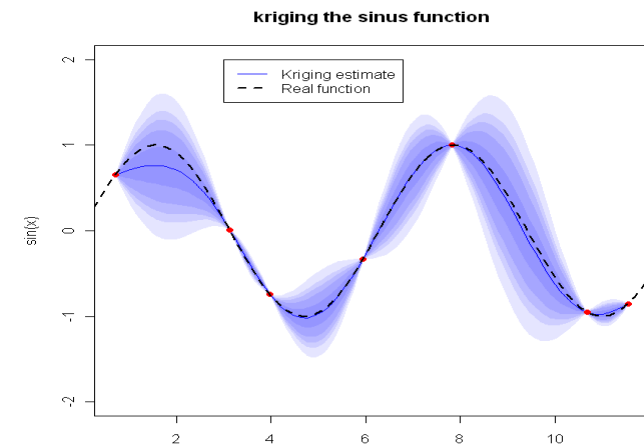


Métamodèle du krigeage

Interpolation par un processus gaussien stationnaire

Avantage : fournit en plus d'une prédiction, une estimation de l'erreur (loi gaussienne en chaque point)

=> Outil algorithmique pour **spécifier des calculs pertinents** (vis-à-vis de la quantité d'intérêt) : **planification adaptative**



Plan

1. Cadre générique pour les études d'incertitudes et de sensibilité
2. Quelques problématiques au-delà du cadre simplifié
3. Exemple en étude de gestion d'actifs

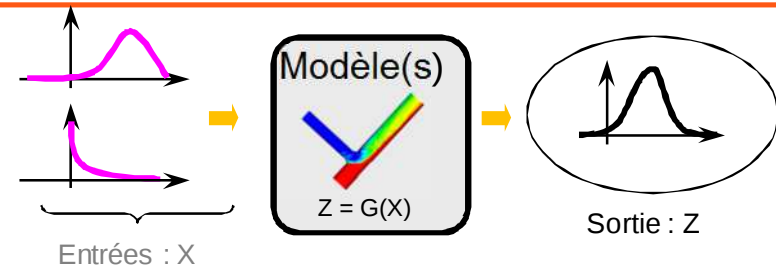
Analyses d'incertitudes et de sensibilité

Impossible d'afficher l'image.

Représentation boîte noire : $Z = G(X)$

Analyse d'incertitude (problématique générale)

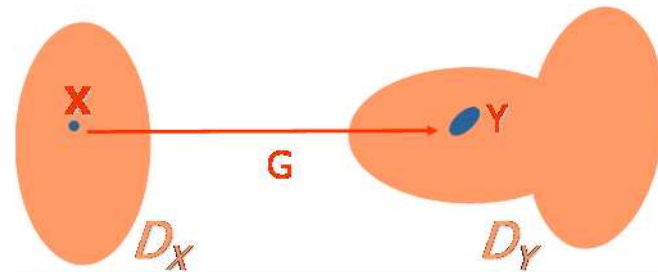
Estimation et étude de la distribution des sorties d'un code induite par les distributions de probabilités des variables d'entrée



Analyse de sensibilité

La sortie Z est-elle plus ou moins variable lorsque l'on fixe une des entrées X_i ?

Quelles variables d'entrée sont les plus influentes ?

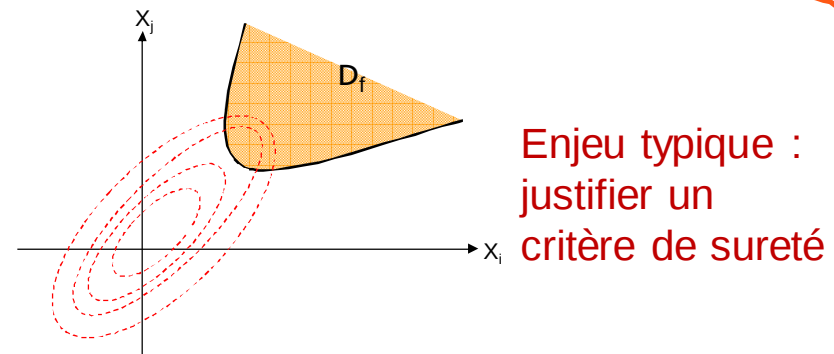


Analyse de fiabilité

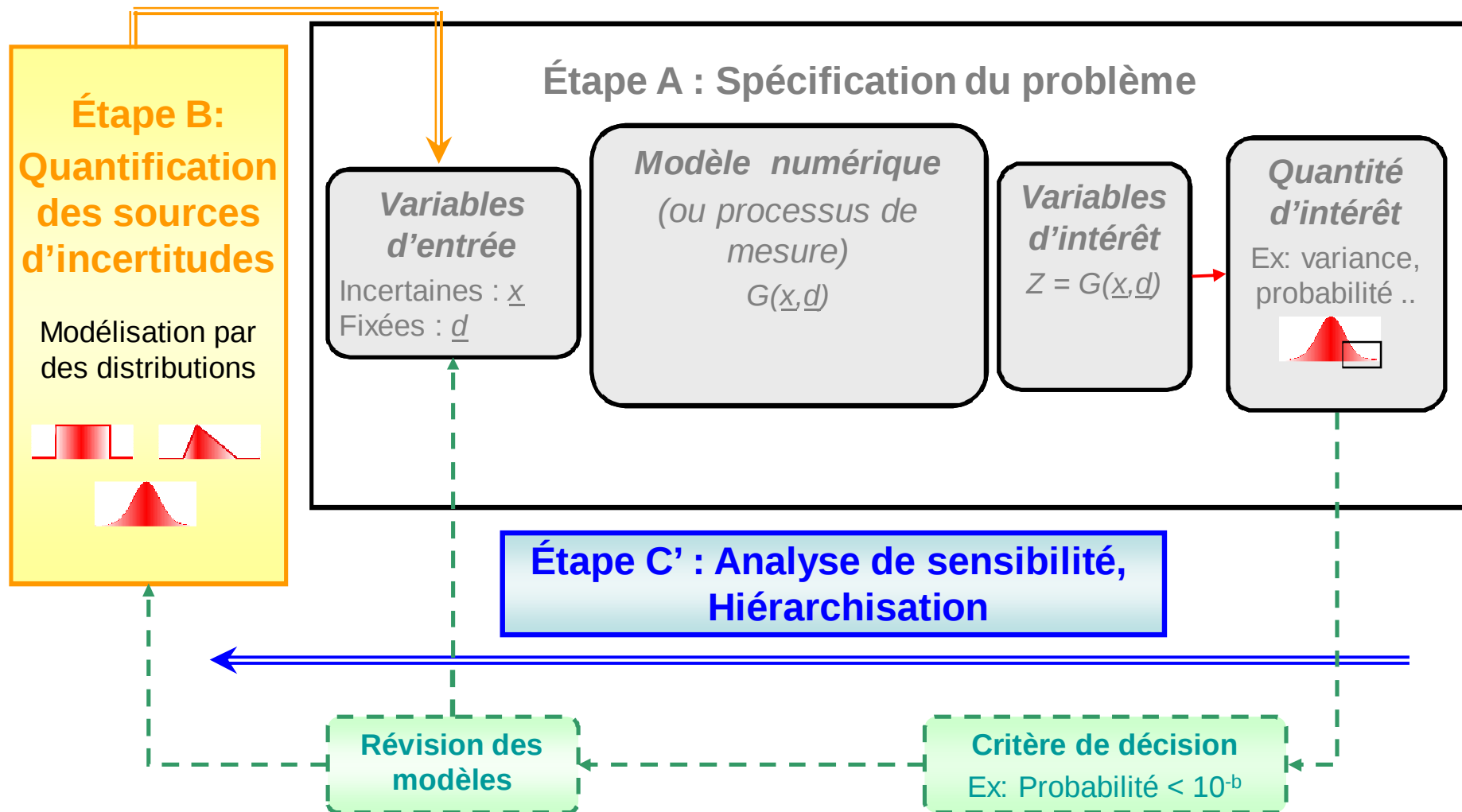
Quelle est la probabilité que la variable Y dépasse un seuil donné ?

$$p_f = \int \mathbb{1}_{\{G(x) \leq 0\}} p(x) dx = \int_{\mathcal{D}_f} p(x) dx$$

Autres problématiques : estimation de quantiles, inversion du domaine de défaillance



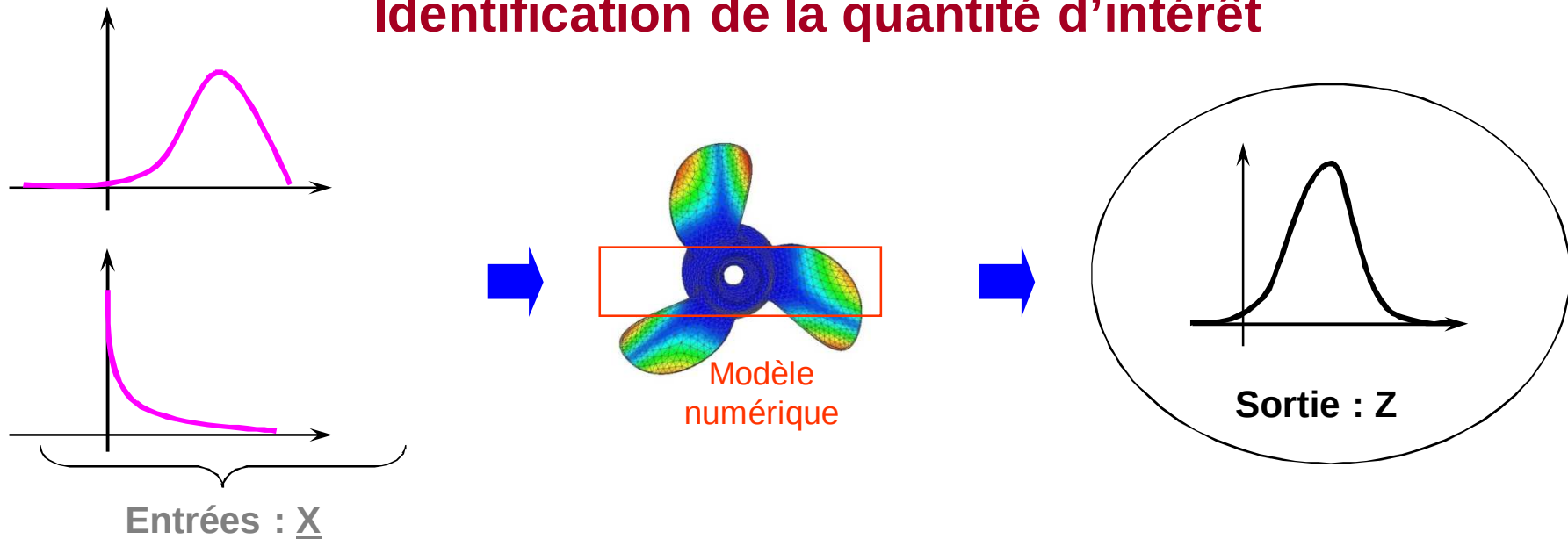
Étape C : Propagation des sources d'incertitude



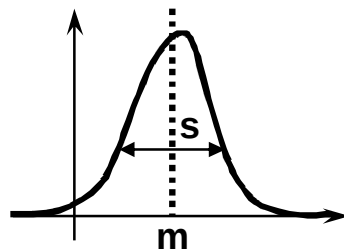
A : Spécification du problème

Impossible d'afficher l'image.

Identification de la quantité d'intérêt

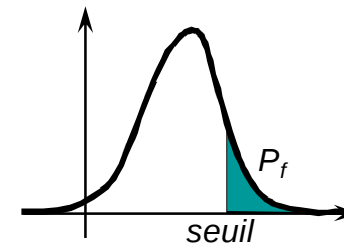


Qu'est ce qui est (vraiment) intéressant pour notre étude ?



Moyenne, médiane, variance, (moments)
de Z

*Contexte d'une étude de conception /
compréhension d'un système*



Quantiles (extrêmes), probabilité de
dépassement d'un seuil fixé,
distribution complète

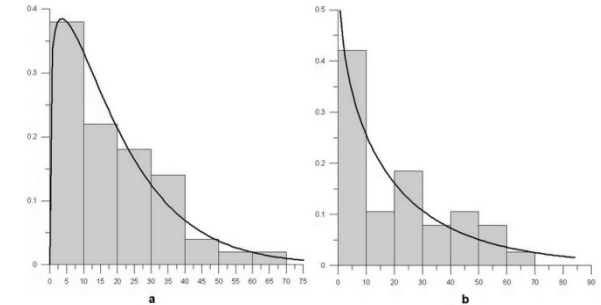
Contexte d'une étude de sûreté

B : Modélisation statistique des sources d'incertitudes

Quantification de la loi jointe nD ...

