Introduction aux modèles de krigeage par processus gaussiens Apport la vérification des Outils de Calcul Scientifique.

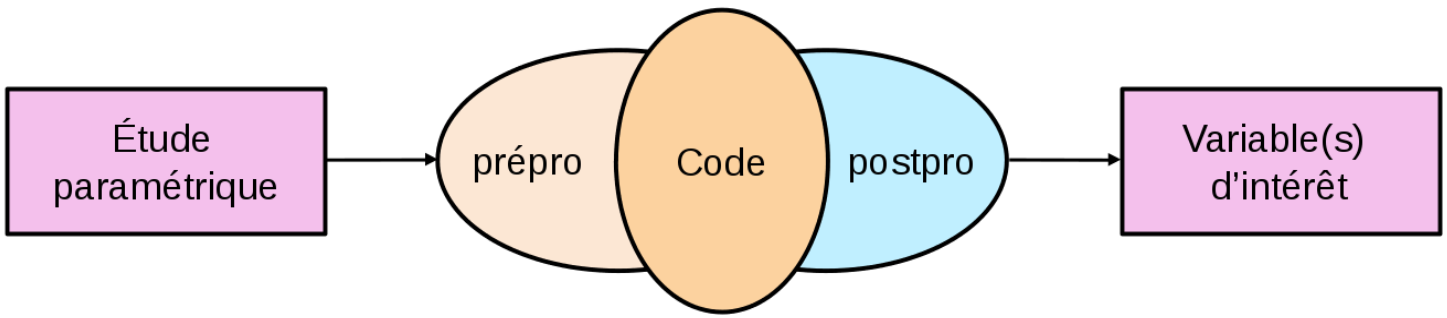
K. Ammar, F. Bachoc, G. Fauchet, J.M. Martinez

CEA-Saclay, DEN, DM2S, F-91191 Gif-sur-Yvette, France.

Séminaire inTheArt, Orme des merisiers, 13 juin 2019

Complexité de la simulation numérique

- Outils de Calcul Scientifique, OCS
- Couplages multi-physiques, multi-échelles
- Pré-traitements : modèles, paramètres, données, ...
- Numérique : maillages (espace, temps), seuils de convergence, nombre de simulations/particules (Monte Carlo)
- Post-traitements : extraction/transformation de données, remontées d'échelles, ...



K. Ammar (thèse CEA/DEN/DM2S/SERMA, 2014)

- Impossibilité de vérifier manuellement tous les calculs spécifiés par un *gros* plan d'expériences numériques

Méta-modèles

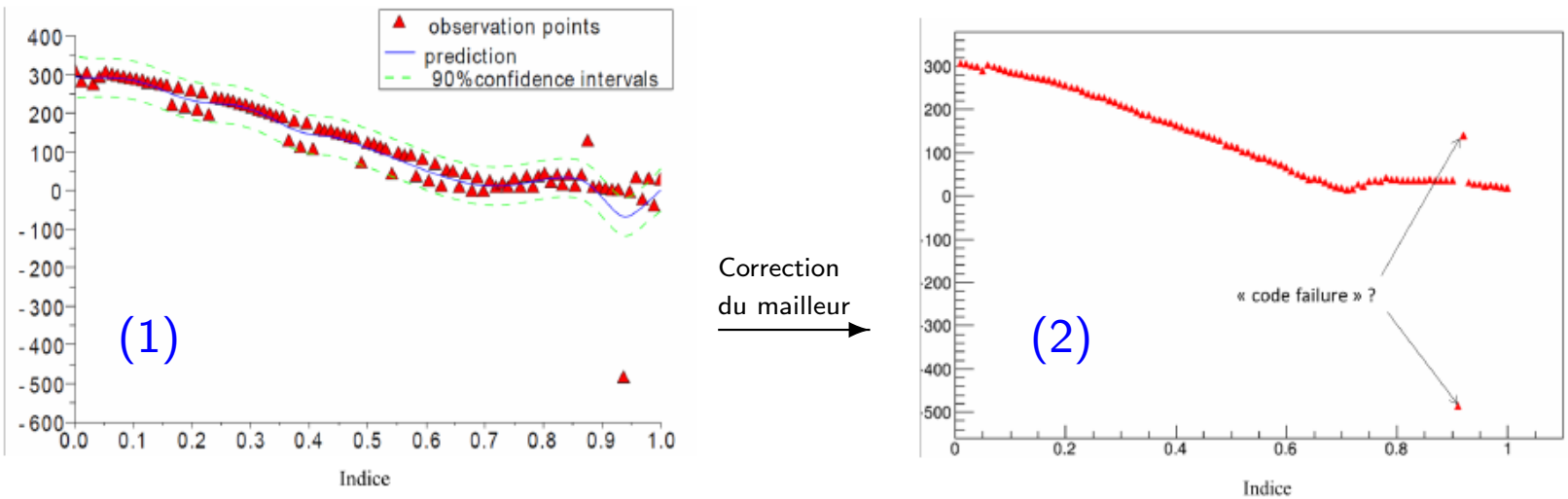
- Travaux au CEA/DEN/DM2S (K. Ammar, F. Bachoc) sur l'apport des méta-modèles dans la *vérification* des OCS
  - métamodèle pour *approcher* l'OCS dans un domaine restreint
  - polynômes, splines, réseaux de neurones, méthodes à noyaux, krigeage
  - temps de calcul fortement réduits (études : conception, optimisation, ...)
- *Grands* écarts OCS - métamodèle peuvent être *expliqués* par
  - des phénomènes fortement non linéaires non captés par le métamodèle
  - des erreurs de calculs, de programmation de pré ou post traitements, ... $\rightarrow$ *outliers* parmi les résultats de l'OCS
- Certaines approximations numériques (discrétisation, maillage, seuils de convergence) peuvent introduire
  - de légères discontinuités $\rightarrow$ pseudo-*bruit* sur les valeurs calculées par l'OCS

Thèse de K. Ammar (CEA/DEN/DM2S/SERMA 2014)

- Méta-modèle krigeage du code thermomécanique **Germinal** simulant le comportement du combustible dans un REP

(1) *Oscillations* du schéma de calcul dues au maillage axial → bruit de mesure anormalement grand

(2) Calculs erronés n'ayant pas été pris en compte par le post-processeur → écarts code-méta-modèle(krigeage) anormaux



- Marge à la fusion (cœur Astrid) fonction de 11 paramètres. Visualisation des variations dans une direction de $\mathbb{R}^{11}$

Références

K. Ammar, Conception multi-physiques et multi-objectifs des coeurs de RNR-Na hétérogènes et développement d'une méthode d'optimisation sous incertitudes, *Paris Sud, thèse 9/12/2014*.

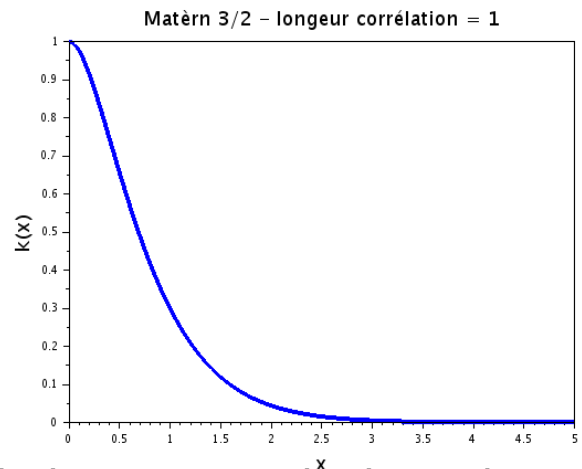
Bachoc F, Ammar K., Martinez J.M, Improvment of code behavior in a design of experiments by metamodeling, *Nuclear Science Engineering, Vol. 183, 2016*.

Processus Gaussiens

Un processus $Z(\mathbf{x})$ est gaussien si $\forall n \in \mathbb{N}, \forall a_{i=1,\ldots,n}, \sum_{i=1}^{n} a_i Z(\mathbf{x}_i)$ est gaussien. $Z(\mathbf{x})$ est caractérisé par ses fonctions moyenne et covariance

- Fonction moyenne : $x \rightarrow m(\mathbf{x}) := \mathbb{E}(Z(\mathbf{x}))$.
- Fonction de covariance : $k(\mathbf{x}_1, \mathbf{x}_2) := \mathtt{Cov}(Z(\mathbf{x}_1), Z(\mathbf{x}_2))$.
- Paramétrée, la fonction de covariance permet à un processus gaussien de s'adapter à un grand nombre de fonctions
- Exemple à une dimension ($x \in \mathbb{R}$), moyenne linéaire et covariance de Matèrn de régularité fixée à 3/2

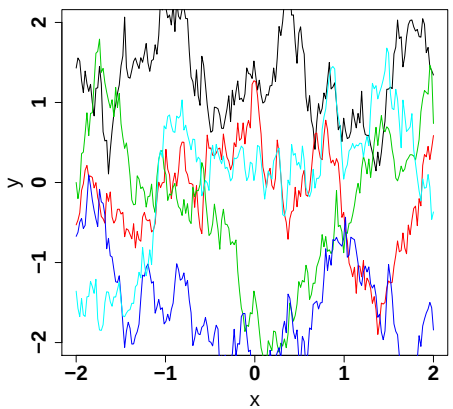
$$
\begin{aligned}
m(x) &= \beta_0 + \beta_1 x \\
k(x_1, x_2) &= \sigma^2 \times \exp[-\sqrt{6}\delta] \times (1 + \sqrt{6}\delta) \\
\delta &= \frac{|x_1 - x_2|}{\mathtt{lc}}
\end{aligned}
$$

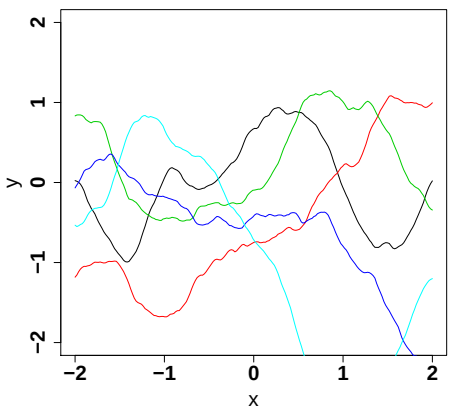


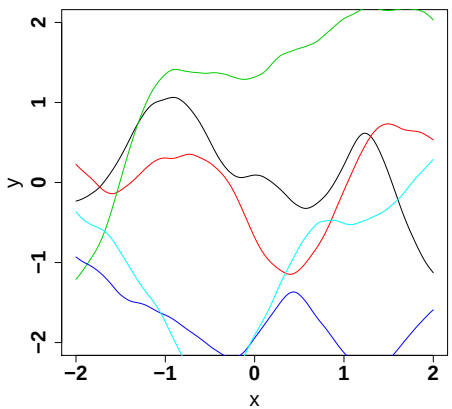
- 4 paramètres à estimer : les coefficients β_0, β_1 de la moyenne linéaire, la variance σ^2 du processus et la longueur de corrélation $\mathtt{lc}$.

