Société de Calcul Mathématique SA
*Outils d'aide à la décision*

# Amélioration d'outils

# pour la Sûreté Nucléaire

## Comparaison entre le krigeage et l'EPH

Rapport adressé à

l'IRSN
*(à l'attention de M. Yann Richet)*

par la

Société de Calcul Mathématique SA

*en application de la commande EX10/12022486, notifiée le 16/04/2015*

rédaction Gottfried Berton, Jean-Baptiste Rouault

# Résumé Opérationnel

Pour reconstituer les données manquantes, et évaluer les situations où aucune mesure n'a été faite, l'IRSN utilise actuellement un algorithme appelé EGO, qui fait appel à une technique standard appelée "krigeage". Lors d'un travail précédent (commande IRSN EX10 / 32001814 du 30.05.2014), la SCM a comparé la technique de krigeage et l'EPH (Experimental Probabilistic Hypersurface) sur quelques situations-types, fournies par l'IRSN. A la demande de l'IRSN, nous abordons un certain nombre de questions techniques plus approfondies, pour étudier les caractéristiques de l'EPH.

## Principe du krigeage et de l'EPH

Le krigeage intègre la distance ainsi que le degré de variation des données, afin d'estimer les valeurs en des points inconnus. Chaque observation est interprétée comme la réalisation d'une variable aléatoire. A chaque point du domaine est associée une variable aléatoire. L'ensemble de ces variables aléatoires forme un processus aléatoire. Le krigeage comporte deux étapes :

- l'estimation d'un modèle de dépendance entre les variables aléatoires de ce processus : elle est basée sur la variabilité des données. Ce modèle est fonction de la distance : deux points proches prennent une valeur semblable, alors que deux points éloignés sont indépendants ;

- la construction pour chaque point, d'une combinaison linéaire des données intégrant la dépendance estimée à l'étape précédente.

La construction de l'EPH peut être vue comme la propagation de l'information provenant des points d'observation vers des points inconnus. La propagation est basée sur un principe d'entropie, qui est une fonction croissante avec la distance au point de mesure.

Le résultat de l'EPH est donné sous la forme d'un ensemble de lois de probabilité dont la variance est maximale à entropie fixée. Les densités sont des fonctions de Dirac aux points de mesure, et deviennent de moins en moins concentrées lorsque l'on s'en éloigne.

A la fin du processus, les lois individuelles sont recombinées en fonction de la distance du point à reconstruire avec chaque mesure. L'interpolation finale est l'espérance de cette loi.

Le présent rapport reprend la division en Lots prévue dans la proposition technique et financière :

## Lot 1. Publication sur l'archive internet du langage R des précédents développements

Afin de valoriser les réalisations logicielles du contrat précédent, nous les publions sur l'archive centralisée du logiciel R : le CRAN. Nous publions l'implémentation de l'EPH dans sa version la plus générale. Nous avons pour cela constitué un "package" regroupant les sources, la documentation du code, et les exemples d'utilisation. Le logiciel R fournit un programme permettant de vérifier :

- que le package est correctement structuré,

- que le formalisme pour la documentation est respecté,

- et que les exemples fonctionnent.

Pour être accepté pour publication, l'exécution de ce programme ne doit produire aucune erreur ou avertissement. Nous avons lancé le programme sur notre package EPH avec succès : aucune erreur ni avertissement n'est renvoyé.

## Lot 2. Prise en compte des incertitudes

Les différents types d'incertitudes portent sur la position des mesures, sur leur valeur et sur la position du point à estimer.

L'un des avantages de l'EPH est que la prise en compte de ces incertitudes est très simple à réaliser. Lorsque les données prennent des valeurs déterministes, le résultat renvoyé par l'EPH pour chaque point est une loi de probabilité. Il est possible de construire l'EPH pour différentes combinaisons des paramètres d'entrée en effectuant un tirage aléatoire à partir des lois de probabilité de chaque paramètre affecté d'une incertitude. Il suffit alors de moyenner les lois de probabilités obtenues. La loi ainsi construite est plus dispersée que si on utilise l'EPH avec des valeurs d'entrée déterministes. La surface reconstruite est plus lisse.