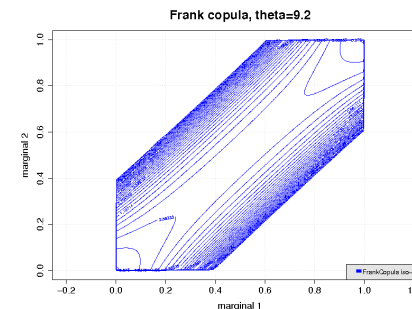
1. Détermination des lois marginales 1D

- Beaucoup de données (typiquement > 30) :
 - ✓ Ajustement de distributions paramétriques ou non
 - ✓ Tests d'adéquation
- Peu de données disponibles => hypothèses, bibliographie, principe du Maximum d'Entropie, Jugement d'expert, modélisation bayésienne

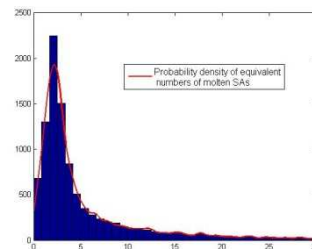


2. Modélisation de la dépendance

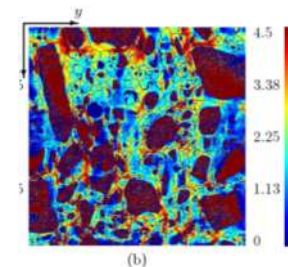
- Utilisation d'une copule
- Validation



3. Modélisation des extrêmes



4. Modélisation de l'hétérogénéité temporelle et/ou spatiale



C : Propagation des incertitudes : Méthodes déterministes vs. stochastiques

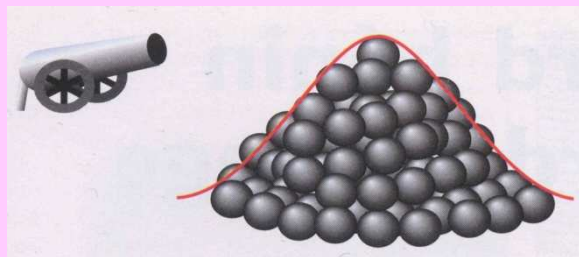
$$Z = G(\mathbf{X}), \mathbf{X} \in \mathbb{R}^p, Z \in \mathbb{R}$$

Approche déterministe (dérivées, adjoint, sensibilités locales)

$$Z(\mathbf{X}) = Z(\mathbf{X}^0) + \sum_{i=1}^p \left(\frac{\partial Z}{\partial X_i} \right)_{\mathbf{X}^0} (X_i - X_i^0) \quad (\text{hypothèse de linéarité})$$

Approche stochastique (échantillonnage Monte-Carlo)

Propagation



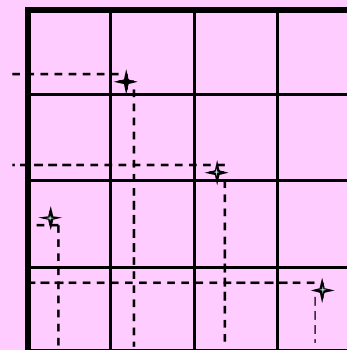
Méthode relativement simple et souple (considère le code en boîte noire)

⇒ coût élevé pour probabilités faibles

⇒ Méthodes probabilistes et plans d'expérience spécifiques pour réduire ce coût

Exemple :

L'hypercube latin



$p = 2$
 $N = 4$

Fiabilité : estimation d'événements rares $Z \leq 0$

Impossible d'afficher l'image.

Objectifs : propagation d'incertitudes permettant de fournir une estimation :

- d'une probabilité de défaillance

$$p_f = \int_{\mathcal{D}_f} f(x) dx = \int_{\mathcal{X}} I_{\mathcal{D}_f}(x) f(x) dx = \mathbb{E} [I_{\mathcal{D}_f}(X)]$$

- d'un quantile z_α tel que $P(Z < z_\alpha) = \alpha$

- d'un domaine de défaillance $\mathcal{D}_f = \{x \in \mathcal{X} : G(x) = z \leq 0\}$

Les estimations doivent être :

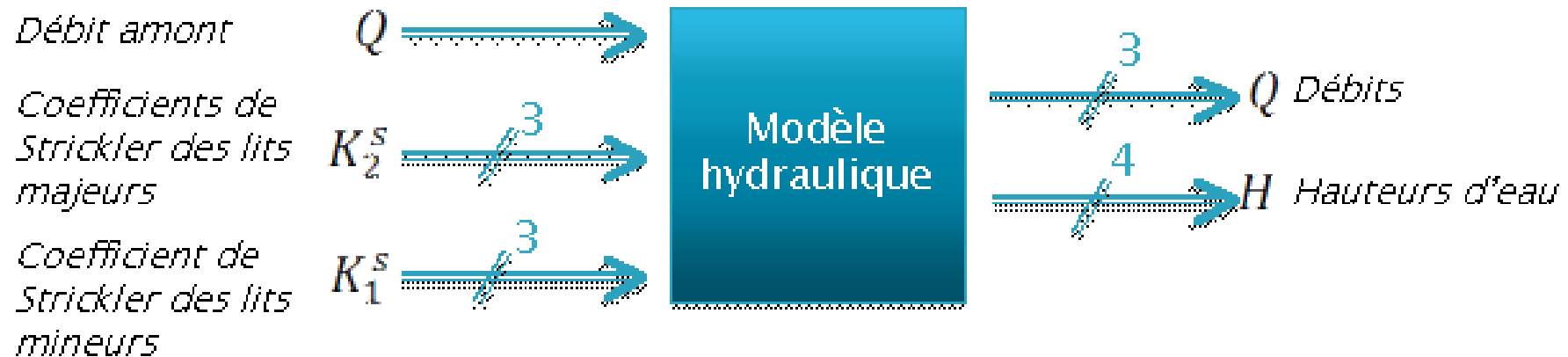
- **Conservatives (ne pas sous-estimer $P_{\text{déf}}$ ou z_α)** => ex : Wilks
- **Parcimonieuses (limiter le nombre d'appels au code G)**
- **Précises (de variance contrôlable ou fixée)**
- **Robustes (à de légères modifications de la loi de la v.a. X ou de G)**

Méthodes efficaces :

- Méthodes de réduction de la variance de Monte Carlo
- Utilisation de **métamodèles** (attention au nouveau terme d'erreur)

Exemple : estimation d'une probabilité d'inondation

- ▶ Estimer le risque que la hauteur d'eau d'un fleuve dépasse la hauteur d'un ouvrage de protection (en aval)
- ▶ Modèle 1D MASCARET (EDF R&D/LNHE) : équations de Saint Venant



- ▶ On cherche à estimer:

$$P_f = \mathbb{P}(H(X) > T)$$

Lois a priori sur les facteurs

- Coefficients de Strickler des deux premiers tronçons fixés :

$$K_{1,1}^S = 50, \quad K_{1,2}^S = 10$$

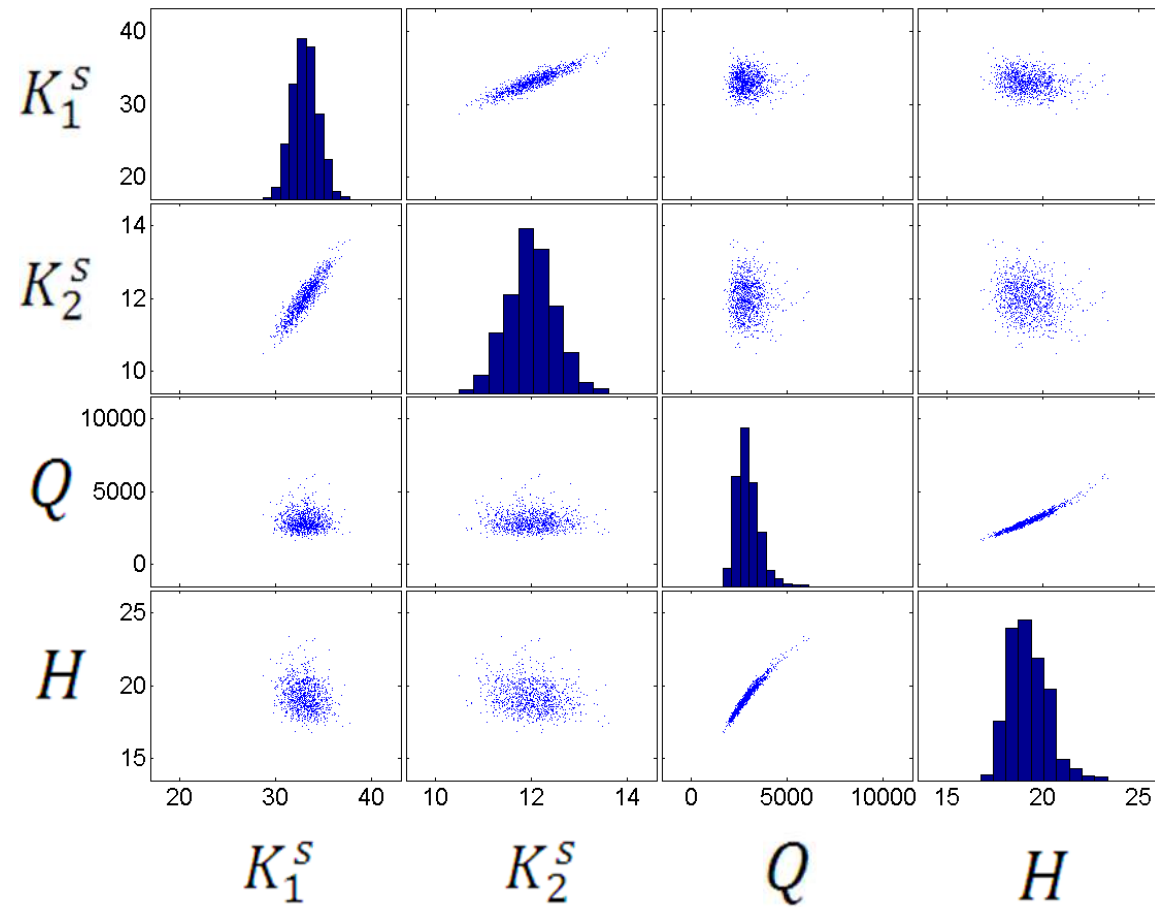
$$K_{2,1}^S = 32, \quad K_{2,2}^S = 12$$

- Loi multivariée $\mathcal{N}(\mu, K)$ pour le 3^{ème} tronçon

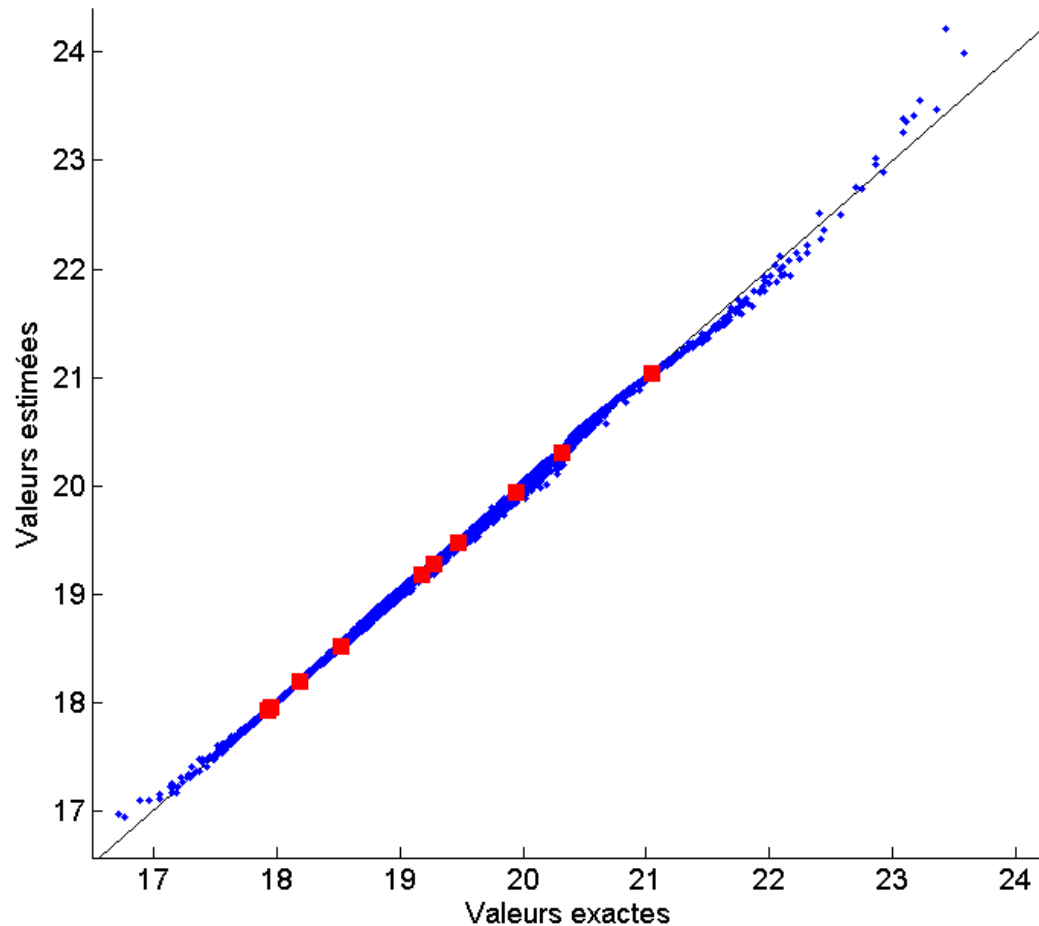
$$\mu = \begin{pmatrix} 33 \\ 12 \end{pmatrix} \quad K = \begin{pmatrix} 1.36 & 0.6 \\ 0.6 & 0.49 \end{pmatrix}$$

- Pour Q : loi de Gumbel de paramètres (2700,470)
- Ces lois sont utilisées pour générer $\mathbb{X}_N$
- Hauteur d'eau à Marmande et $T = 21.34$ m