Emulateurs de fonctions déterministes

- Méta-modèle krigeage : *la fonction à approcher est supposée être une réalisation d'un processus gaussien ! ! !*
- On dispose d'une large classe de fonctions de covariance permettant de représenter un large spectre de fonctions plus ou moins régulières

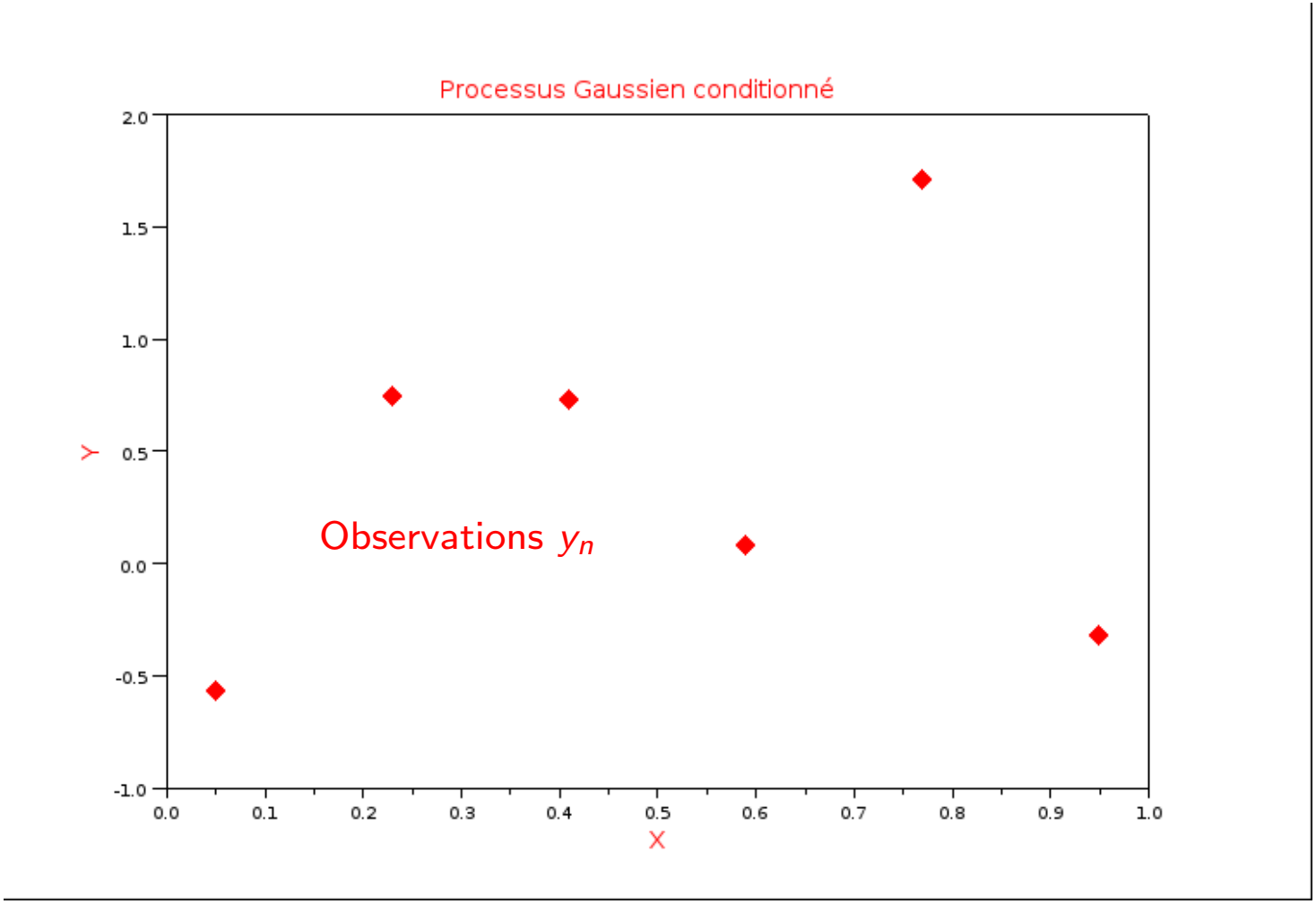




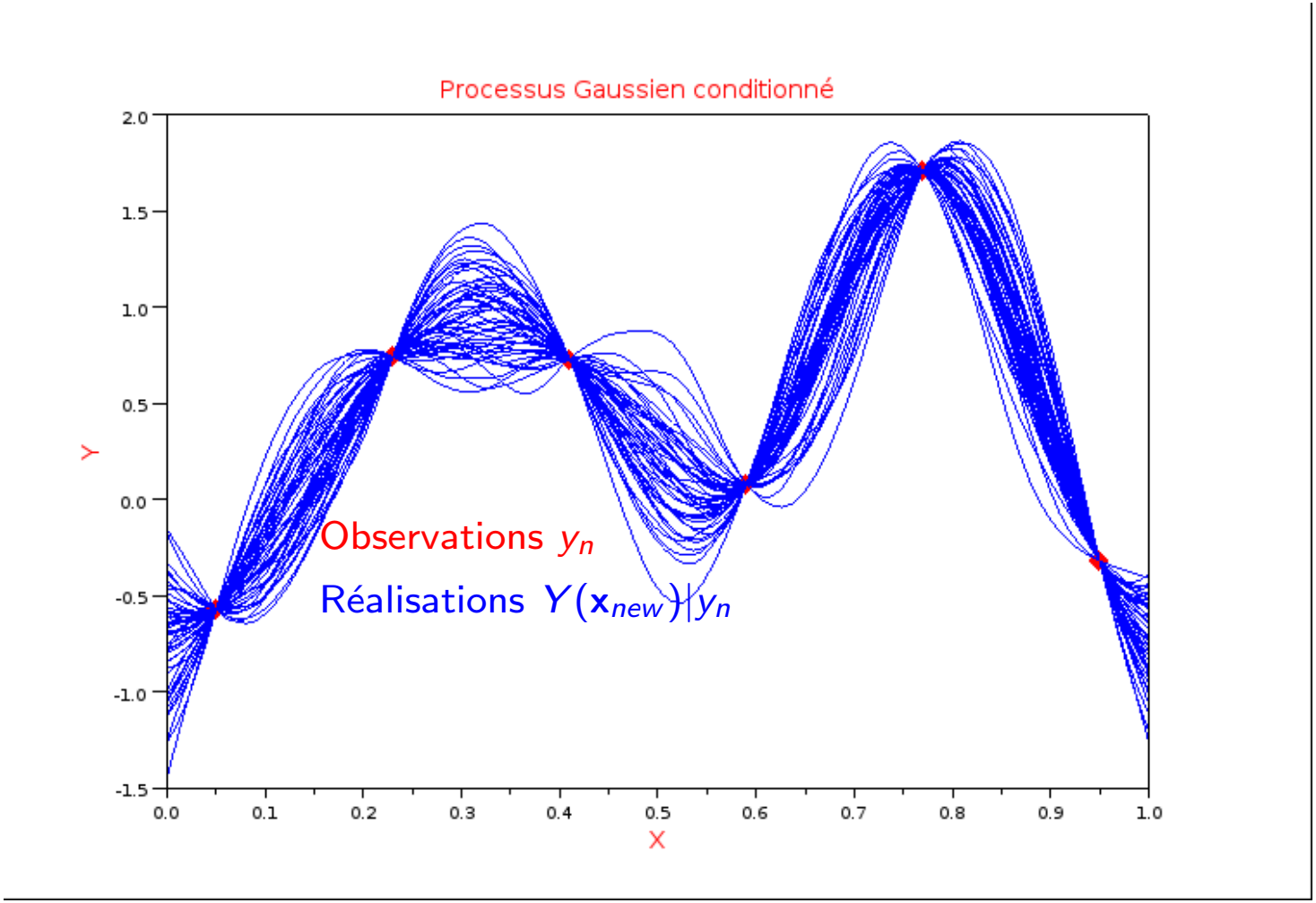


