Benchmark « construction de métamodèles prédictifs »
GdR MASCOT NUM

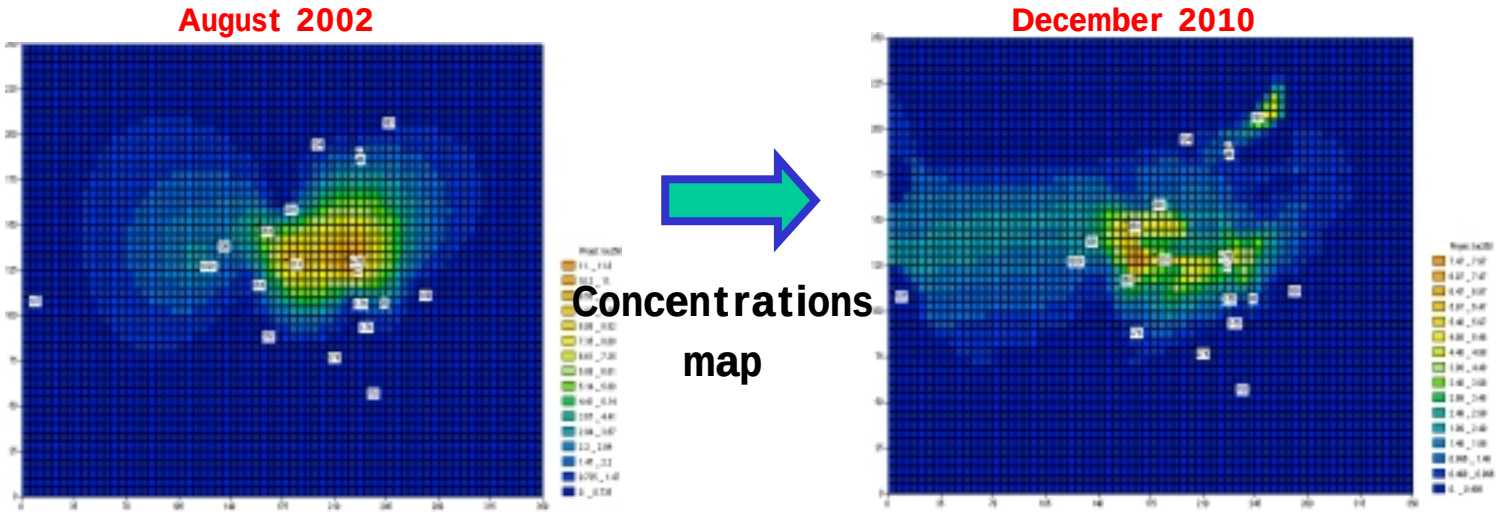
Exemple du CEA Cadarache – Données MARTHE

B. Iooss et A. Marrel

Avril 2008

Présentation du cas d'étude

En 2005, le CEA et l'Institut Kurchatov (Russie) ont collaboré au développement d'une modélisation du transport de strontium 90 (^{90}Sr) en milieu poreux saturé en eau, pour le cas d'un site de stockage temporaire de déchets radioactifs (STDR) à Moscou. Le but principal était de modéliser le transport de ^{90}Sr à des fins prédictives entre 2002 (où le terme source était connu) et 2010 afin de déterminer le degré de contamination potentielle de la nappe. La simulation numérique du transport de ^{90}Sr, espèce la plus mobile du terme source, dans l'aquifère supérieur du site a été réalisée à l'aide du code de calcul MARTHE (BRGM). La figure ci-dessous illustre l'évolution du panache de concentration en ^{90}Sr sur le site. La figure de gauche correspond au champ de concentration initiale (déduit de mesures), la figure de droite correspond au champ de concentration calculé par le code. Les petits rectangles blancs correspondent à l'emplacement sur le site des différents piézomètres (sondes poreuses permettant de prélever l'eau en profondeur et de relever le niveau de la nappe).



Afin d'identifier les paramètres d'entrée du code les plus influents sur le résultat du calcul, des méthodes statistiques d'analyses d'incertitudes et de sensibilité ont été utilisées. Vu la complexité du modèle numérique et son temps de calcul élevé, une phase intermédiaire de construction de métamodèles sur un nombre restreint de simulations du code a dû être mise en œuvre.

20 paramètres d'entrée scalaires du modèle numérique ont été considérés dans l'analyse d'incertitudes. Le tableau suivant synthétise les types de distribution des données et les intervalles de distribution de chacun de ces 20 paramètres. La loi de Weibull utilisée est la suivante :

$$f(x) = \alpha\beta^{-\alpha} x^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right),$$

où α est le paramètre de forme, et β le paramètre d'échelle de la loi.

	Paramètres d'entrée	Indicateur	Type de distribution	Intervalle ou paramètres de distribution
1	Perméabilité couche 1	per1	Uniforme	1 - 15
2	Perméabilité couche 2	per2	Uniforme	5 - 20
3	Perméabilité couche 3	per3	Uniforme	1 - 15
4	Perméabilité zone 1	perz1	Uniforme	1 - 15
5	Perméabilité zone 2	perz2	Uniforme	1 - 15
6	Perméabilité zone 3	perz3	Uniforme	1 - 15
7	Perméabilité zone 4	perz4	Uniforme	1 - 15
8	Dispersivité longitudinale couche 1	d1	Uniforme	0,05 - 2
9	Dispersivité longitudinale couche 2	d2	Uniforme	0,05 - 2
10	Dispersivité longitudinale couche 3	d3	Uniforme	0,05 - 2
11	Dispersivité transversale couche 1	dt1	Uniforme	0,01*d1 - 0,1*d1
12	Dispersivité transversale couche 2	dt2	Uniforme	0,01*d2 - 0,1*d2
13	Dispersivité transversale couche 3	dt3	Uniforme	0,01*d3 - 0,1*d3
14	Coefficient de partage volumique couche 1	kd1	Weibull	α=1.1597, β=19.9875
15	Coefficient de partage volumique couche 2	kd2	Weibull	α=0.891597, β=24.4455
16	Coefficient de partage volumique couche 3	kd3	Weibull	α=1.27363, β=22.4986
17	Porosité de toutes les couches	poros	Uniforme	0,3 - 0,37
18	Infiltration type 1	i1	Uniforme	0 - 0,0001
19	Infiltration type 2	i2	Uniforme	i1 - 0,01
20	Infiltration type 3	i3	Uniforme	i2 - 0,1