Corrélations des variables

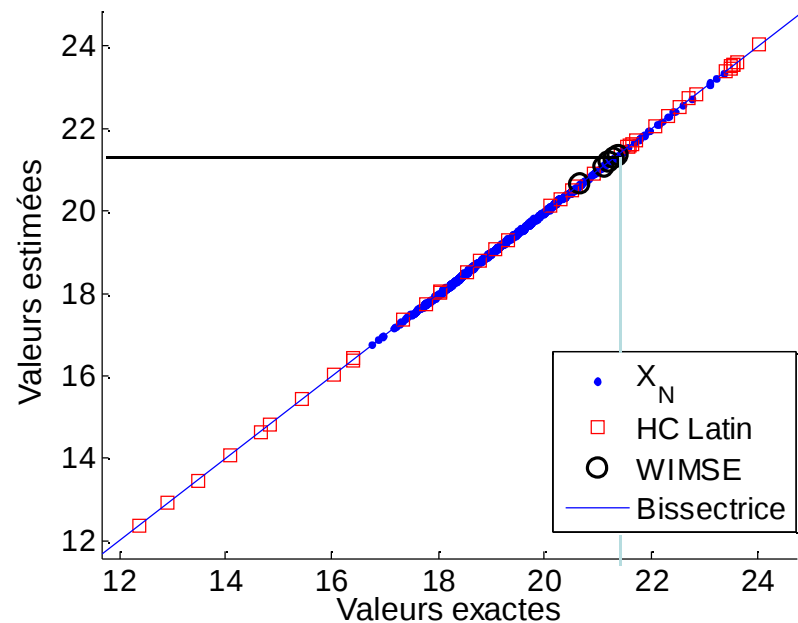
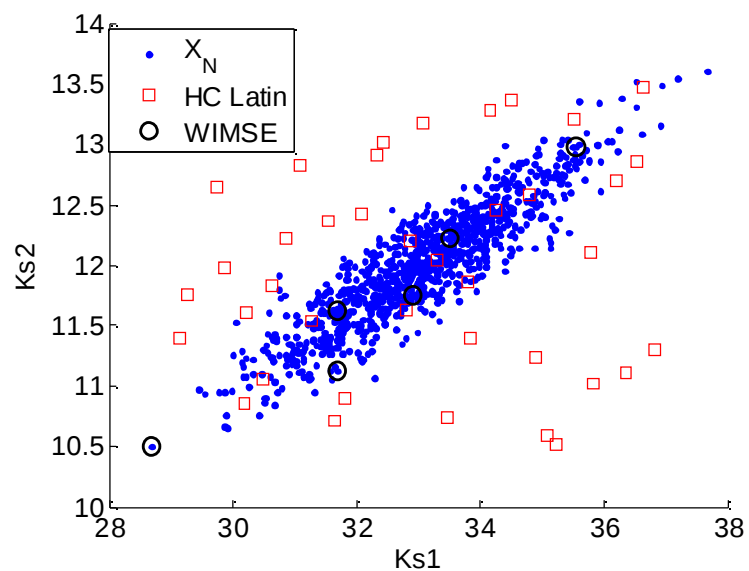
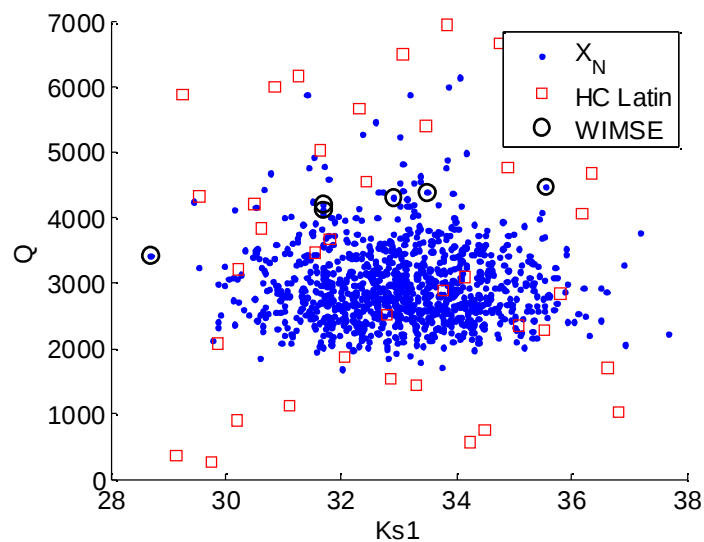
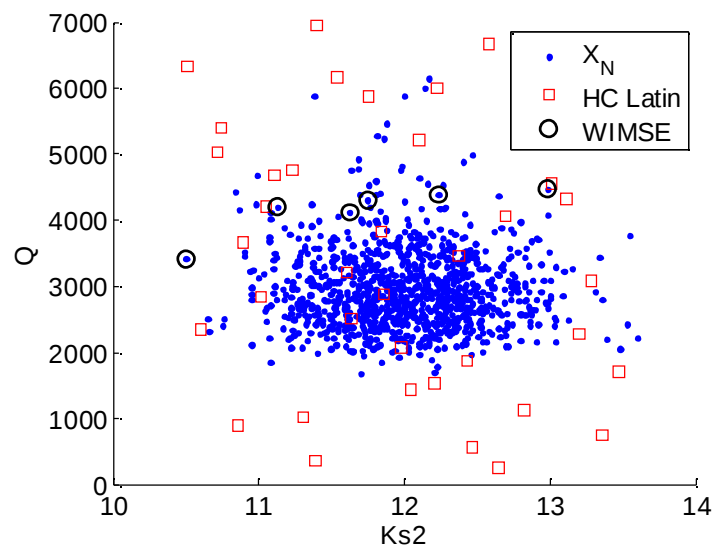


Performances de l'estimation à partir de 10 évaluations

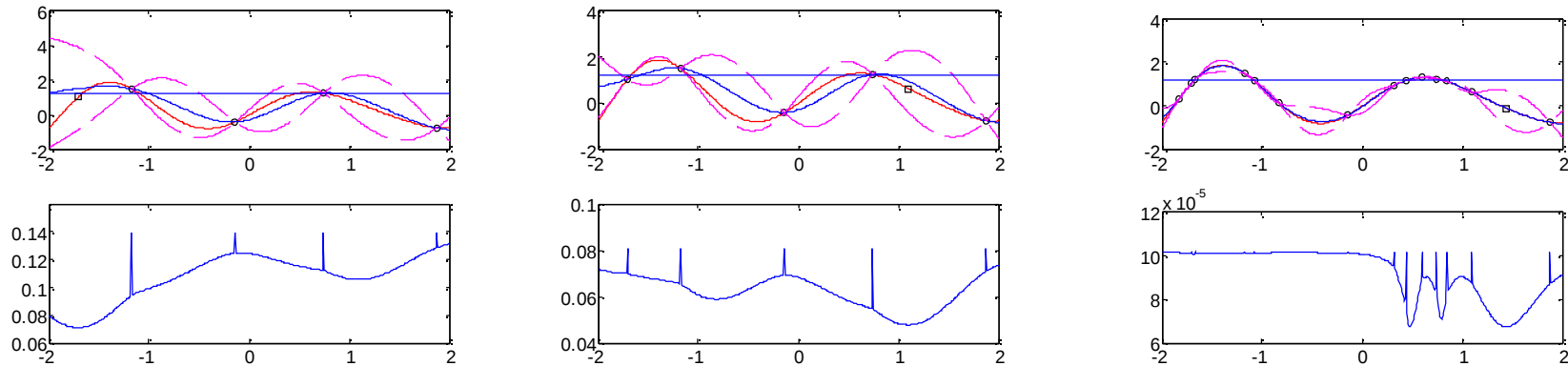


- Fonction de covariance estimée *par ailleurs* avec 40 points (Hypercube Latin)
- N=1000 points dans la base de test

Résultats de la planification



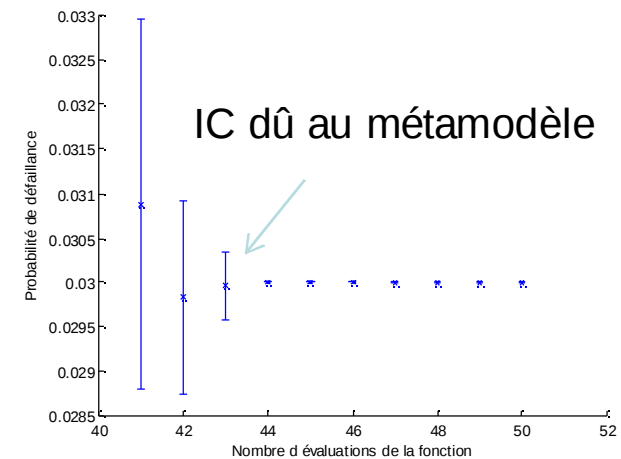
RÉSULTATS



► Résultats

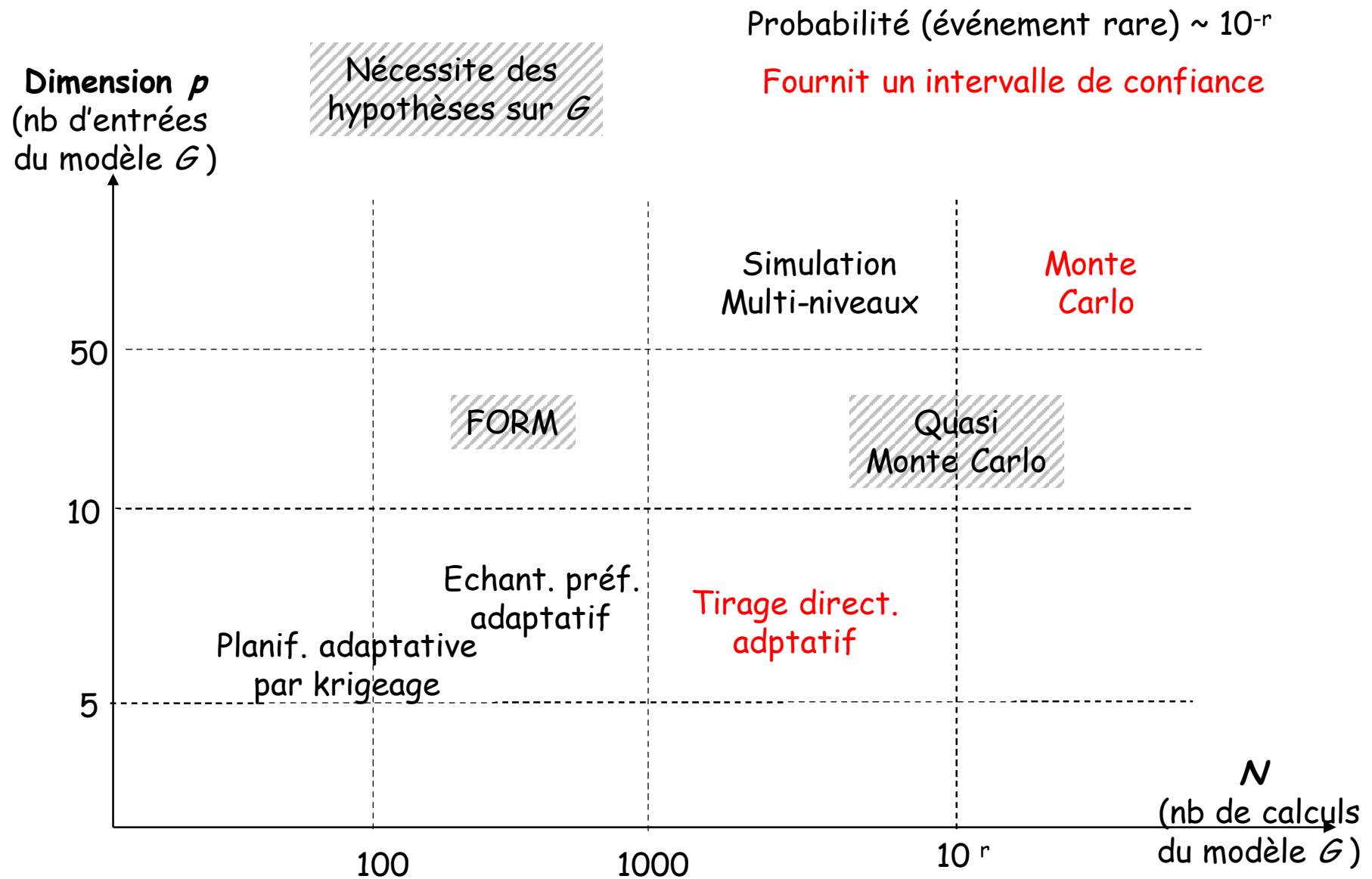
	Monte Carlo simple	Méthode exposée
Nombre de points évalués	$N = 1000$	$n = 44$
Valeur estimée	0.030	0.030
Intervalle de confiance	[0.0192, 0.0408]	[0.0192, 0.0408]

IC dû au MC

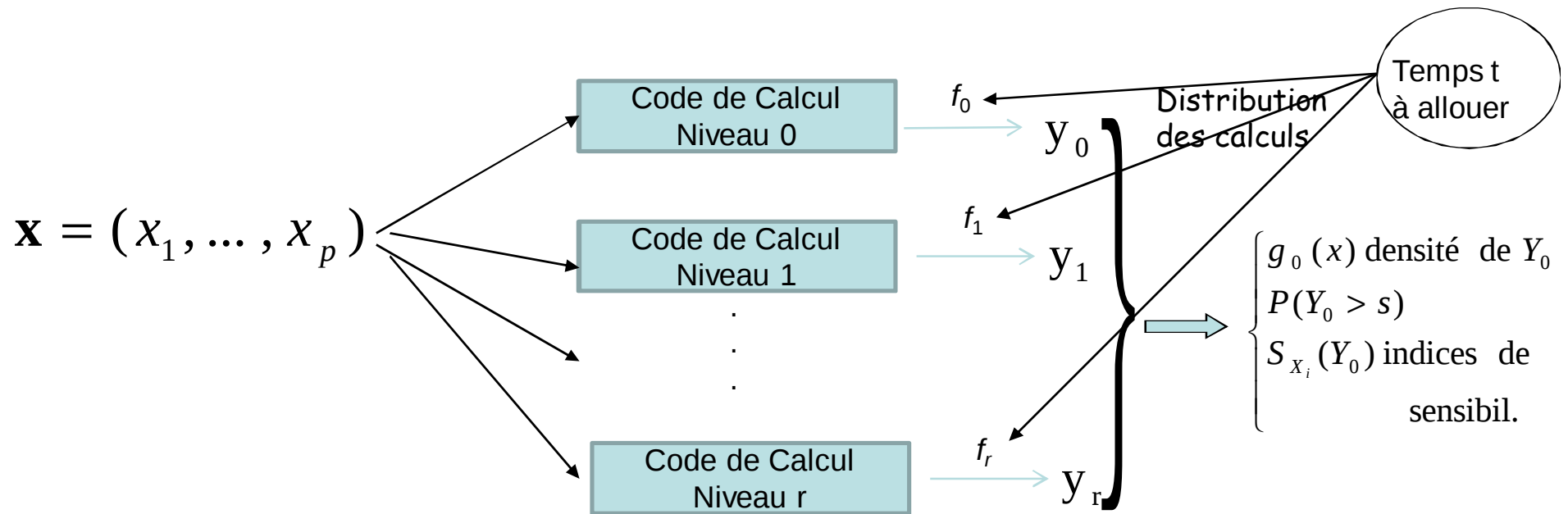


- Méthode qui donne de très bon résultats, mais qui peut s'avérer fausse si la covariance du processus gaussien est mal estimée