- En multi-dimension : prise en compte de l'anisotropie

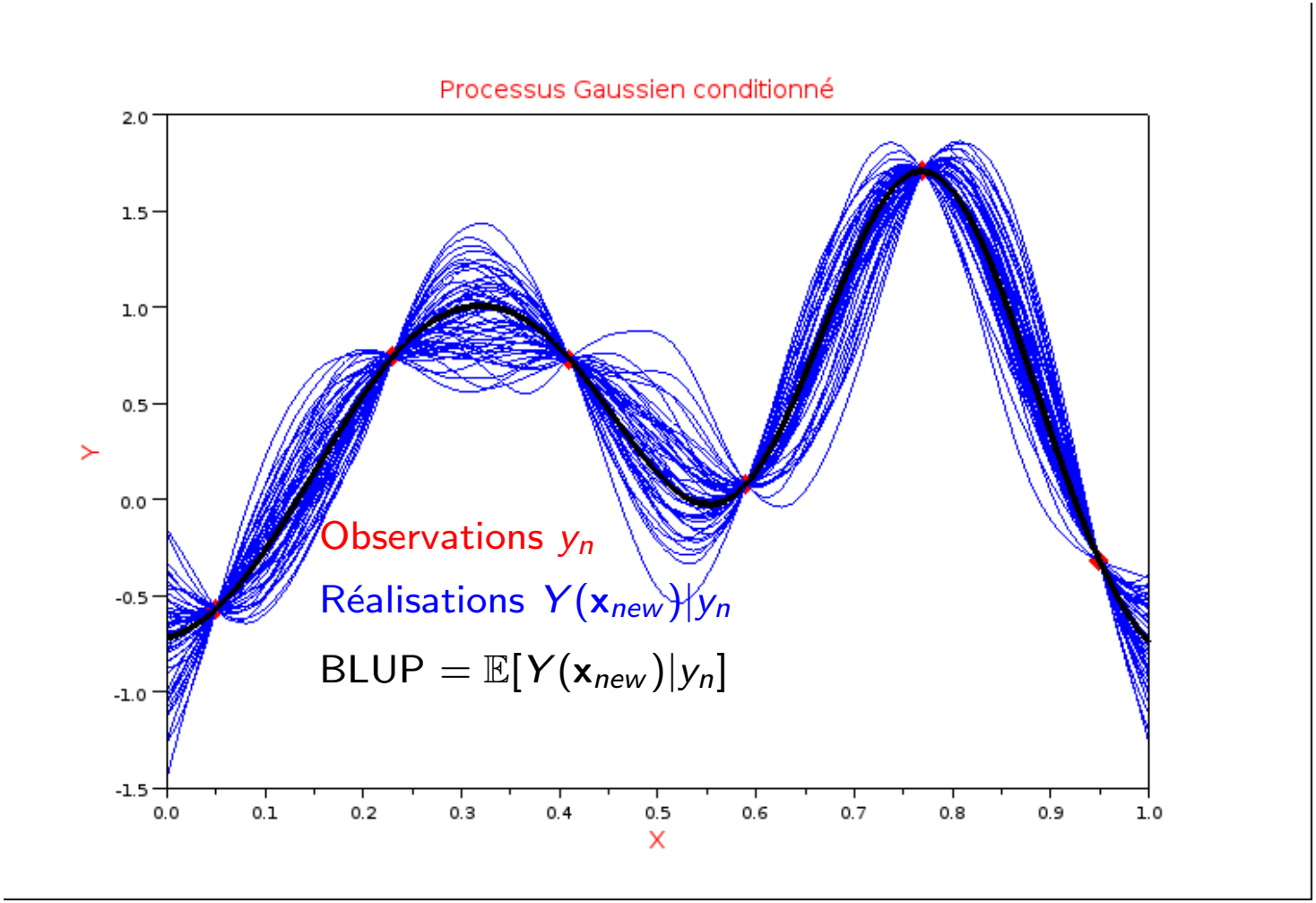
Principe du krigeage



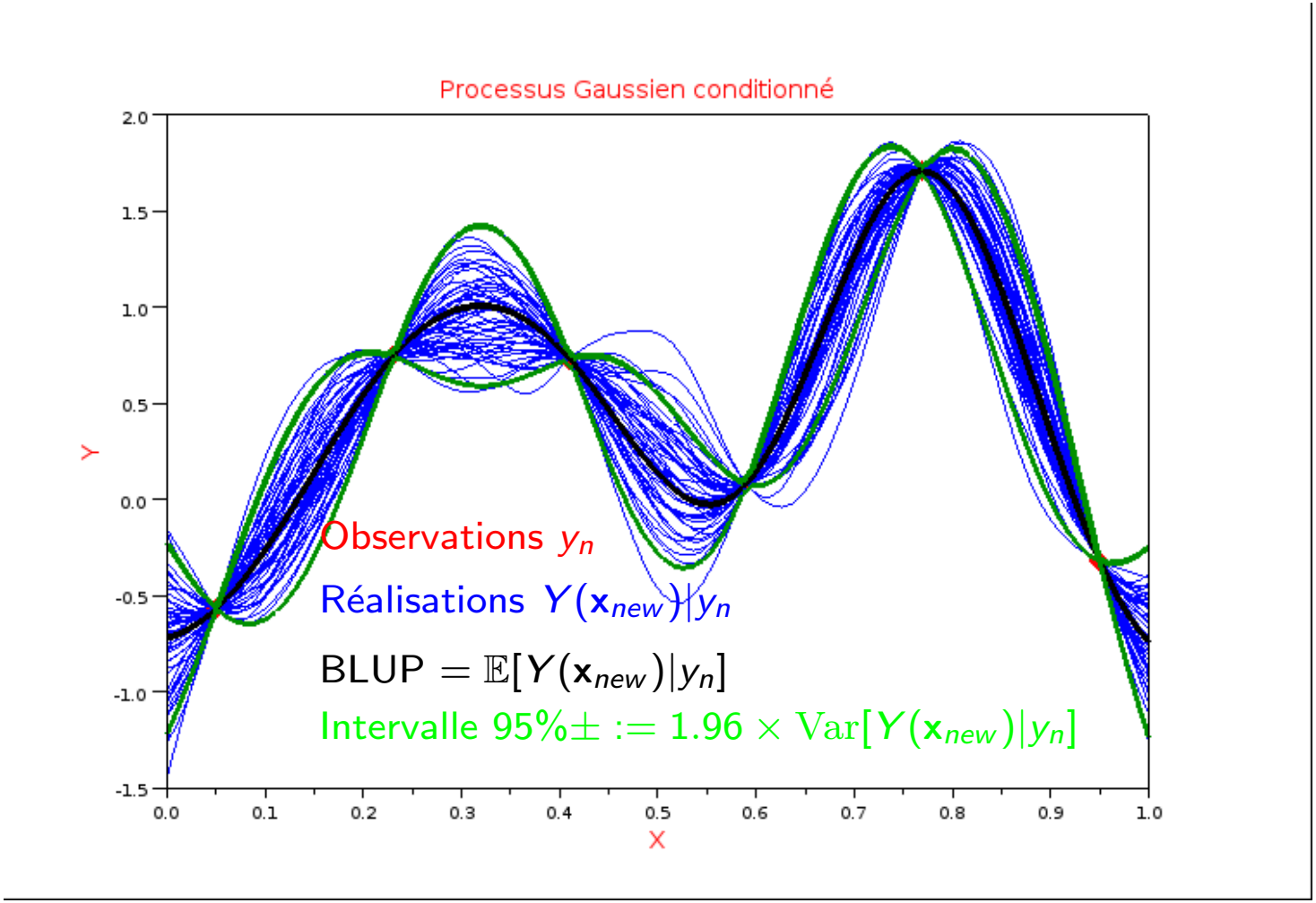
Principe du krigeage



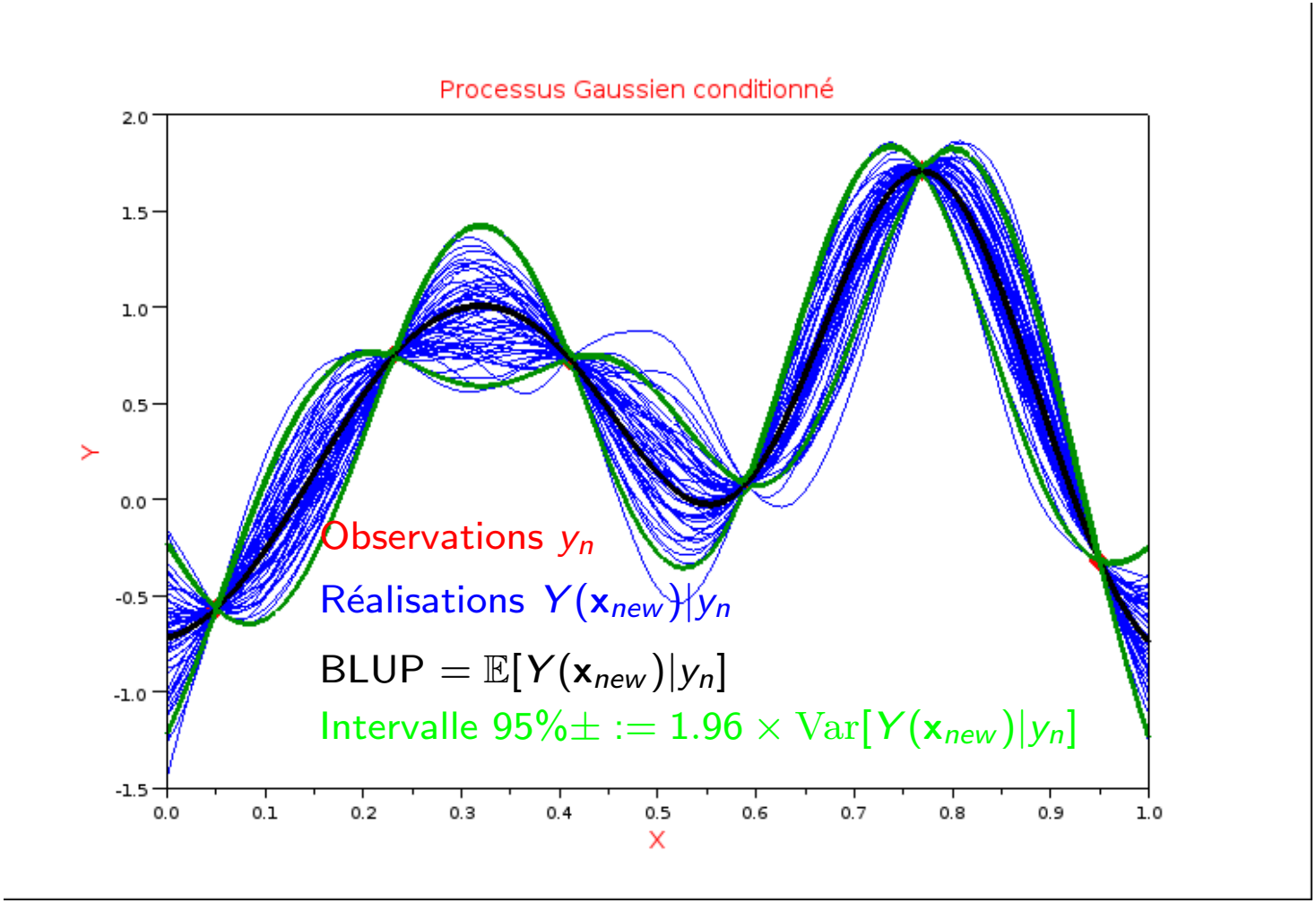
Principe du krigeage



Principe du krigeage



Principe du krigeage



- Expressions analytiques des espérances et variances conditionnelles
- Point fort du krigeage : intervalle de confiance de l'estimation