20 variables de sortie ont été considérées. Elles correspondent aux concentrations calculées aux emplacements de chaque piézomètre.

Le coût en temps de calcul du modèle MARTHE ne nous a permis de faire que 300 simulations. Il s'agit de la combinaison de 3 plans LHS réalisés successivement, indépendamment et comprenant chacun 100 simulations. Pour le calcul des indices de sensibilité de Sobol, une phase de construction d'un métamodèle associé à chaque sortie a été décidée. Celle-ci s'est révélée particulièrement difficile, essentiellement du fait de la faible taille de l'échantillon (300) au vu de la dimension des entrées (20) et de certains effets de seuil et autres non linéarités présents dans le modèle hydrogéologique. Il peut être également intéressant d'uniformiser les entrées (en appliquant la fonction de répartition inverse) afin d'obtenir une répartition uniforme des points dans l'espace des entrées.

A l'issue de nos premières études, il s'avère que certaines sorties n'ont pas d'intérêt physique (par exemple si les concentrations prédites sont très faibles du fait d'un piézomètre situé en dehors du panache) ou d'intérêt statistique (par exemple si les concentrations sont bien prédites par un modèle linéaire). Certaines sont également fortement corrélées entre elles (piézomètres proches les uns des autres). On restreint donc le problème à l'étude des 10 variables de sortie suivantes :

p102K, p104, p106, p2-76, p29K, p31K, p35K, p37K, p38, p4b

Objectifs du benchmark

Il s'agit de construire le métamodèle le plus prédictif possible vis-à-vis des variables de sortie dans tout l'espace des paramètres d'entrée (défini par leur distribution). Il est possible de construire un métamodèle par variable de sortie, mais aussi d'utiliser des métamodèles multiples (modélisant plusieurs sorties à la fois).

Le code de calcul n'étant plus accessible, on restreint l'étude à l'utilisation de la base des 300 simulations. Pour la validation, nous proposons à l'utilisateur de balayer les 300 simulations par une validation croisée. A chaque validation, il devra prendre une base de test de 1 à 50 points maximum (au libre choix de l'utilisateur en fonction du coût de construction du métamodèle utilisé).

Les critères de validation génériques et intéressants pourraient être les suivants :

- le coefficient de prédictivité (que l'on nomme Q_2), qui correspond à un R^2 calculé sur la base de test ;
- l'erreur moyenne en valeur absolue (moyenne des valeurs absolues des résidus de la base de test) ;
- le biais des résidus de la base de test ;
- le maximum des résidus de la base de test ;
- la distribution des résidus de la base de test ;
- des critères plus robustes à d'éventuels « outliers » (utiles s'il y a quelques résidus très élevés dans la base de test) : par exemple le biais géométrique $\mathrm{MG} = \exp\left[\mathrm{E}(\ln X) - \mathrm{E}(\ln Y)\right]$ et la variance géométrique $\mathrm{VG} = \exp\left\{\mathrm{E}\left[(\ln X - \ln Y)^2\right]\right\}$ où X et Y sont les deux échantillons que l'on compare (les sorties du code de calcul et les sorties prédites par le métamodèle).

D'autres critères de validation peuvent être bien entendu proposés par les participants.

Nota

Il est clair que cet exercice n'a pas valeur de jugement sur la qualité de tel ou tel métamodèle (chacun étant fortement dépendant du choix du plan d'expériences et du modèle physique sous-jacent), mais il nous semble qu'il illustre bien les pratiques industrielles : les expériences n'ont pas forcément été faites dans l'objectif de construire un métamodèle particulier (ici une propagation d'incertitudes était désirée avant toute chose), le plan choisi peut être loin d'être optimal (ici il s'agit d'un échantillon issu de trois échantillons indépendants de type LHS) et le code de calcul n'est plus disponible.

Mise en œuvre

Tous les détails du scénario sont disponibles dans la publication de Volkova et al. (2008). Une mise en oeuvre possible est décrite dans Marrel et al. (2008). Contacter B. Iooss (bertrand.iooss[at]cea.fr) pour tout renseignement complémentaire.

Résultats

Nous souhaiterions que les résultats nous soient communiqués : bertrand.iooss[at]cea.fr

Remerciements

On remercie E. Volkova (Institut Kurchatov) pour l'autorisation d'utiliser ses données.

Références

A. Marrel, B. Iooss, F. Van Dorpe and E. Volkova, An efficient methodology for modeling complex computer codes with Gaussian processes, *Computational Statistics and Data Analysis*, 52:4731-4744, 2008.

E. Volkova, B. Iooss and F. Van Dorpe, Global sensitivity analysis for a numerical model of radionuclide migration from the "RRC" Kurchatov Institute radwaste disposal site, *Stochastic Environmental Research and Risk Assessment*,22:17-31, 2008.