Classification des méthodes d'événements rares



Vaincre le coût cpu : simulateurs multi-fidélités



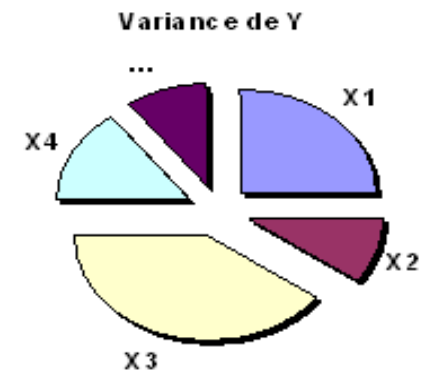
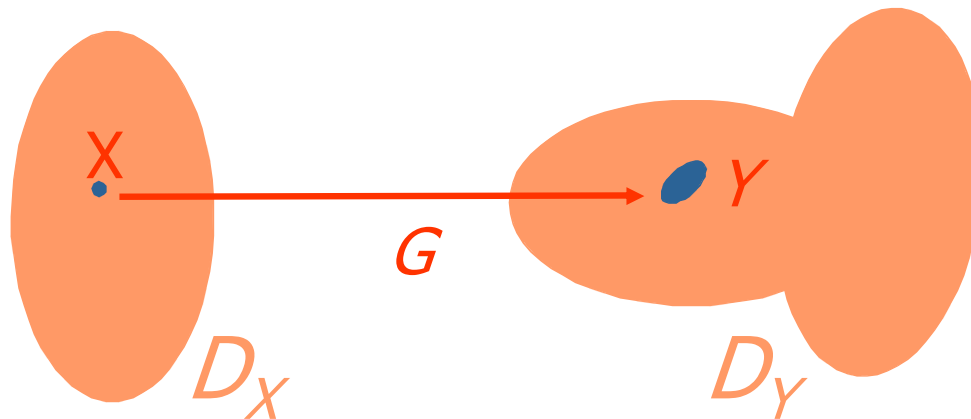
=> Stratégies optimales d' allocation de calculs
 (par exemple en utilisant le cadre du cokrigage)

C' : Analyse de sensibilité

[cf Clémentine]

Objectifs :

- Hiérarchiser les sources d'incertitudes afin de réduire la variabilité de la sortie

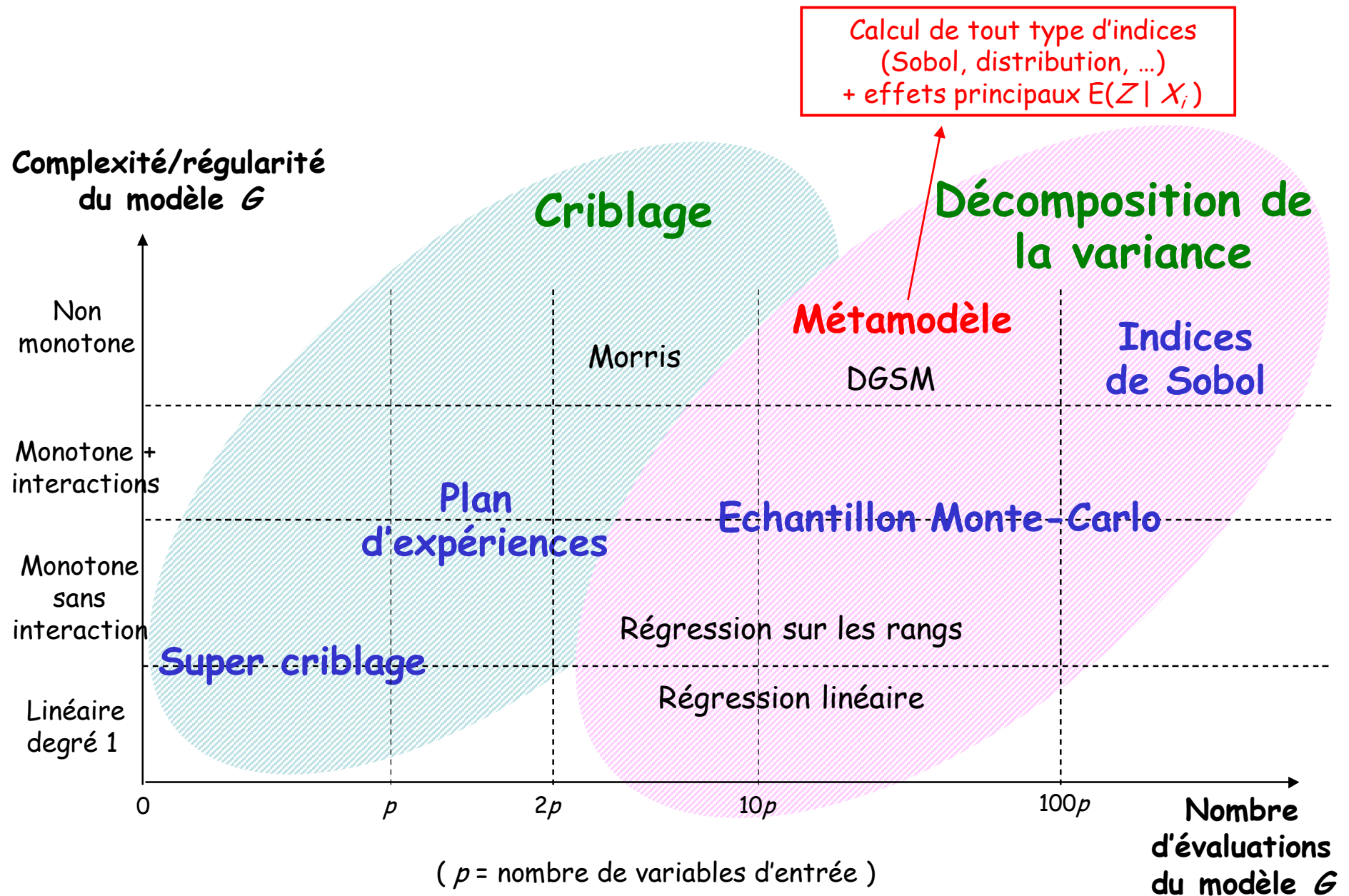


- Simplification du modèle

- **Déterminer les variables aléatoires non-influentes**, qui peuvent être considérées déterministes sans influence la variabilité de la sortie (méthode de Morris, calculs des dérivées)
- Construire un modèle simplifié, un métamodèle

Classification des méthodes d'analyses de sensibilité

Impossible d'afficher l'image.



Needs some *a priori* knowledge

Main effects (1st order)

Main and total effects

All effects (at all order)

visible
per l'image.

Visualization of main effects

***p*-independent** ($N \sim 100$)

Moment-independent

Smoothing

Sobol RBD-RLHS

Statistical tests

Complexity/Regularity
of the model G

Screening

**Variance
decomposition**

Non
monotonic
discontinuous

Non
monotonic
continuous

Monotonic
with
interactions

Monotonic
without
interaction

Linear 1st
degree

Morris

Classification
+metamodel

Quasi-MC

Monte Carlo

DGSM

Metamodel

FAST
eEFAST

Fract. fact.
RIII design

RIV
design

Sequential
bifurcations

Screening by groups

Rank regression

Linear regression

Supersaturated
design

One-At-a-Time

Number of model
evaluations

0

$p/2$

p

$2p$

$10p$

$100p$

$1000p$

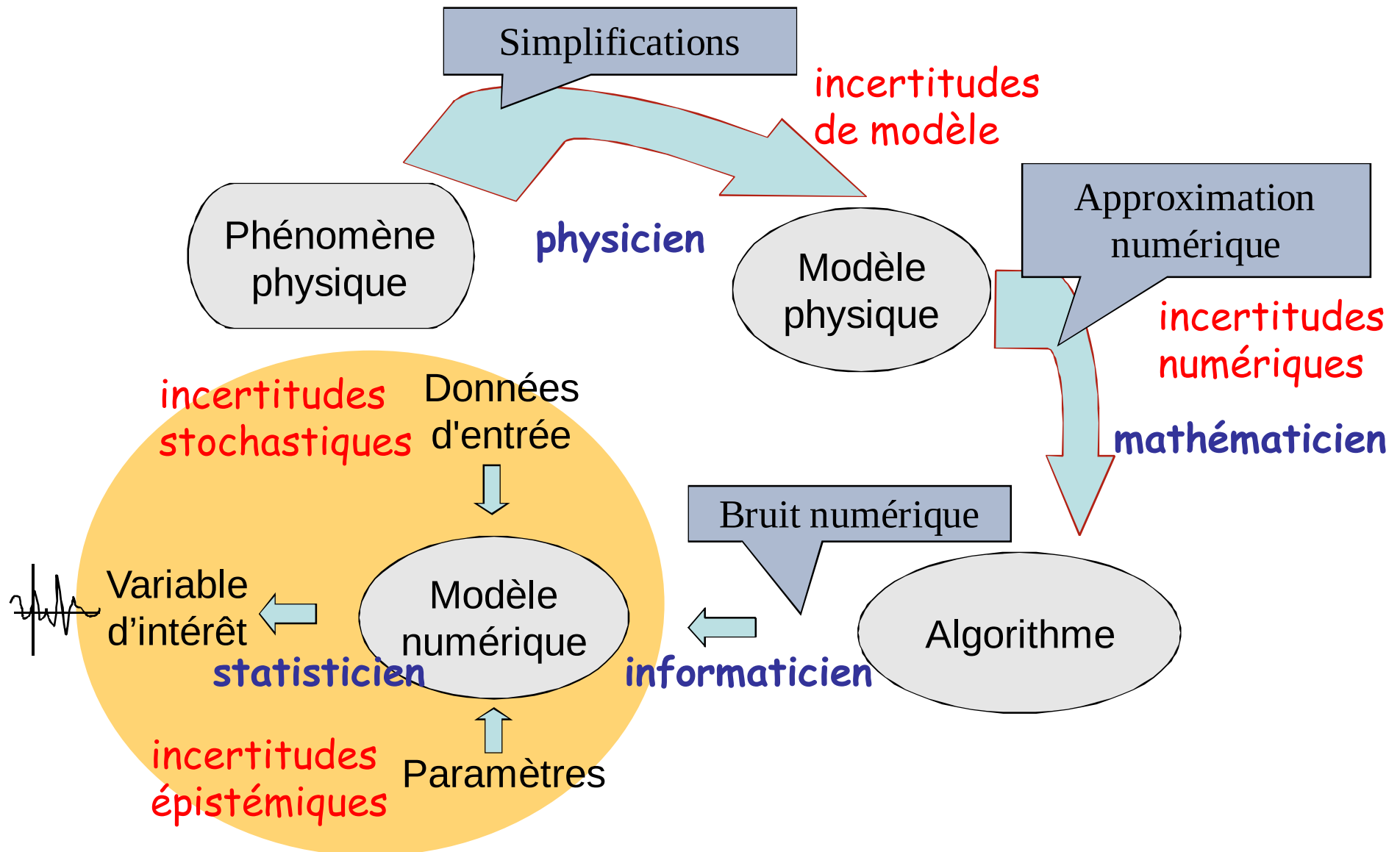
N

Plan

1. Cadre génériques pour les études d'incertitudes et de sensibilité
2. Quelques problématiques au-delà du cadre simplifié
3. Exemple en étude de gestion d'actifs

Des incertitudes dans toute la chaîne de modélisation

Problème majeur : fiabilité des prédictions



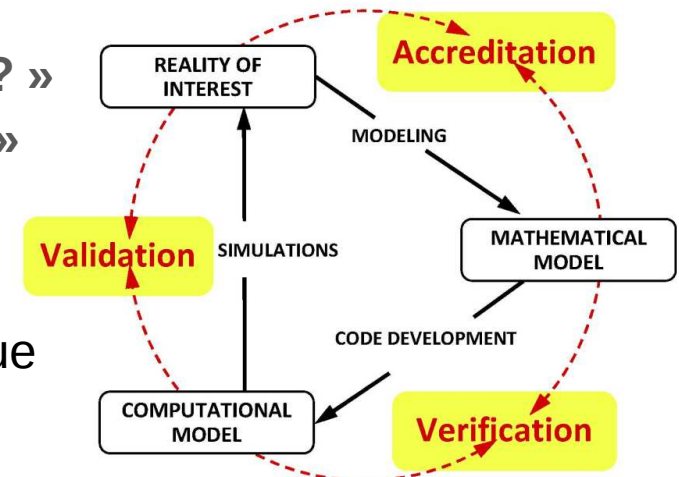
Incertitudes de modèle & validation

Approche V&V

- **Vérification** : « Do I solve the equations right ? »
- **Validation** : « Do I solve the right equations ? »

Apport des méthodes statistiques

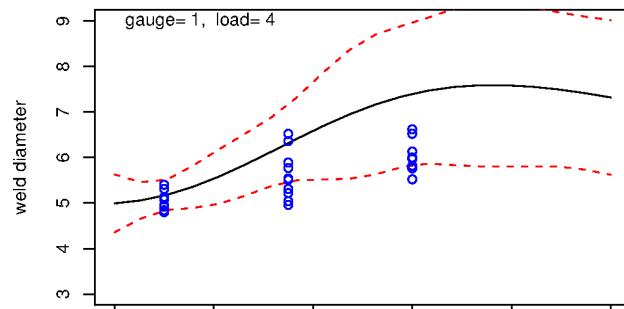
- Mise en place de biais forfaitaires,
- Détection de résultats aberrants par analyse statistique
- Inférence bayésienne