Exemple d'un krigeage de moyenne nulle

- Observations $\mathbf{y}_{obs} = [y_{obs,i}]_{1 \leq n}$ aux points $[\mathbf{x}_i]_{1 \leq i \leq n}$
- Choix d'une fonction paramétrique de covariance $k(\mathbf{x}_i, \mathbf{x}_j; \theta)$
- Estimation des paramètres θ par maximum de vraisemblance
- Calcul de la matrice de covariance $\boldsymbol{K} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{1 \leq i,j \leq n}$
- Loi de la prédiction conditionnelle aux observations $Y(\mathbf{x})|\mathbf{y}_{obs}$

$$
\begin{aligned}
\mathbf{k}(\mathbf{x}) &= [k(\mathbf{x}, \mathbf{x}_i)]_{1 \leq i \leq n} \\
\texttt{moyenne} : y(\mathbf{x}) &= \mathbf{k}^t(\mathbf{x}) \underbrace{\boldsymbol{K}^{-1}\mathbf{y}_{obs}}_{\texttt{coefficients} \in \mathbb{R}^n} \\
\texttt{variance} : \sigma^2(\mathbf{x}) &= k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^t(\mathbf{x})\boldsymbol{K}^{-1}\mathbf{k}(\mathbf{x})
\end{aligned}
$$

- Moyenne $y(\mathbf{x})$ $\rightarrow$ estimation par le méta-modèle
- Variance $\sigma^2(\mathbf{x})$ $\rightarrow$ intervalles de confiance (quantiles)

$$
Y(\mathbf{x})|\mathbf{y}_{obs} \sim \mathcal{N}(y(\mathbf{x}), \sigma^2(\mathbf{x}))
$$

Validation du modèle

- Lorsqu'il n'y a pas de bruit de mesure $\rightarrow$ interpolation. Le modèle de krigeage passe par tous les points $[\mathbf{x}_i]_{1 \leq i \leq n}$

$$\begin{aligned} y(\mathbf{x}_i) &= y_{obs,i} \\ \sigma^2(\mathbf{x}_i) &= 0 \end{aligned}$$

- La validation du krigeage doit se faire par des techniques de Validation Croisée
  - on construit le modèle sur une partie de la base d'exemples puis on le teste sur l'autre partie
  - erreur est estimée par la moyenne des erreurs de test en effectuant plusieurs découpages différents de la base
- La méthode de validation la plus adaptée au krigeage est le Leave One Out
  - on construit le modèle sur la base d'exemples à laquelle on a retiré l'exemple $(\mathbf{x}_i, y_i)$ et on calcule l'erreur d'estimation $\epsilon_{loo,i} = y_i - \hat{y}_i$
  - erreur de généralisation (fonctionnelle) est estimée par la moyenne des erreurs Leave One Out

$$\mathtt{EQM}_{loo} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_{loo,i}^2$$

Validation par Leave One Out

- Base d'exemples $\mathcal{L} = \{(\mathbf{x}_1, y_{obs,1}), \ldots, (\mathbf{x}_n, y_{obs,n})\}$
- Procédé d'apprentissage : $\mathcal{L} \rightarrow$ méta-modèle $\mathcal{M}_{\mathcal{L}}$
- Construire un méta-modèle sans utiliser l'exemple $(\mathbf{x}_i, y_{obs,i})$ puis calculer l'erreur d'estimation pour cet exemple

Algorithm 1 Erreurs par Leave One Out

for $1 \leq i \leq n$ **do**
  Retirer l'exemple i : $\mathcal{L}_{(-i)} = \mathcal{L} \setminus \{(\mathbf{x}_i, y_{obs,i})\}$
  Construire le méta-modèle $\mathcal{M}_{\mathcal{L}_{(-i)}}$ en utilisant la base $\mathcal{L}_{(-i)}$
  Calculer l'erreur Leave One Out $\epsilon_{loo,i} = y_{obs,i} - \mathcal{M}_{\mathcal{L}_{(-i)}}(\mathbf{x}_i)$
end for

- les erreurs Leave One Out s'obtiennent à partir du méta-modèle $\mathcal{M}_{\mathcal{L}}$ sans avoir a construire tous les méta-modèles $[\mathcal{M}_{\mathcal{L}_{(-i)}}]_{1 \leq i \leq n}$

$\rightarrow$ c'est une propriété des méta-modèles lorsque leurs estimations sont des formes linéaires des observations (méthodes à noyaux, RKHS : Reproducing Kernel Hilbert Space)

Calcul des erreurs Leave One Out du krigeage

- Formule du cas avec moyenne nulle

$$\boldsymbol{\epsilon}_{loo} \quad = \quad [\mathtt{diag}(\boldsymbol{K}^{-1})]^{-1}\boldsymbol{K}^{-1}\mathbf{y}_{obs}$$

- De plus le krigeage permet d'estimer la variance des erreurs LOO

$$\boldsymbol{\sigma}^2_{loo} \quad = \quad [\mathtt{diag}(\boldsymbol{K}^{-1})]^{-1}$$

- Une erreur LOO suit donc une loi normale de moyenne (nulle dans notre cas) et de variance connue.
- Les erreurs LOO standardisées doivent donc suivre une loi normale centrée réduite

$$\tilde{\epsilon}_{loo,i} = \frac{\epsilon_{loo,i}}{\sigma_{loo,i}}$$

qu'on utilise pour détecter d'éventuelles erreurs de simulation

- Exemple de détection d'*outliers* par $|\tilde{\epsilon}_{loo,i}| > 3$ correspondant à une probabilité de 1.35×10^{-3}

Modèle physique testé

- Frottement F fonction du diamètre hydraulique D_h, du débit U, de la masse volumique ρ, de la viscosité dynamique μ
- Modèle du coefficient de frottement C_f fonction du Reynolds Re

$$
\begin{aligned}
\mathtt{Re} &= \frac{U \times D_h}{\nu}, \quad \nu = \frac{\mu}{\rho} \ \mathtt{viscosité\ cinématique} \\
C_f &= \begin{cases} 64 \times \mathtt{Re}^{-1} & \mathtt{si} \quad \mathtt{Re} < 1000 \\ \mathtt{Coef} \times \mathtt{Re}^{-0.25} & \mathtt{sinon} \end{cases} \\
F &= \frac{C_f}{2 \times D_h} \times U^2
\end{aligned}
$$

- Valeur exacte de Coef $= 64 \times 1000^{-0.75} = 0.3598984$.

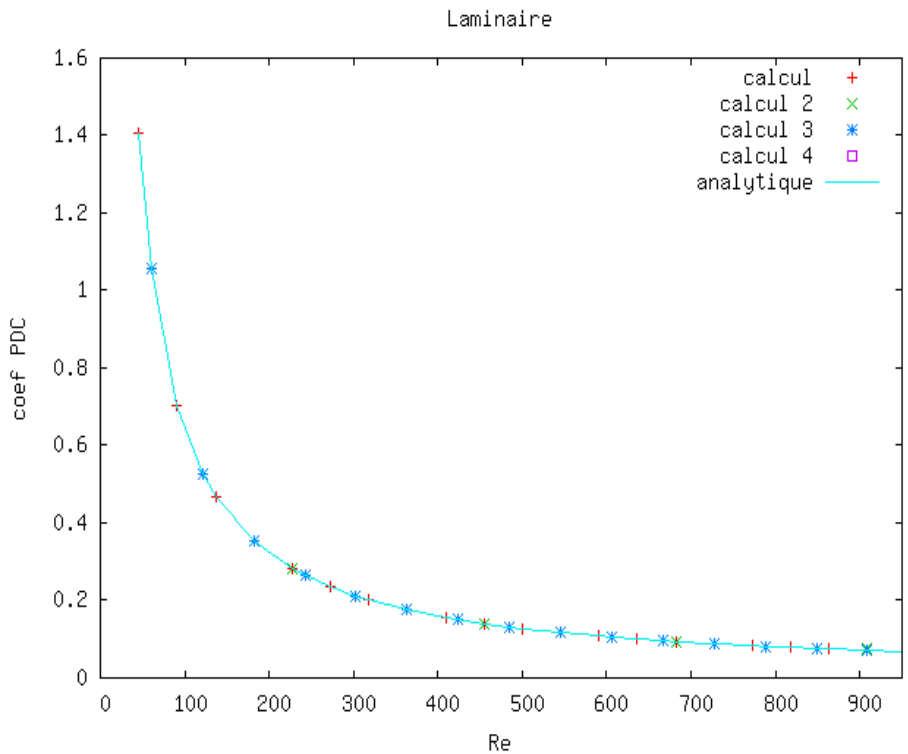
Erreurs possibles ... malgré une fiche de vérification