=> Evaluer l'erreur de modélisation par rapport aux incertitudes paramétriques

calibration

$$Y_{\text{réel}} = G(\mathbf{X}, \theta^*) + \varepsilon$$

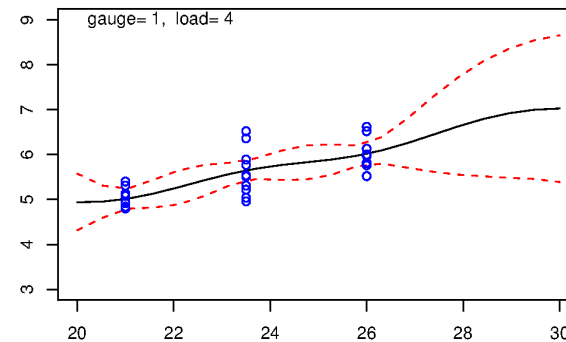


Pure-model predictions

calibration & validation

$$Y_{\text{réel}} = G(\mathbf{X}, \theta^*) + b(\mathbf{X}) + \varepsilon$$

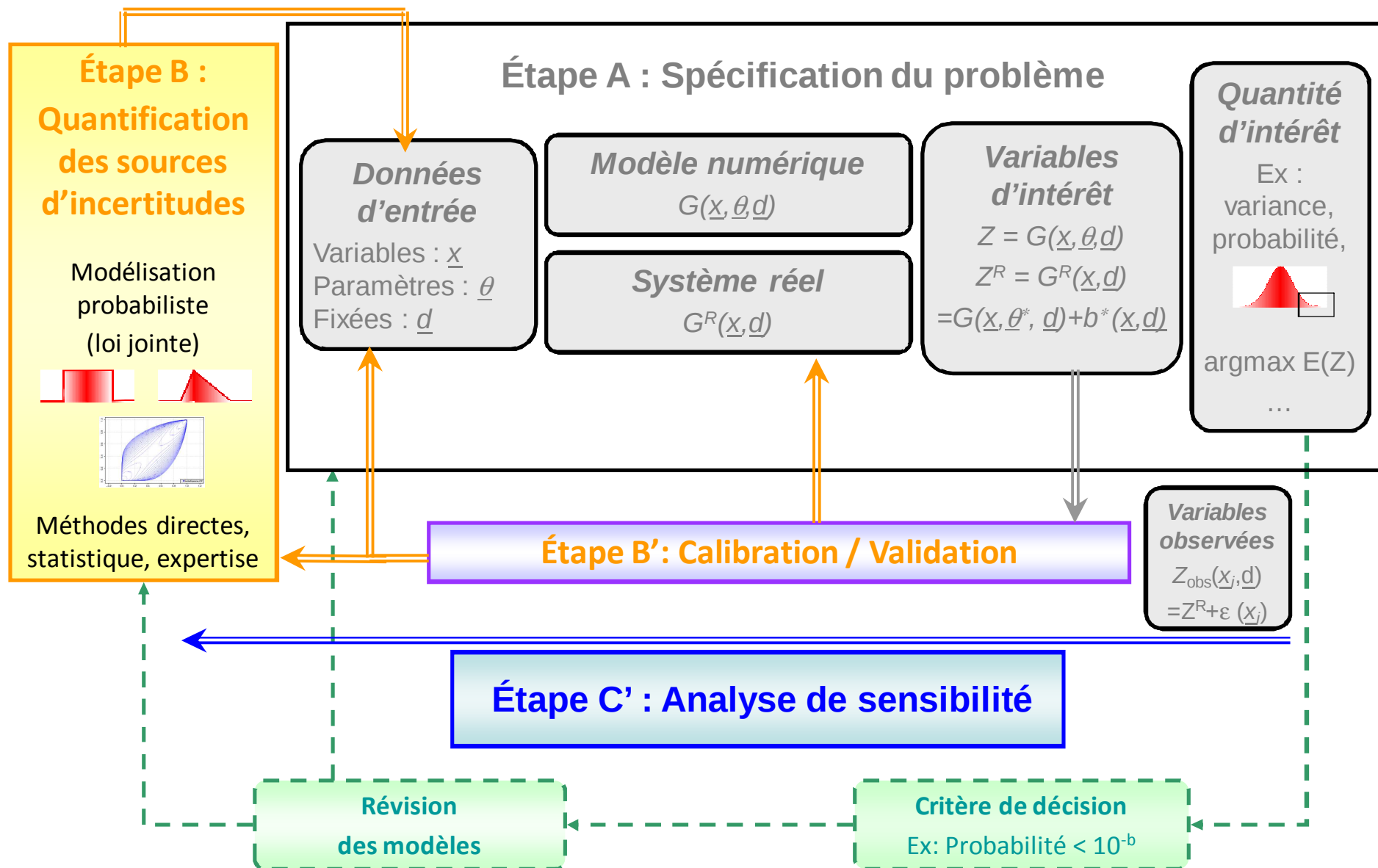
! Problème d'identifiabilité !



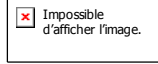
Bias-corrected predictions

Étape C : Propagation des sources d'incertitude

Impossible d'afficher l'image.



Problématiques d'optimisation ou d'inversion



[cf Victor]

Optimisation : $x^* = \arg \min_{x \in D} G(x)$

Optimisation robuste : $x^* = \arg \min_{x \in D} E_U [G(x, U)]$

où U est un vecteur de variables incertaines que l'on peut pas contrôler

- Autres formulations :
- optimisation multi-critères (+ $\arg \min \text{Variance}_U (G)$)
 - optimisation sous contrainte (+ $\text{Variance}_U (G) < \text{seuil}$)
 - optimisation fiabiliste (+ $P[H(x, U) < 0] < \text{seuil}$)

Inversion : $D = \{x \in \mathcal{X} : G(x) \leq \text{seuil}\}$

Inversion robuste : $D = \{x \in \mathcal{X} : \forall U, G(x, U) \leq \text{seuil}\}$

Autres formulations : inversion sous contrainte, inversion fiabiliste

Le problème est particulièrement important dans le contexte « robuste » pour éviter une boucle Monte Carlo à chaque étape d'optimisation

=> Algorithmes de planification adaptative de calculs

Grandes problématiques en analyse de sensibilité

P0a) $G(.)$ complexe (interactions, non monotone, discontinu, ...)

⇒ Sobol (variance), entropie, Borgonovo (densité)

P0b) $G(.)$ coûteux (nb de simulations : $N \ll 1000$)

⇒ criblage, plans parcimonieux, métamodèles

P1) Z est de grande dimension ou à support temporel, spatial, ...

[cf Nathalie]

⇒ agrégation, décompositions fonctionnelles, outils de visualisation

P2) X est de grande dimension ou fonctionnel ($p \gg 10$)

[cf Nathalie]

⇒ criblage, indices de groupes, plans parcimonieux, exploitation des dérivées

P3) La **quantité d'intérêt est liée aux événements rares** (proba de déf., quantile)

⇒ perturbations de loi des X , fcts de contraste différentes

P4) Dépendance entre entrées

P5) $G(.)$ est stochastique

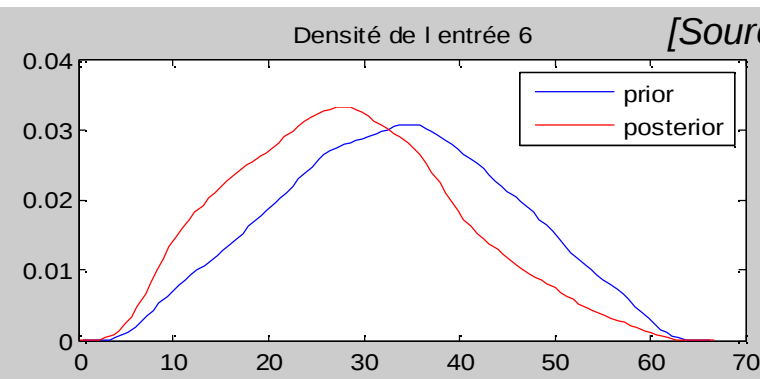
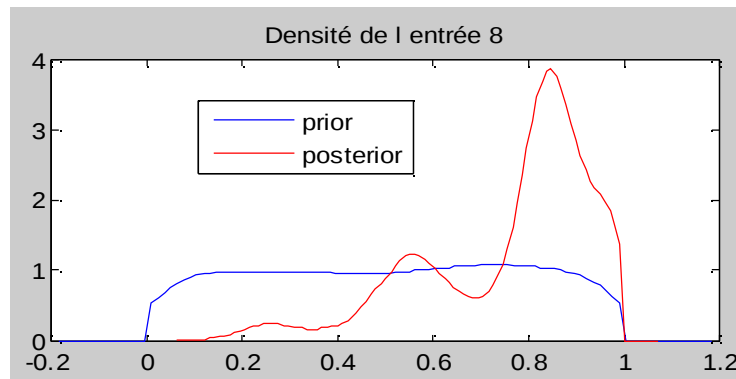
Analyse de sensibilité pour les calculs fiabilistes



Problème : disposer d'une méthode d'estimation de $P = \int 1_{G(\mathbf{X}) \geq 0} f(\mathbf{X}) d\mathbf{X}$ associée à des indices de sensibilité des entrées sur P

Quelques méthodes :

- FORM et facteurs d'importance : hyp. limitatives, espace standard, peu robuste
- Indices de Sobol sur l'indicatrice $1_{G(\mathbf{X}) \geq 0}$
=> pas terrible : estimation coûteuse, peu d'infos (influence interactions)
- Indices de Sobol sur $P_f(Q)$ par rapport aux Q , paramètres des lois des entrées
=> il faut fixer des lois sur les paramètres des lois des X_i ☹
- Comparer les lois des entrées et les lois des entrées sachant défaillance



[Source : Marrel]

Méthode coûteuse => utilisation du subset simulation, métamodèle, ...

PLI : Perturbation Law-based sensitivity Indices



- Soit $\mu_{i\delta}$ une loi perturbée d'un X_i (perturbation δ de moy., var., ...), telle que minimisant (sous contrainte) sa divergence Kullback-Leibler avec la loi initiale
- **Indices de sensibilité :** [Lemaître et al. 2015]

$$S_{i\delta} = \left(\frac{P_{i\delta}}{P} - 1 \right) 1_{P_{i\delta} \geq P} + \left(1 - \frac{P}{P_{i\delta}} \right) 1_{P_{i\delta} < P}, \quad P_{i\delta} = \int 1_{f(\mathbf{X}) \geq 0} \frac{\mu_{i\delta}(X_i)}{\mu_i(X_i)} \mu_{\mathbf{X}} d\mathbf{X}$$

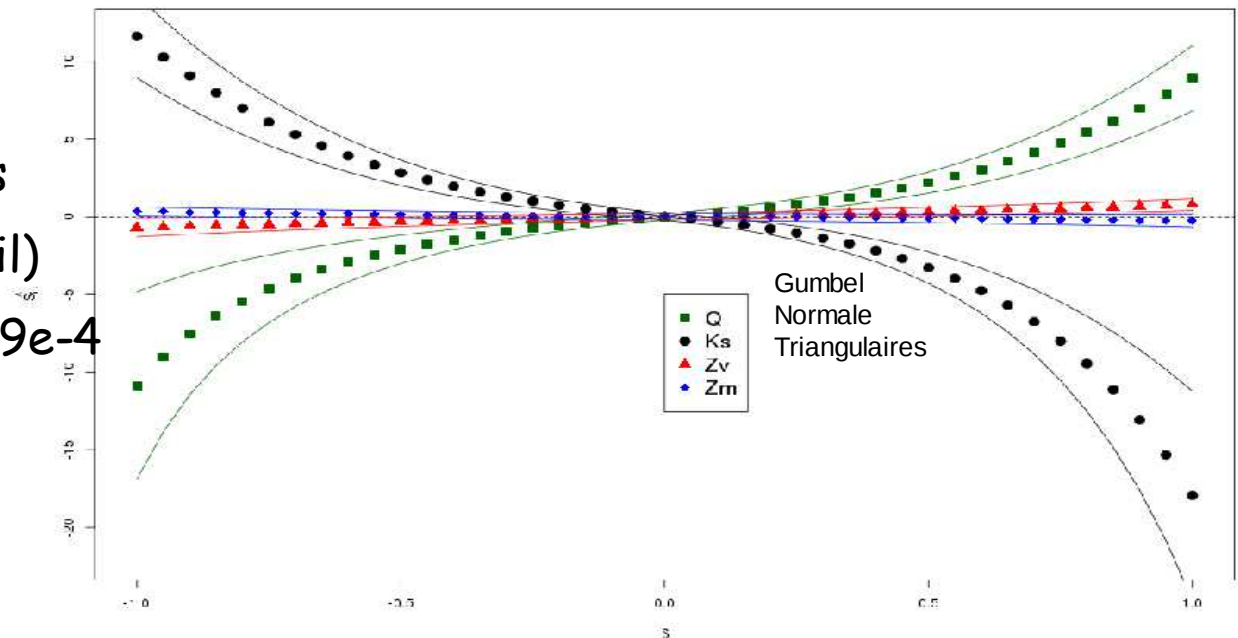
- P et $P_{i\delta}$ sont estimés avec le même échantillon (par exemple Monte Carlo ou tirage d'importance) $\Rightarrow$ pas de nouveaux appels au code

Exemple : Modèle de crues

$\rightarrow$ Proba(hauteur eau $>$ seuil)

Monte Carlo : $N = 1e5$, $P = 9e-4$

Décalage sur moyenne

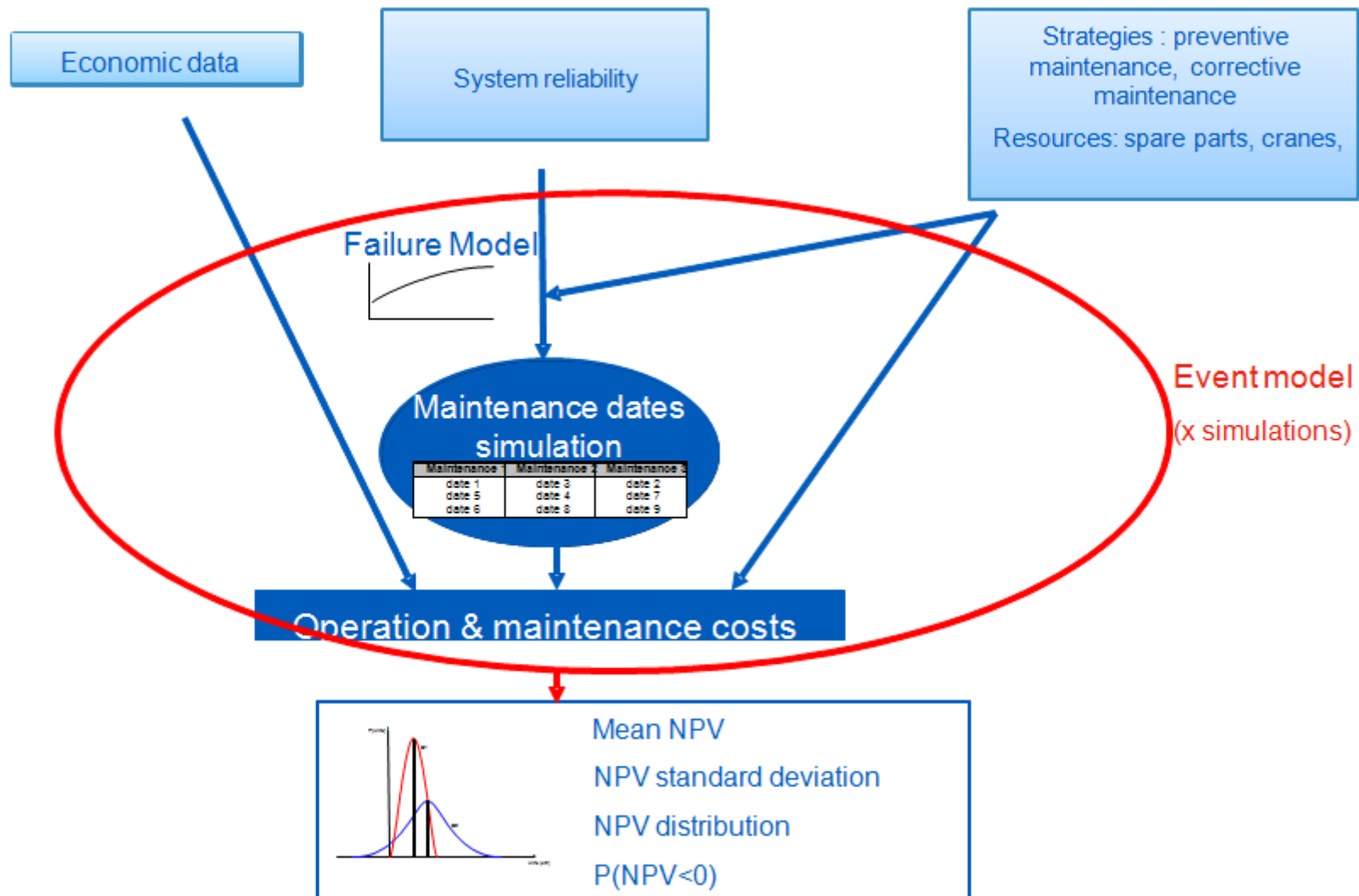


Plan

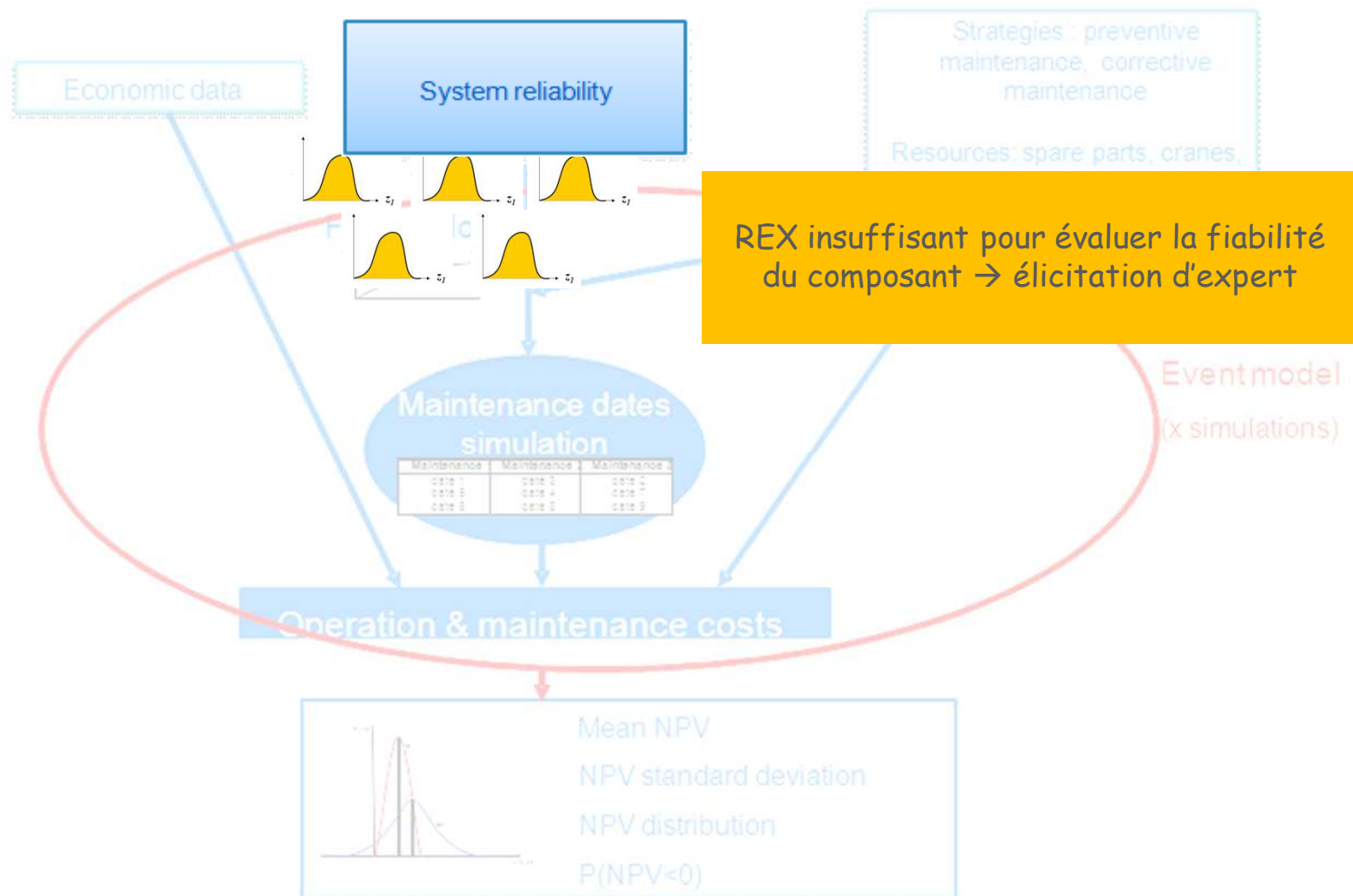
1. Etudes d'incertitudes et de sensibilité
2. Quelques problématiques au-delà du cadre simplifié
3. Exemple en étude de gestion d'actifs

(travaux avec Hélène Bickert, Fanny Douard, Jérôme Lonchampt, Emmanuel Remy, Mahmoud Ezzaki)