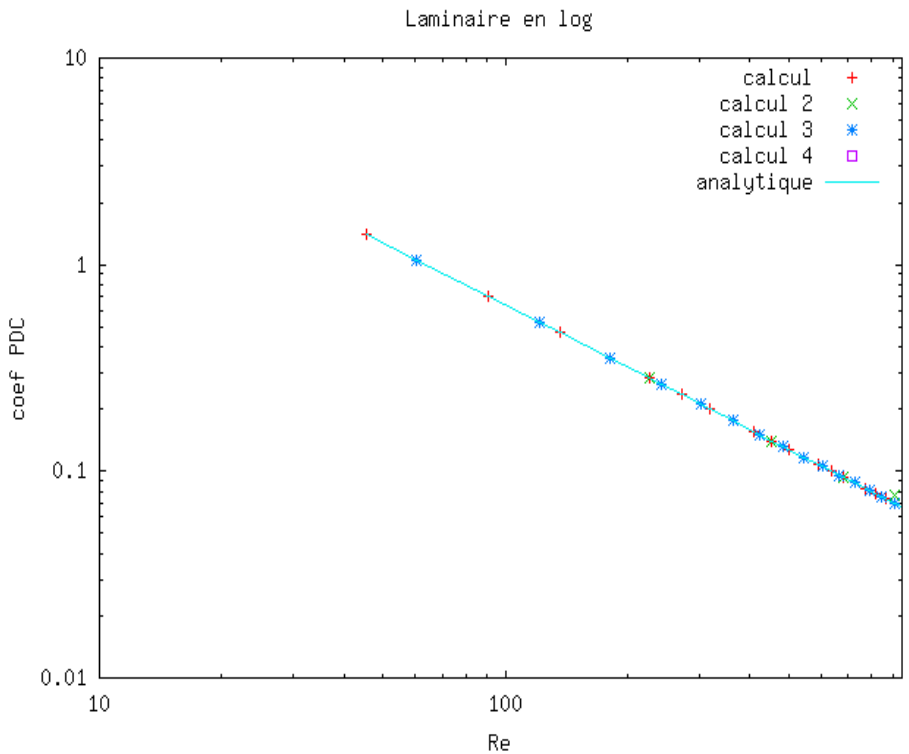
Dans la fiche de vérification associée à ce modèle, on recalcule le coefficient de frottement à partir du postraitement du terme source. On teste différentes valeurs de (μ, ρ, D_h, U) (400 points de fonctionnement).

- 100% des lignes de code sont testées par la fiche de vérification
- MAIS cela ne veut pas dire que 100% des situations sont testées (dimension, direction non orthogonale...)
- NI que la fiche en elle même est juste : exemple on remplace partout (src,doc,fiche) 0.35 par 0.75
- NI que le modèle est juste au départ (est il continu ?)

Erreurs possibles ... malgré une fiche de vérification

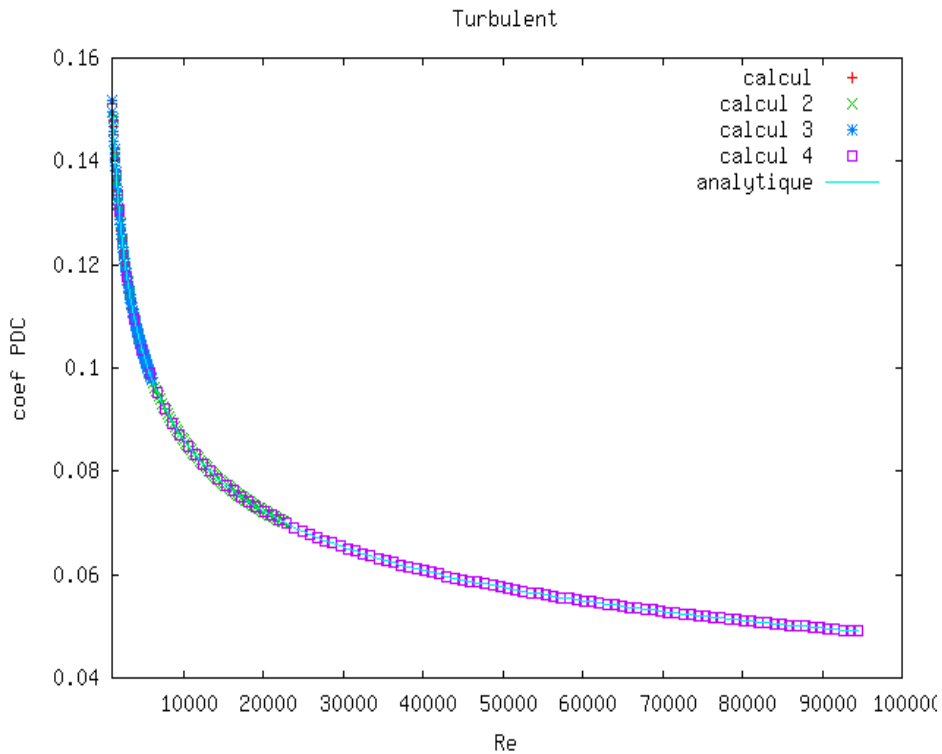
Images de la fiche de vérification : régime laminaire ..