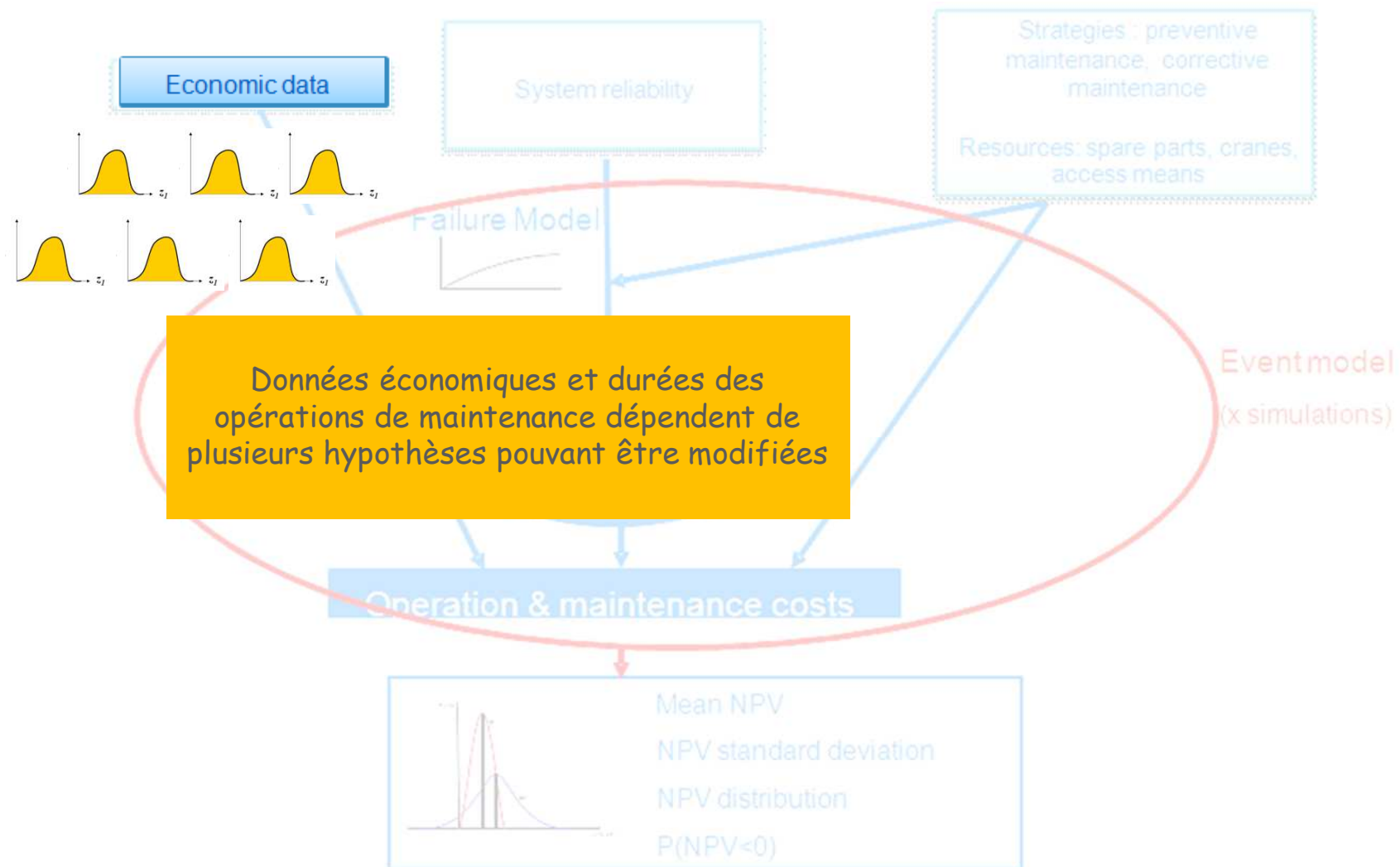
Contexte - Etudes d'Asset management



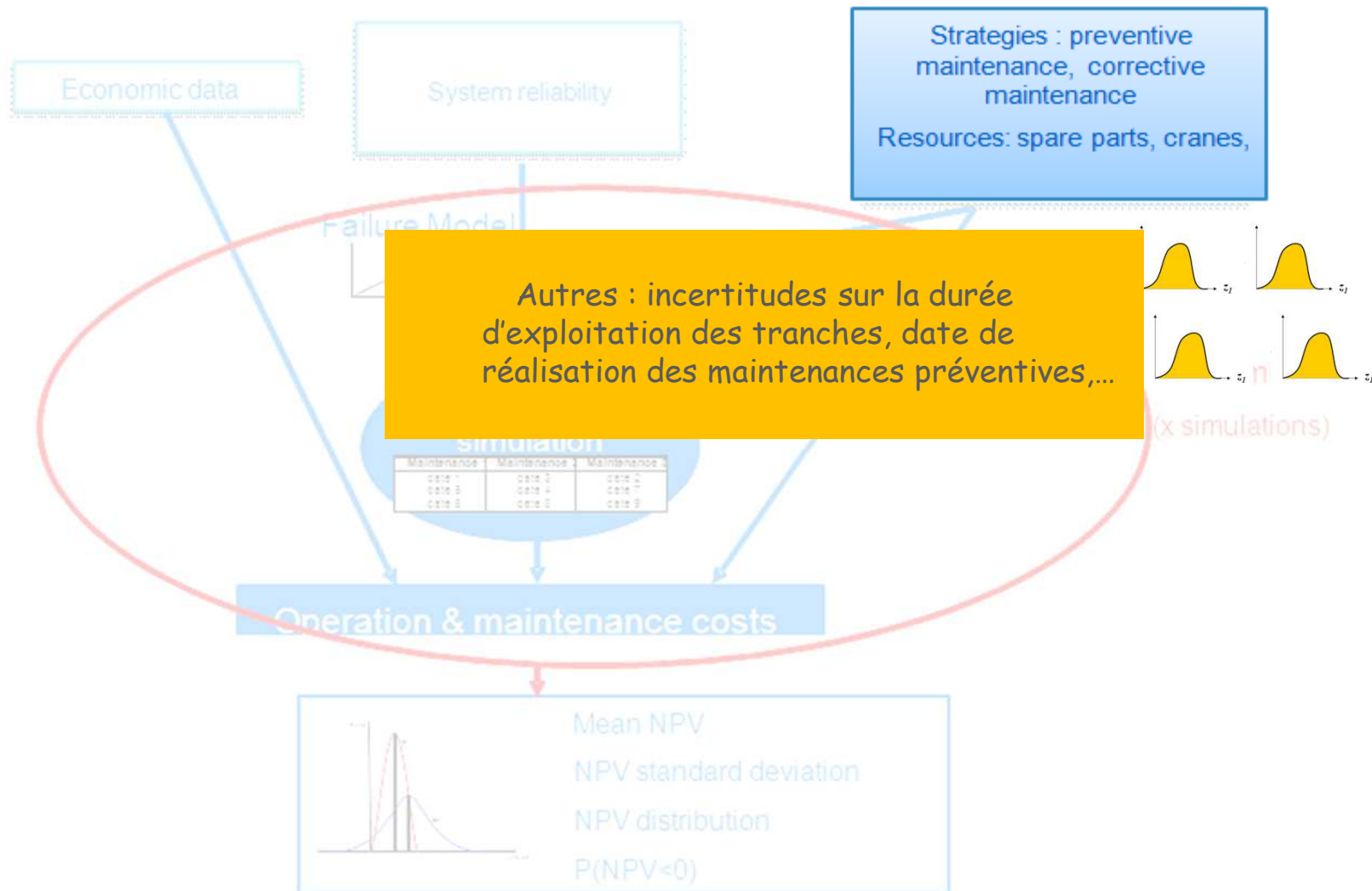
Contexte – Sources d'incertitudes



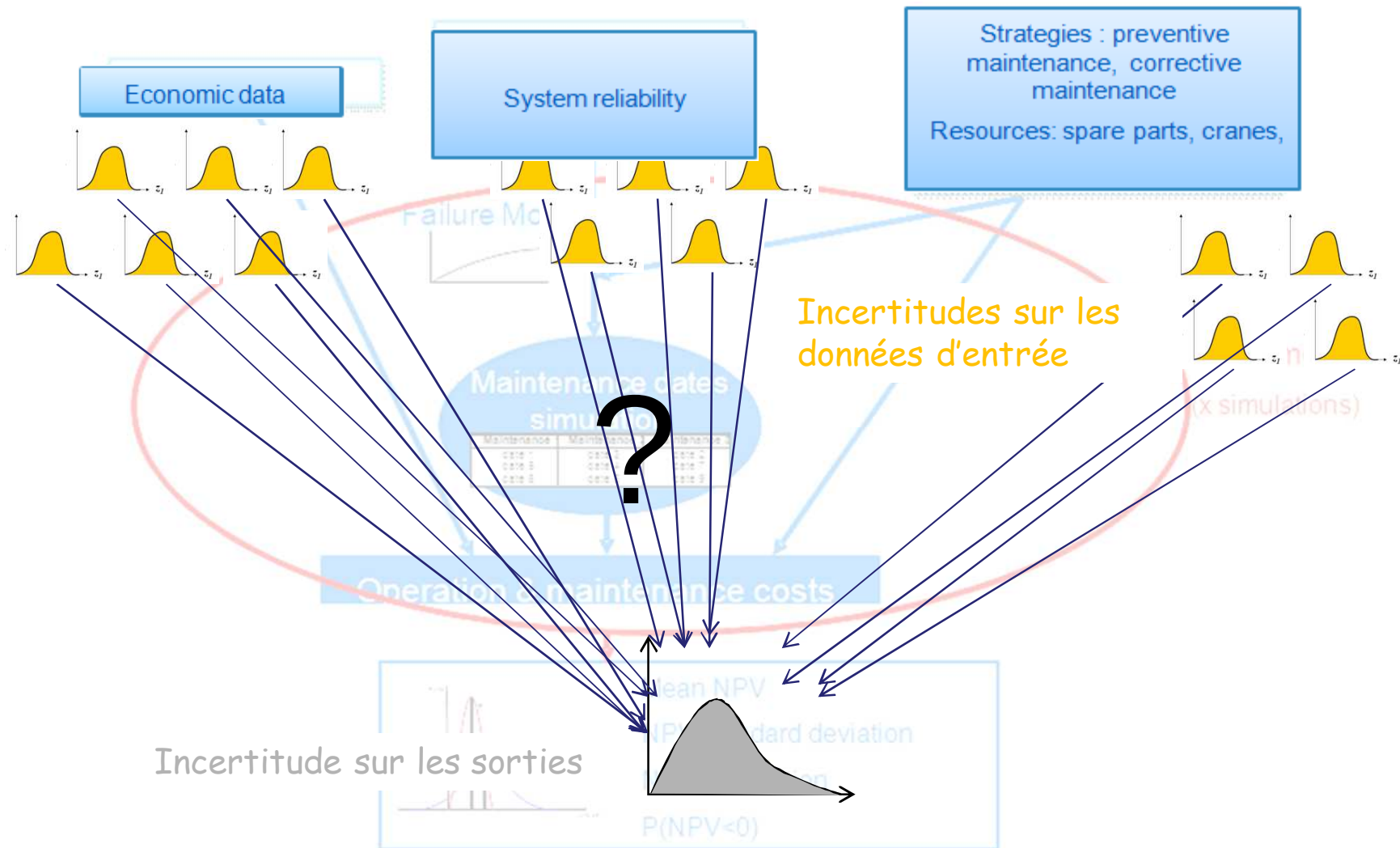
Contexte – Sources d'incertitudes



Contexte – Sources d'incertitudes

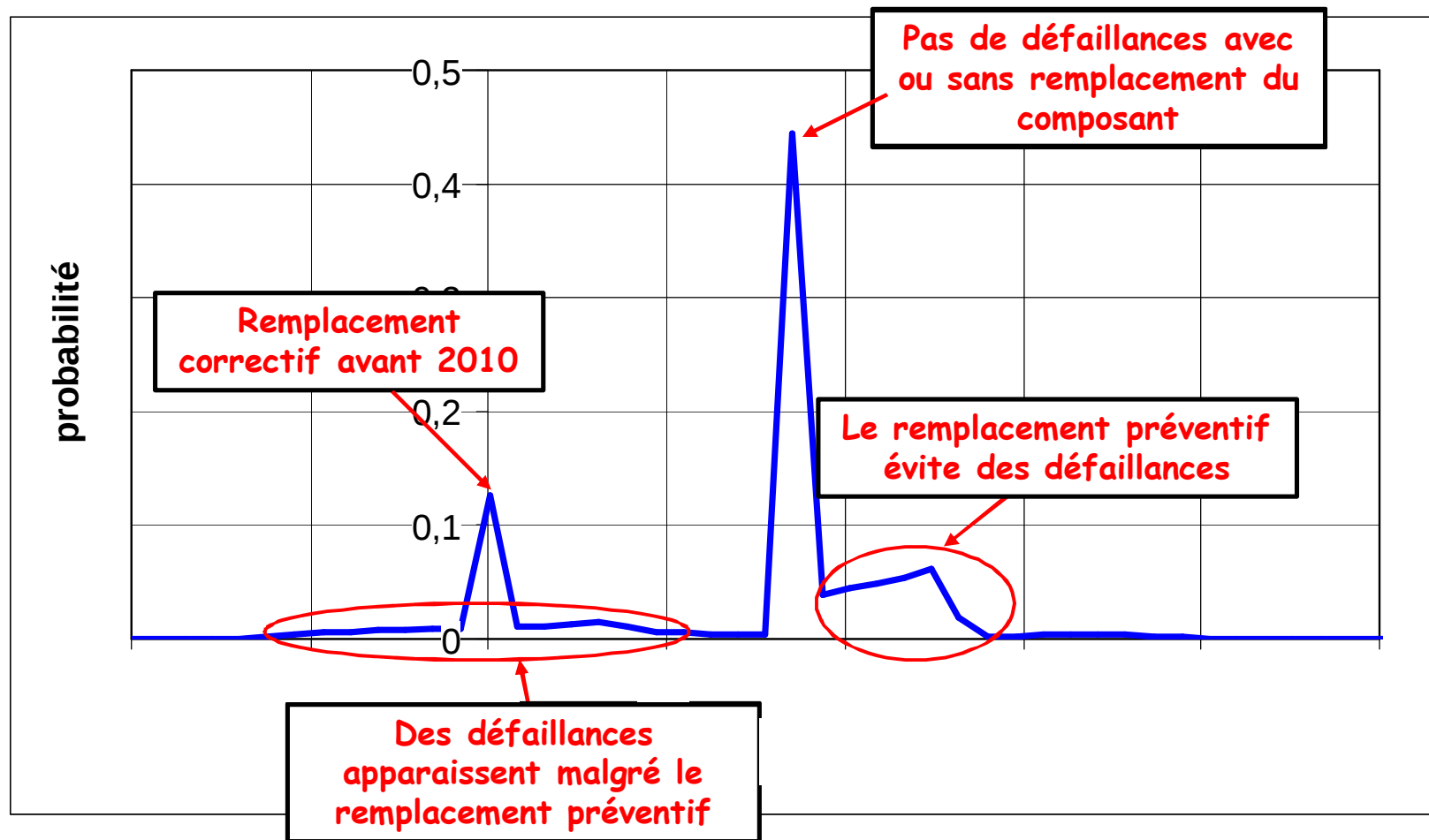


Contexte – Besoin d'analyse de sensibilité



Etude exemple - Définition

- Etude initiale : valorisation du remplacement en 2010 d'un composant d'une centrale thermique
- Résultats :



Probabilité de regretter l'investissement = $6 \cdot 10^{-2}$

Etude exemple - Dimensionnement de l'étude de sensibilité

➤ 11 facteurs étudiés :

- Durée d'indisponibilité et coûts induits par les replacements (fortuits et programmés)
- Valorisation de l'indisponibilité (fortuite et programmée)
- Fiabilité

➤ Quantités d'intérêt étudiées

- E (VAN)
- P (VAN<0)
- σ (VAN)

➤ Méthodologie

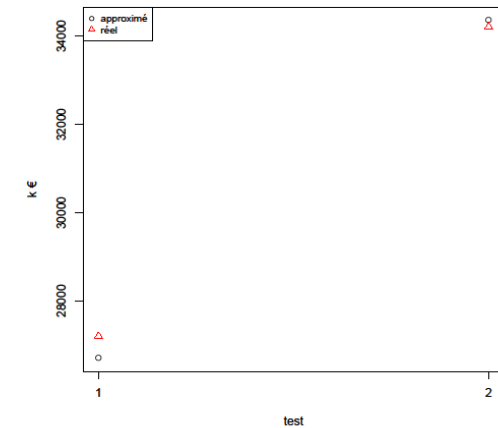
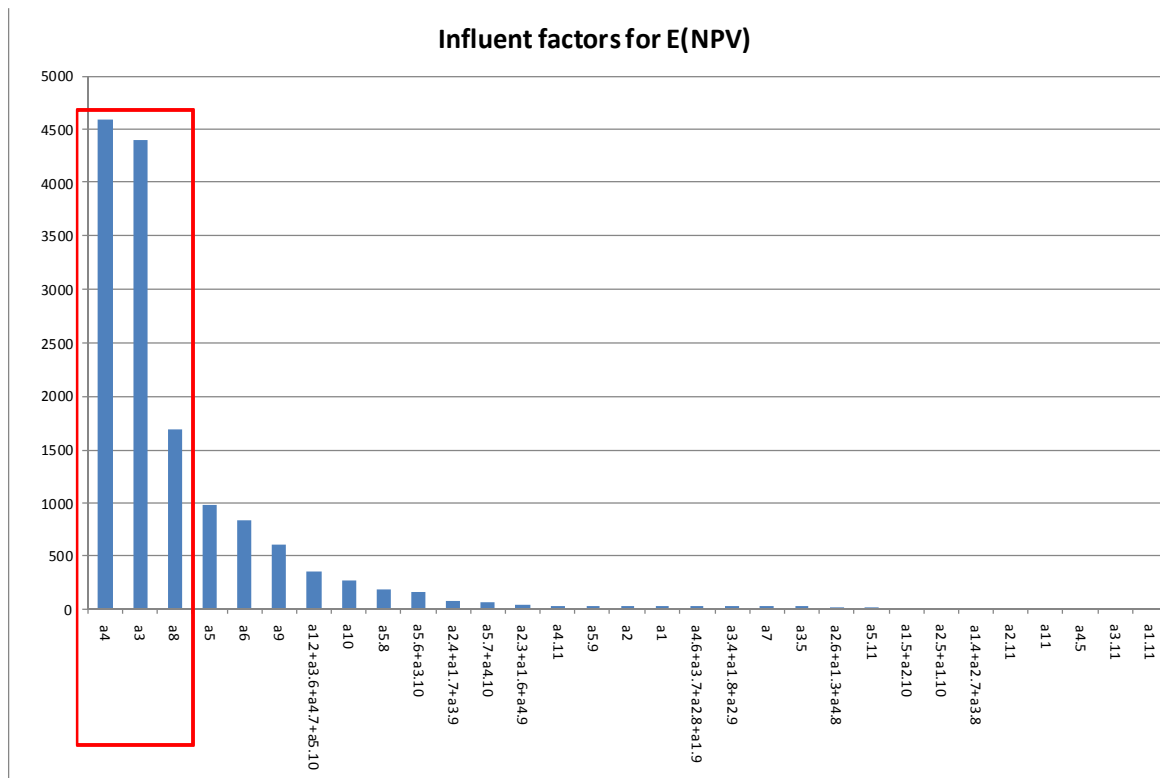
- 1) Criblage
- 2) Estimation d'indices de sensibilité quantitatifs

Etude exemple – Criblage (1/2)

➤ Plan factoriel fractionnaire de résolution IV → 32 runs

Type de résultats obtenus

Validation



Etude exemple – Criblage (2/2)