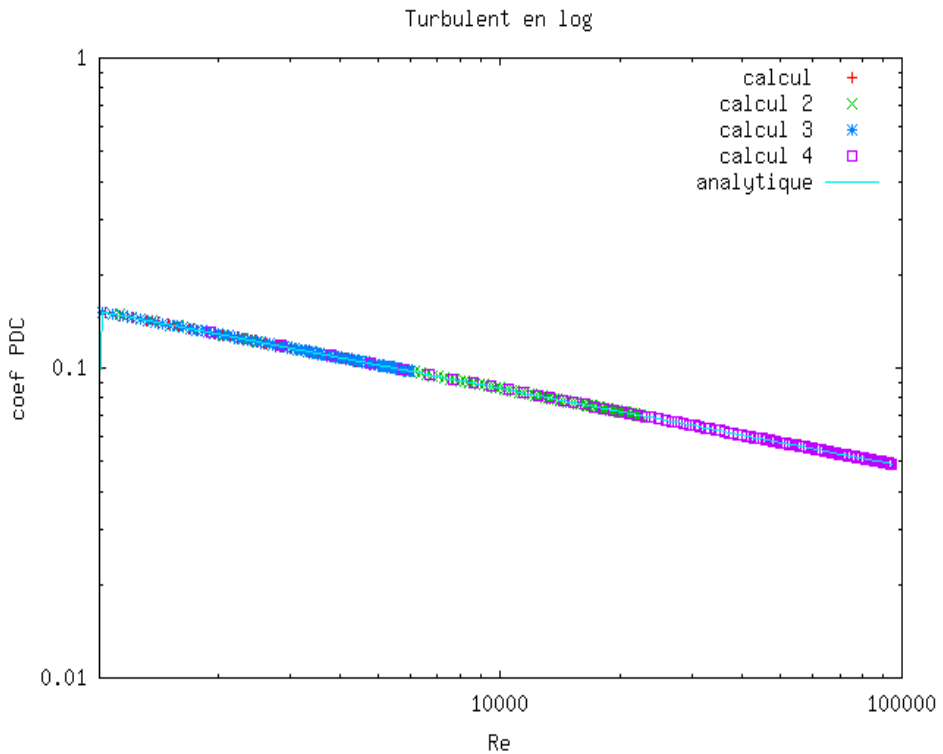




Erreurs possibles ... malgré une fiche de vérification

Images de la fiche de vérification régime turbulent





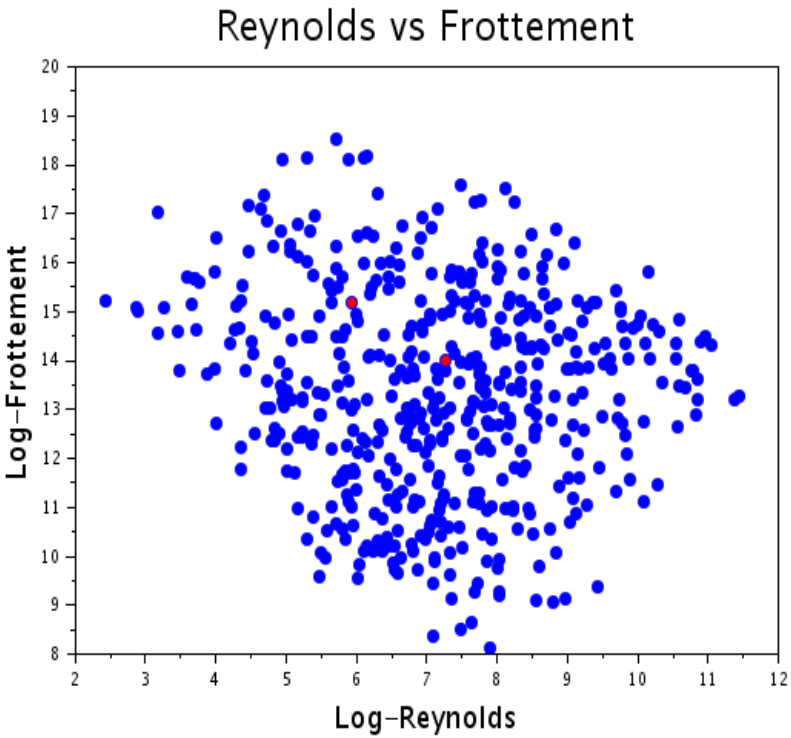
Modèle de krigeage utilisé

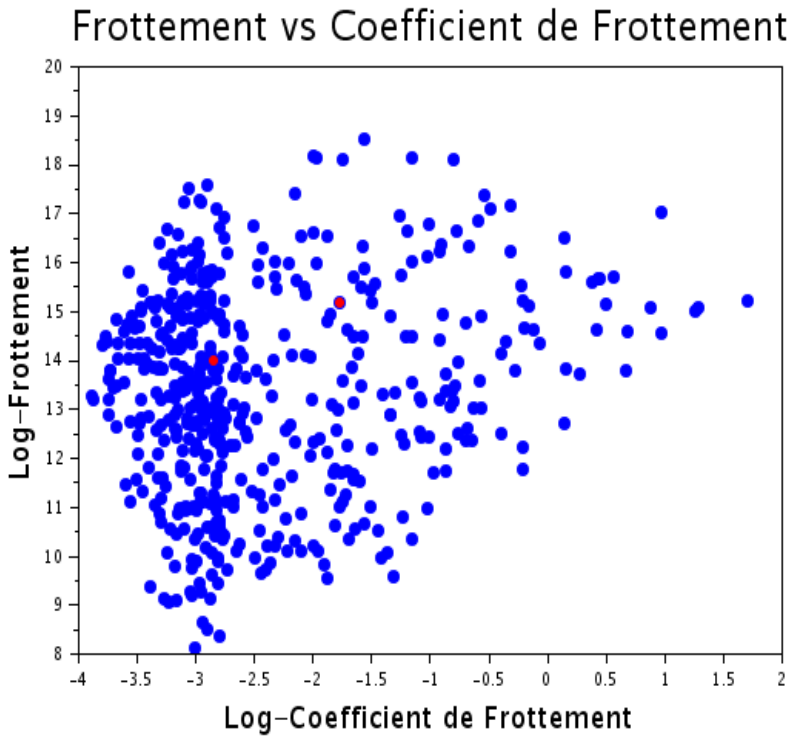
- Le comportement du modèle de frottement a été étudié en faisant varier les paramètres $\ln(\mu, \rho, D_h, U)$
  - $\rightarrow$ analyse exploratoire en loi *Log-Uniforme*
- L'échantillon sur les paramètres (μ, ρ, D_h, U) a été spécifié par plans d'expériences de type LHS MaxiMin (Log-Uniforme)
  - un plan LHS (Latin Hypercube Sampling) est aléatoire $\rightarrow$ risque de non couverture de l'espace d'exploration (grande discrépance)
  - LHS MaxiMin : plusieurs tirages LHS, on retient celui qui maximise la distance minimale entre points (faible discrépance)
- Modèle du krigeage
  - fonction de corrélation Matèrn 7/2
  - moyenne linéaire : $\beta_0 + \beta_1 \ln \mu + \beta_2 \ln \rho + \beta_3 \ln D_h + \beta_4 \ln U$
  - bruit de mesure supposé gaussien centré de variance inconnue à estimer (bon *numériquement*)
  - Estimation de ln(Frottement)

Différents cas simulés

1. Erreur de calcul du frottement en 2 points
   - 500 calculs réalisés puis permutation de 2 valeurs calculées
2. Légères discontinuités dans le calcul du frottement
   - ajout d'un bruit gaussien sur les 500 calculs du frottement
3. Erreur sur un paramètre du coefficient de frottement
   - erreur sur la valeur de Coef $= 0.3598984 \rightarrow 0.7598984$
   - 15 calculs en ne faisant varier que le débit U
4. Erreur sur un paramètre du coefficient de frottement
   - erreur sur la valeur de Coef $= 0.3598984 \rightarrow 0.7598984$
   - 625 calculs en faisant varier les 4 paramètres (μ, ρ, D_h, U)

Cas 1 - Erreur sur 2 calculs par permutation

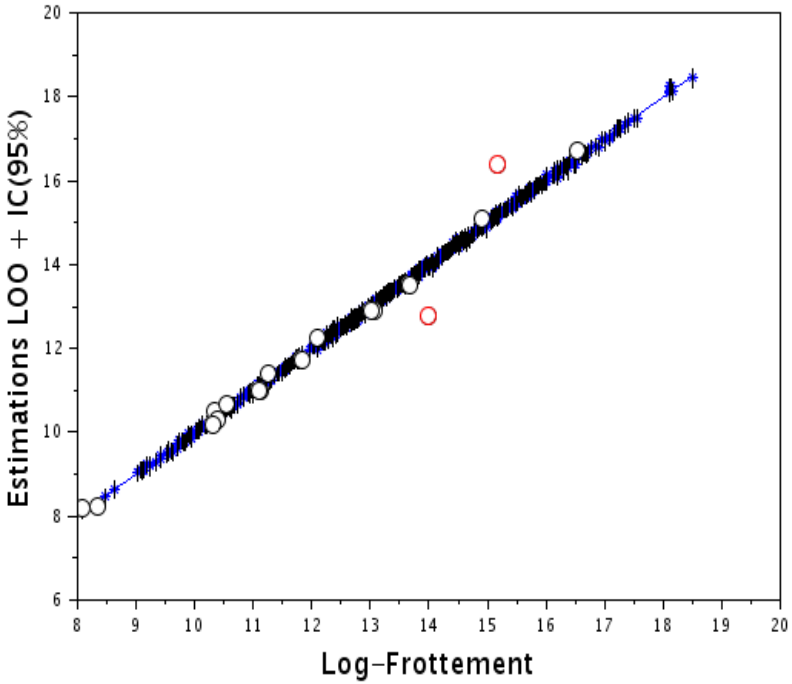


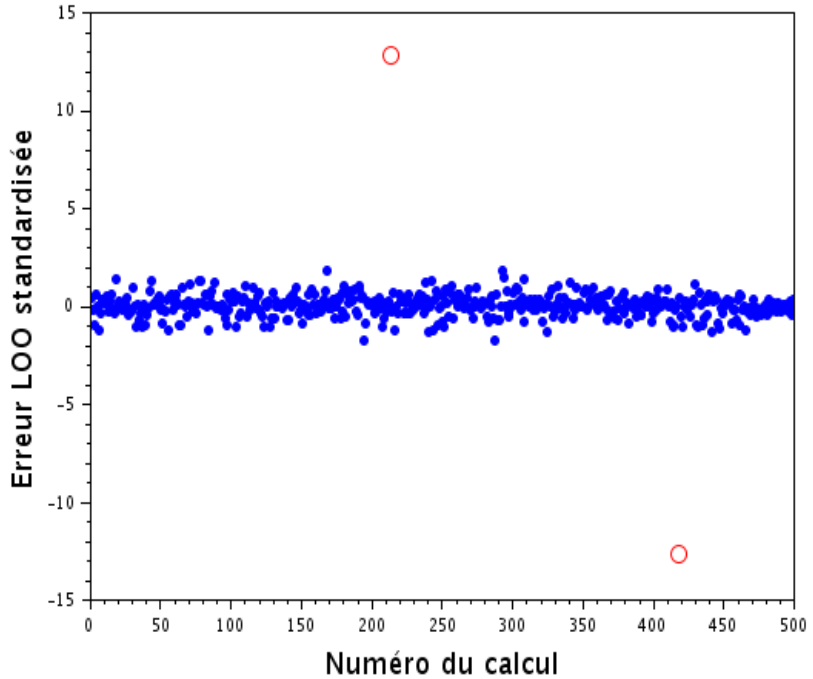


Visualisation des 2 calculs erronés en rouge, difficilement discernables dans $\mathbb{R}^4 \times \mathbb{R}$

Cas 1 - Erreur sur 2 calculs par permutation

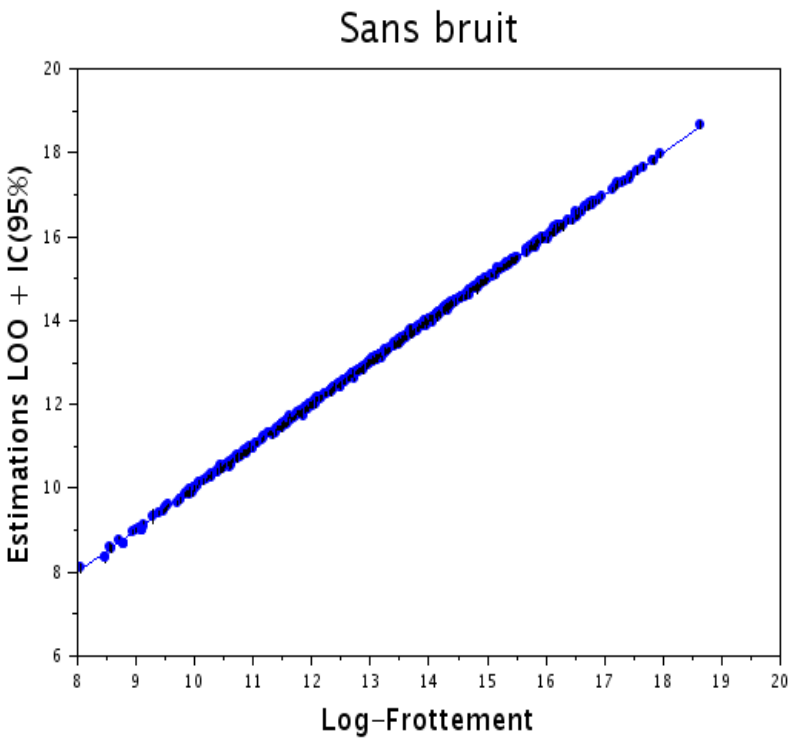
Estimations par LOO (gauche) + erreurs LOO standardisées (droite)