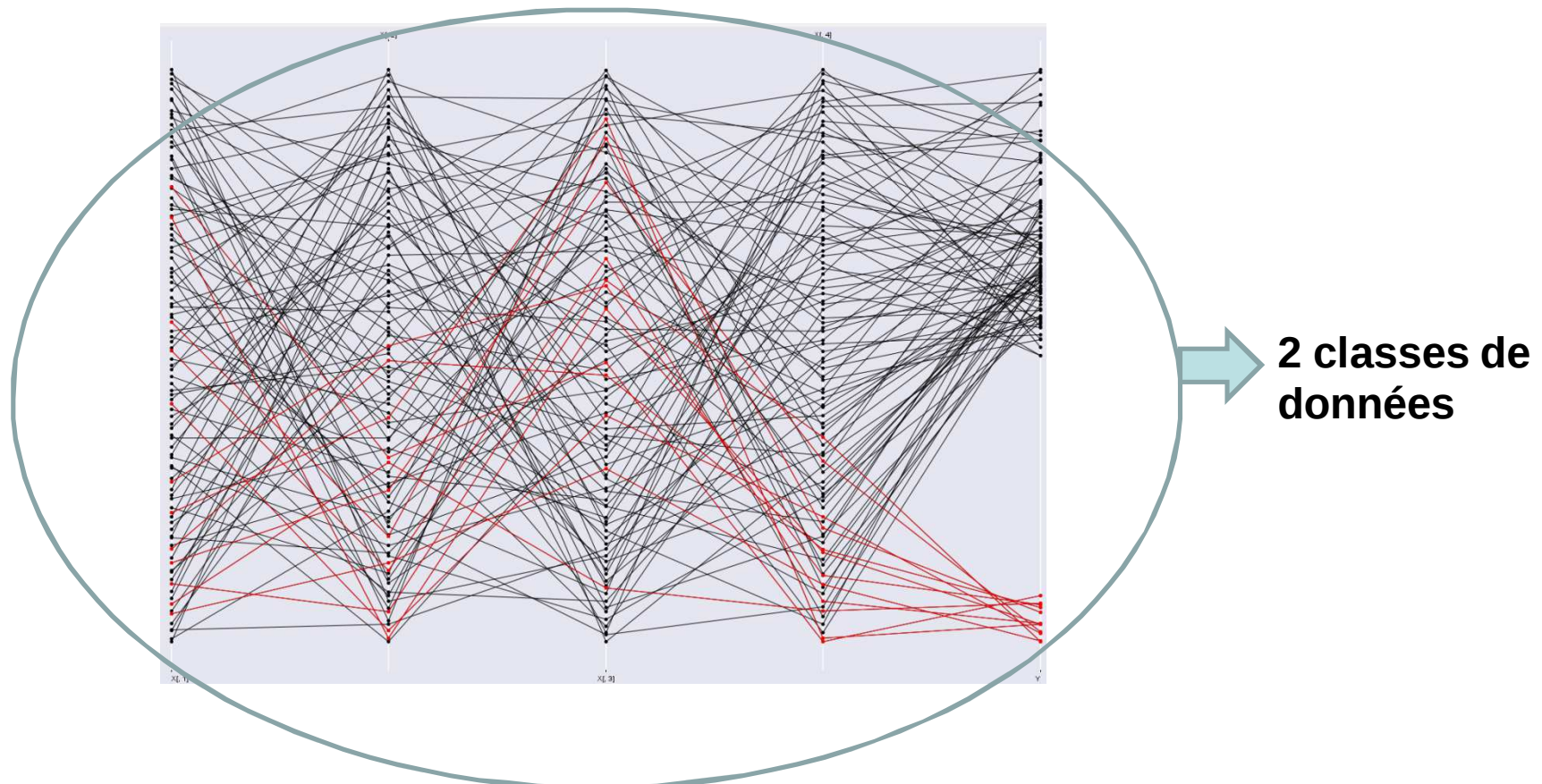
Quantité d'intérêt		Facteurs et interactions influents	
		Facteurs	Interactions
Variables ponctuelles	$E(VAN)$	X_3, X_4, X_8	
	$P(VAN < 0)$	X_3, X_4, X_5, X_8	
	$\sigma(VAN)$	X_4, X_5, X_8	$X_5 X_8$

Facteurs influents pour les sorties économiques	
X3	Coût associé au remplacement du composant
X4	Durée d'indisponibilité (de prolongation d'arrêt) liée au remplacement préventif du composant
X5	Durée d'indisponibilité (fortuite) associée au remplacement fortuit du composant (en correctif)
X8	Facteur multiplicatif sur la valorisation de l'indisponibilité fortuite prise en compte dans l'étude

- 4 facteurs sélectionnés sur 11 en 32 runs
- Tous types de facteurs sélectionnés (coûts, durée de remplacement, données économiques)
- Les facteurs de fiabilité ne sont pas forcément les plus influents

Etude exemple – Analyse de sensibilité

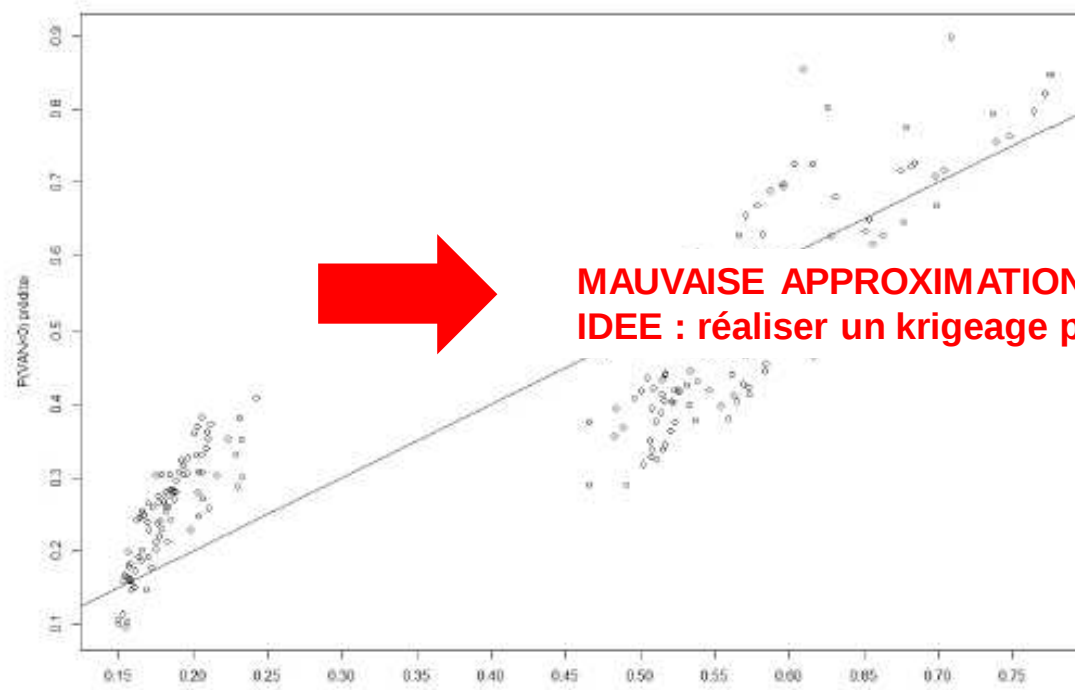
- On étudie l'influence des 4 facteurs sur $P(VAN < 0)$
- Ici code (VME) coûteux → Utilisation d'un métamodèle
 - Plan d'expériences « space filling » (100 calculs) générées par suite de Sobol
- Analyse graphique (cobweb plot) des 100 réponses vs entrées



Etude exemple – Analyse de sensibilité

➤ Métamodèle

- 1^{ère} idée: Krigeage classique



MAUVAISE APPROXIMATION

IDEE : réaliser un krigeage pour chaque classe de points

- Validation croisée :

$$Q^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^{pred})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 62\%$$

Etude exemple – Analyse de sensibilité

➤ Métamodèle

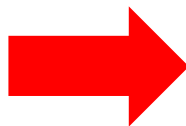
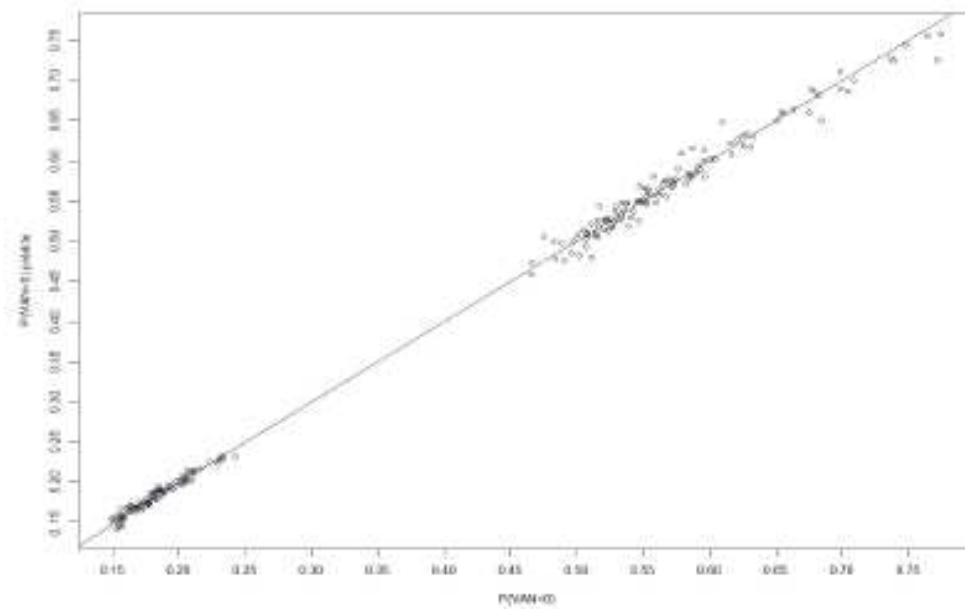
- 2^{ème} idée: Krigeage avec classification :
 - **Principe** : réaliser un krigeage pour chaque classe de données d'entrée.
 - ✓ Définir comment attribuer une classe à une donnée d'entrée
 - ✓ Réaliser un Krigeage pour chaque classe
 - ✓ Méta-modèle associé:



- **Problème** : pas assez de données dans la seconde classe → **solution** :
 - ✓ Repartir les sorties de l'ensemble d'apprentissage en deux classes et affecter chaque entrée du code à la classe correspondante à sa sortie
 - ✓ Générer plusieurs entrées (1000 dans notre cas) à l'aide des suites à faible discrétance
 - ✓ Prédire les classes d'appartenance de ces entrées.
 - ✓ Garder seulement les points se situant dans la classe des probabilités faibles et relancer les calculs pour générer les sorties associées

Etude exemple – Analyse de sensibilité

- **Métamodèle : Krigeage avec classification :**
 - Résultats : $Q^2 = 99\%$



METAMODELE VALIDE

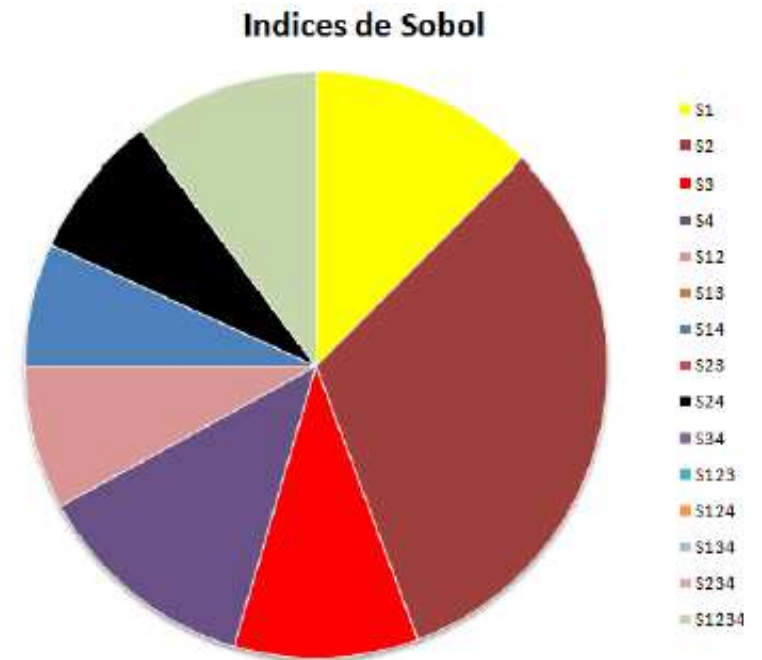
Etude exemple – Analyse de sensibilité

➤ Calcul des indices de sensibilité avec le métamodèle

	$\tilde{S}_i$	$\tilde{S}_{T_i}$	$\tilde{\delta}_i^b$	δ_i^{Ku}	δ_i^{Ks}
$i = 1$	0.118	0.259	0.164	0.153	0.167
$i = 2$	0.287	0.466	0.253	0.246	0.250
$i = 3$	0.093	0.112	0.142	0.143	0.143
$i = 4$	0.270	0.470	0.210	0.196	0.198

$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	$S_{2,3}$	$S_{2,4}$	$S_{3,4}$
0.088	0.001	0.069	0.011	0.081	0.008

$S_{1,2,3}$	$S_{1,2,4}$	$S_{1,3,4}$	$S_{2,3,4}$	$S_{1,2,3,4}$
-0.068	-0.026	-0.027	0	0.095



Etude exemple – Analyse de sensibilité



➤ Conclusion

- 11 facteurs étudiés
- 4 facteurs influents
- Pour $E(VAN)$ et $P(VAN < 0)$ deux facteurs sont deux fois plus influents que les autres
- Pas d'influence forte des interactions

Facteurs influents pour les sorties économiques	
X_1	Coût associé au remplacement du composant
X_2	Durée d'indisponibilité (de prolongation d'arrêt) liée au remplacement préventif du composant
X_3	Durée d'indisponibilité (fortuite) associée au remplacement fortuit du composant (en correctif)
X_4	facteur multiplicatif sur la valorisation de l'indisponibilité fortuite prise en compte dans l'étude

Quantité d'intérêt	Facteurs influents
$E(VAN)$	X_4 : 47% X_2 : 33% X_1 : 19% X_3 : 1%
$\sigma(VAN)$	X_1 : 60% X_2 : 30% X_3 : 8% X_4 : 2%
$P(VAN < 0)$	X_2 et X_4 deux fois plus influents qu' X_1 et X_3

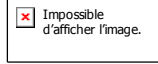
Merci

Cours disponible sur :

<http://www.gdr-mascotnum.fr/iooss1.html>



Biblio



De Rocquigny, Devictor & Tarantola (eds), *Uncertainty in industrial practice*, Wiley, 2008

Rubinstein, *Simulation and the Monte Carlo method*, Wiley, 1981

Lemaire, *Fiabilité des structures*. Lavoisier, 2005

Fang et al., *Design and modeling for computer experiments*, Chapman & Hall, 2006

Kleijnen, *The design and analysis of simulation experiments*, Springer, 2008

Koehler & Owen, *Computer experiments*, 1996

Faivre et al., *Analyse de sensibilité et exploration de modèles – Applications aux sciences de la nature et de l'environnement*, Editions Quaé, à paraître

Saltelli et al., *Sensitivity analysis*, Wiley, 2000

B. Iooss and P. Lemaître. *A review on global sensitivity analysis methods*. In *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, C. Meloni and G. Dellino (eds), Springer, 2015

Lemaître et al., *Density Modification-based reliability sensitivity analysis*, JSCS, 2015

B. Iooss. *Revue sur l'analyse de sensibilité globale de modèles numériques*. *Journal de la Société Française de Statistique*, 152:1-23, 2011