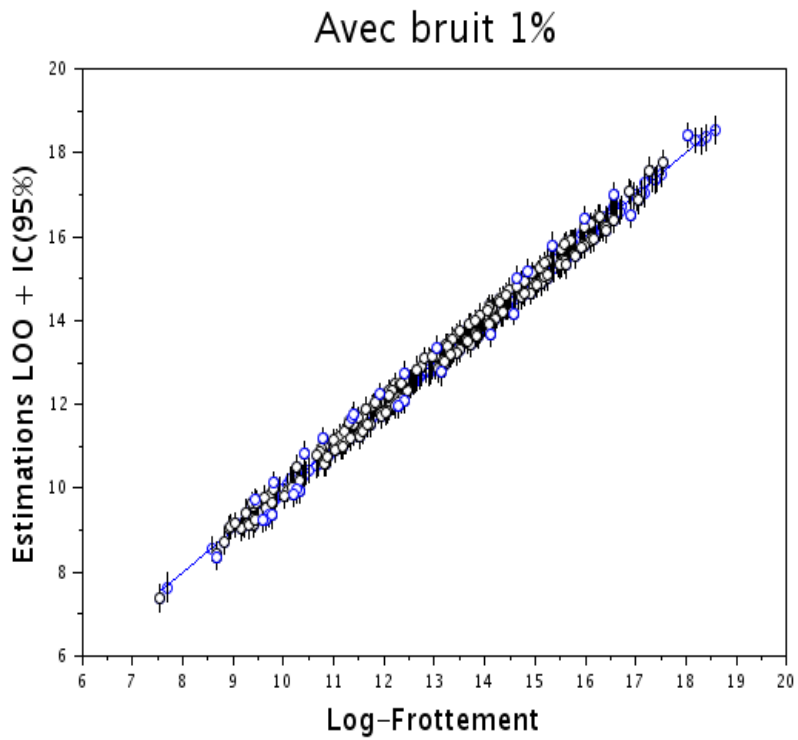




Détection des 2 *outliers* par comparaison Calculs-Estimations et par analyse de l'erreur LOO standardisée (2 écarts-types > 10)

Cas 2 - Ajout d'un bruit sur le calcul du frottement

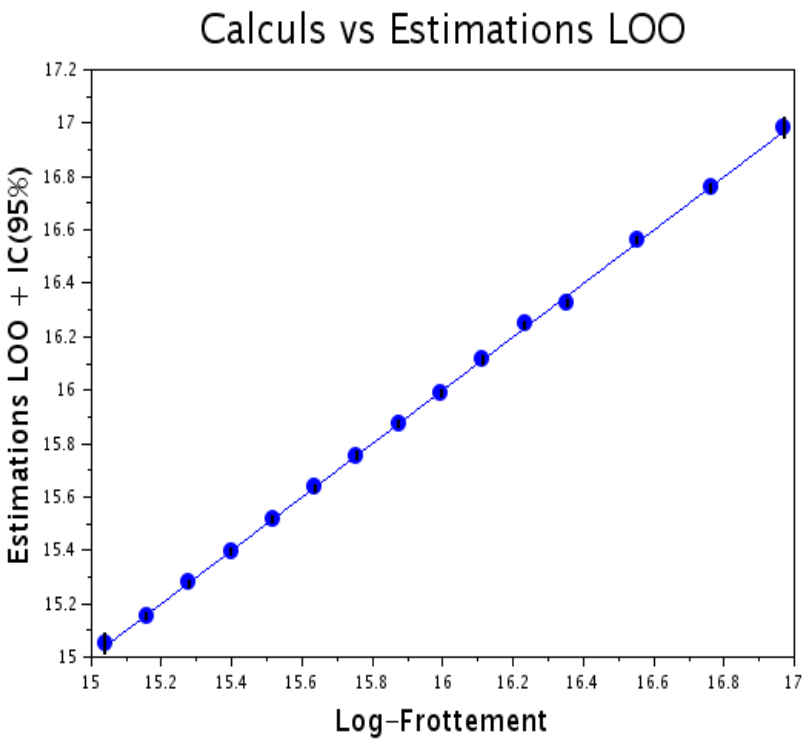




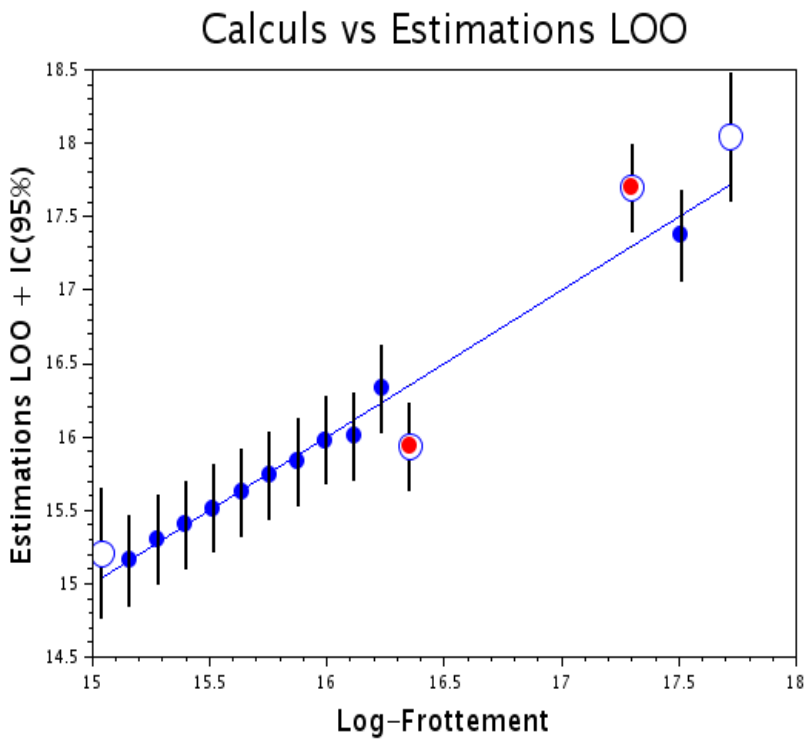
écart-type bruit additif (% de la moyenne)	écart-type estimé
0% $\rightarrow$ 0	0.0205725
1% $\rightarrow$ 0.1334038	0.1416509
2% $\rightarrow$ 0.2668077	0.2747681

Cas 3 - Erreur sur Coef - Débit variable

Modèle exact



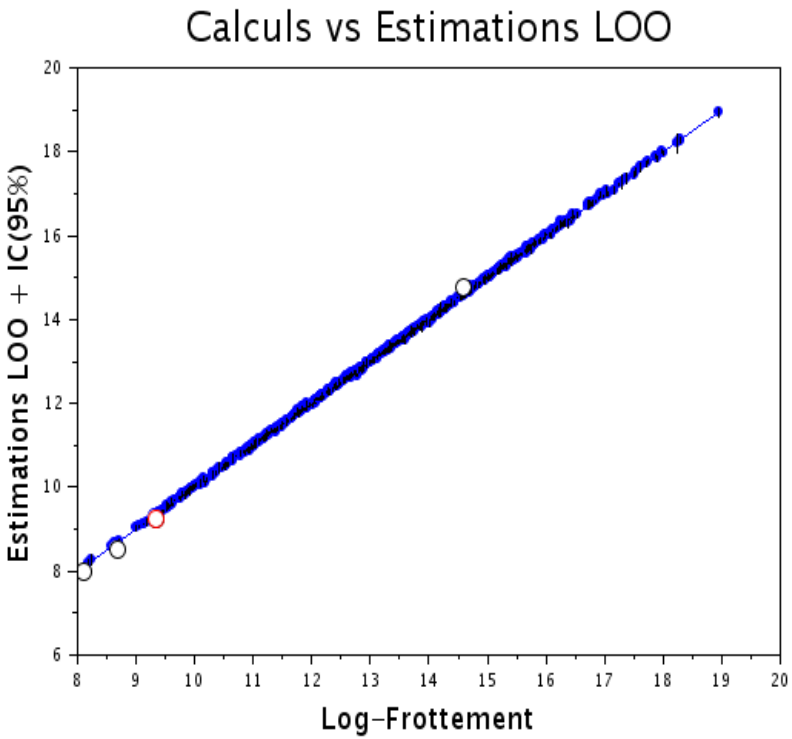
Erreur sur Coef



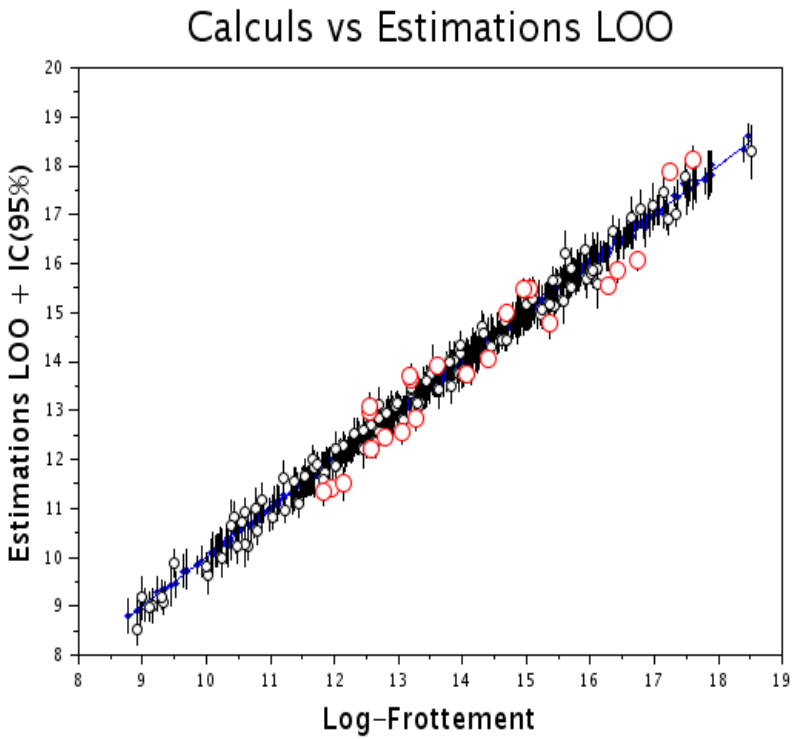
- erreur relative $> 1\%$
- erreur relative $> 1\%$ ET $|\epsilon_{loo}| > 1.96\sigma_{loo} \rightarrow$ [IC à 95%]

Cas 4 - Erreur sur Coef - Tous les paramètres varient

Modèle exact



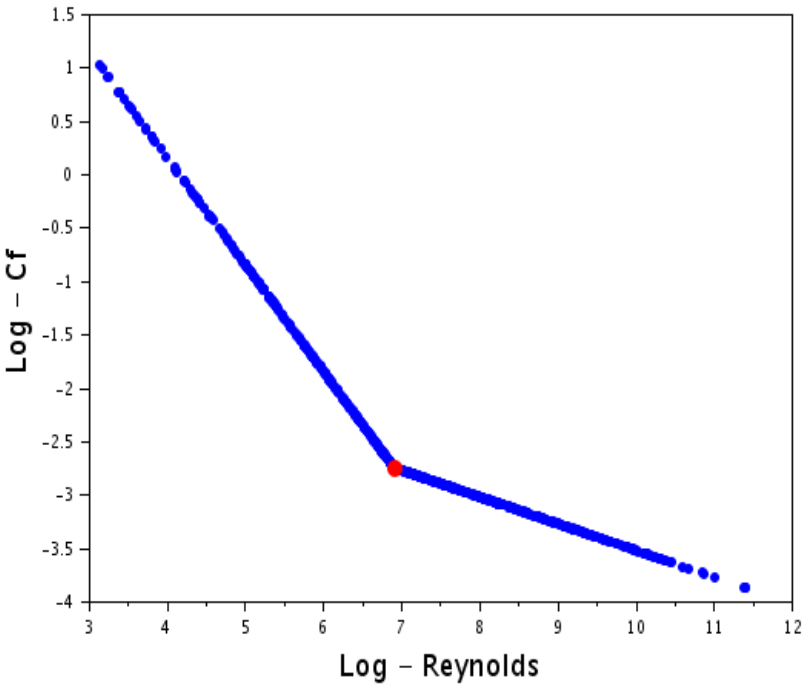
Erreur sur Coef



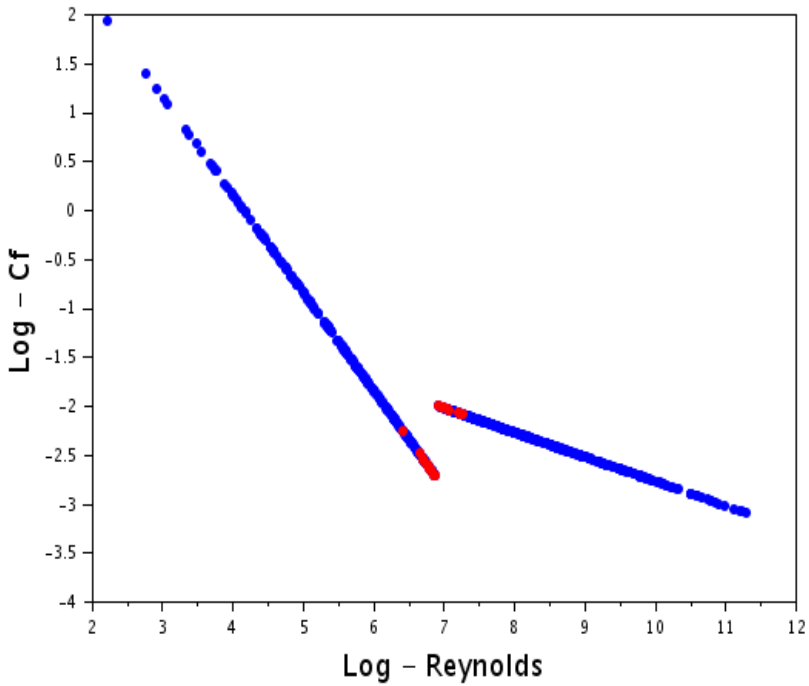
- erreur relative $> 1\%$
- erreur relative $> 1\%$ ET $|\epsilon_{loo}| > 3\sigma_{loo}$

Cas 4 - Erreur sur Coef - Tous les paramètres varient

Modèle exact



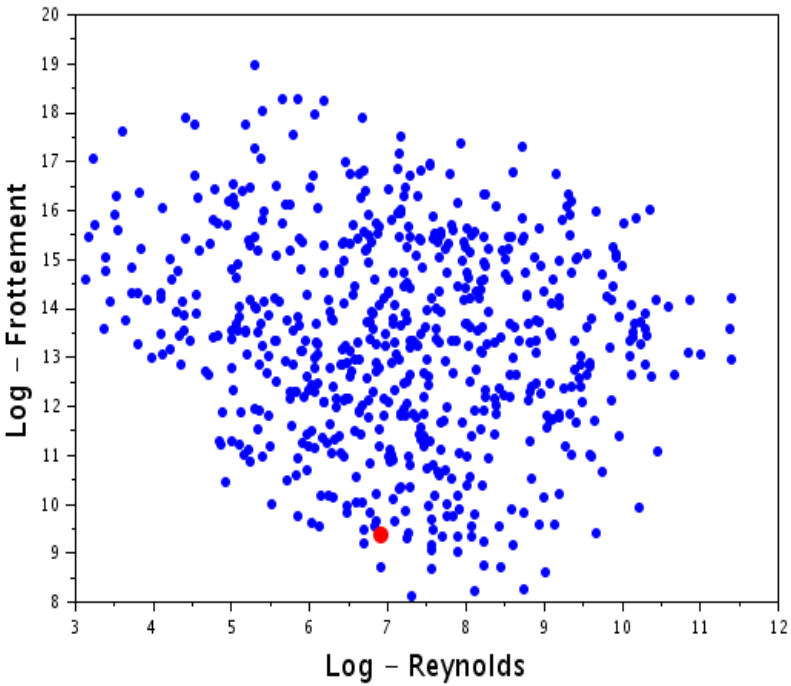
Erreur sur Coef



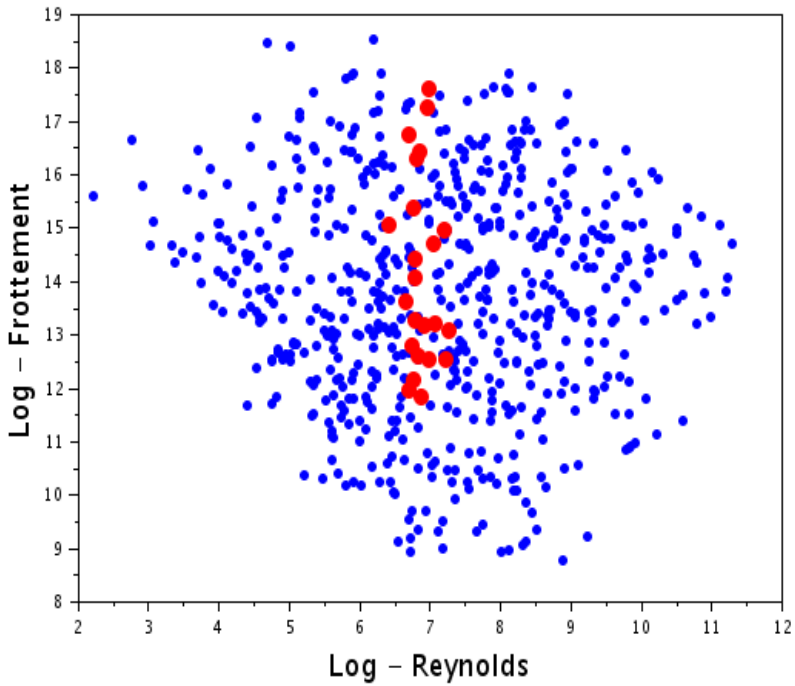
- erreur relative $> 1\%$ ET $|\epsilon_{loo}| > 3\sigma_{loo}$

Cas 4 - Erreur sur Coef - Tous les paramètres varient

Modèle exact



Erreur sur Coef



- erreur relative $> 1\%$ ET $|\epsilon_{loo}| > 3\sigma_{loo}$

Stratégie de détection des *outliers*

- Par krigeage, détecter les éventuels *outliers* en analysant la distribution statistique des erreurs LOO et celle des erreurs LOO standardisées
- Remarque : des cas où la variance prédictive est petite peuvent entraîner de grandes valeurs sur les erreurs LOO standardisées
  - ne retenir que les erreurs LOO (relatives) suffisamment importantes

Algorithm 2 Méthode de détection des outliers

Se fixer une ErreurRelative maximale
Se fixer un Quantile de la loi normale centrée réduite
Outlier_level_1 = find $(|\epsilon_{loo,i})| > y_{obs,i} \times$ ErreurRelative)
Outlier_level_2 = Outlier_level_1 $\cap$ (find $(|\tilde{\epsilon}_{loo,i}| >$ Quantile)

Plate-forme URANIE du CEA/DEN

- Pour vérifier vos codes ou analyser le comportement de vos modèles

  → utiliser sans modération la plate-forme URANIE

- en mode exploration paramétrique suivi d'un krigeage
- en ne faisant travailler que vos processeurs de calcul

  → avec le soutien de l'UTF du DEN/DM2S/STMF/LGLS

- URANIE

  → open source

  → http ://sourceforge.net/projects/uranie


Thomas J. Santner, Brian J. Williams, William I. Notz, The Design and Analysis of Computer Experiment, *Springer Series in Statistics, 2003*.


Carl E. Rasmussen, Christopher K.I. Williams, Gaussian Processus for Machine Learning, *MIT Press, 2006*.


François Bachoc, Thèse, publications, HDR dans le domaine des processus gaussiens. *https ://www.math.univ-toulouse.fr/ fbachoc/*.


Jean-Marc Martinez, Tutorial du krigeage dans URANIE, *Rapport CEA, DEN/DANS/DM2S/STMF/LGLS/NT-15-005/A*.

Cas 3 - Erreur sur Coef - Débit variable