Dans le krigeage, la dépendance spatiale est modélisée par un variogramme. L'utilisation d'un variogramme moyen permet de prendre en compte les incertitudes sur la localisation des observations. Pour le construire, il faut :

- générer aléatoirement un nombre fini de positions possibles pour chaque point d'observation ;

- construire un variogramme pour chaque répartition obtenue ;

- moyenner les variogrammes.

Les incertitudes sur la valeur des observations sont prises en charge par un variogramme spécifique permettant de modéliser un écart entre les observations et le processus.

L'EPH renvoie une densité de probabilité en chaque point ; il donc est facile d'obtenir une incertitude sur le résultat obtenu. On peut extraire de cette loi une probabilité de dépassement de seuil, des quantiles ou un intervalle de confiance. Cet intervalle est bien plus large que celui du krigeage.

Les versions de base du krigeage ne permettent d'obtenir en sortie qu'une variance d'estimation. Elle ne dépend pas de la valeur des observations, mais seulement de leur position et

ne reflète que la densité des mesures environnantes : plus il y a de mesures proches, plus la variance d'estimation est faible.

Pour construire un intervalle de confiance à partir de cette variance, on suppose que les variables aléatoires suivent une loi gaussienne, ce qui n'est pas toujours pertinent. De plus l'intervalle de confiance construit ne dépend pas des valeurs des mesures mais que de leurs positions.

Le krigeage par indicatrice permet reconstruire la loi de probabilité conditionnée aux observations.

### Lot 3. Convergence de l'algorithme lorsque le nombre de points de mesure augmente

Lorsque la fonction à reconstruire est continue, le krigeage et l'EPH convergent vers la valeur vraie en tout point quand le nombre de mesures tend vers l'infini.

### Lot 4. L'écart entre valeur prédite et valeur réelle

Nous avons défini deux mesures pour évaluer la qualité de l'interpolation :

- L'écart moyen entre la valeur estimée et réelle.
- L'écart quadratique moyen entre valeur estimée et réelle.

Nous avons comparé ces mesures pour le krigeage et pour l'EPH dans plusieurs situations :

### Prise en compte de la variabilité des données

Le krigeage présente l'avantage d'intégrer le degré de variation des données. Cependant il ne parvient pas à estimer correctement la variabilité des données lorsque :

- l'information est pauvre ;
- les données varient fortement.

Dans ces deux situations, l'algorithme suppose que toutes les variables aléatoires du processus à reconstruire sont indépendantes, et l'estimation obtenue est constante, ce qui n'est pas acceptable. L'utilisateur doit alors corriger les paramètres de dépendance spatiale du krigeage.

La faible quantité d'information ne pose aucun problème dans l'EPH. Lorsque la variabilité des données est inconnue, l'EPH ne cherche pas à inventer cette information, contrairement au krigeage. L'EPH n'incorpore que l'information existante et ne fait aucune hypothèse artificielle.

Cependant lorsque les données varient fortement, l'interpolation réalisée par l'EPH est mauvaise. La figure ci-dessous montre la reconstruction faite par l'EPH en rouge, les observations sont représentées par des triangles bleus :

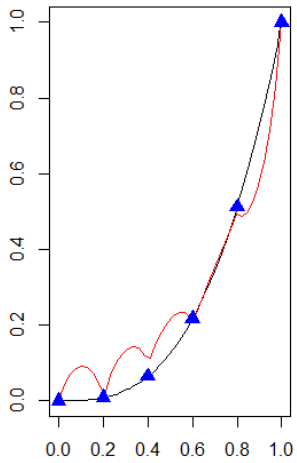


*Figure 1 : interpolation faite par l'EPH*

L'EPH attribue un poids trop important aux données éloignées du point à estimer. Le concept même de l'EPH est d'ajouter le moins d'information arbitraire possible, donc la variabilité des données n'est pas prise en compte. Les poids sont les mêmes quelle que soit la variation. Ici la variation est forte, donc les données situées à une distance de 0.8 sont très différentes. Les observations situées à cette distance devraient avoir un poids très faible, ce qui n'est pas le cas dans l'EPH. Il en résulte une interpolation bosselée qui est insatisfaisante.

Dans cet exemple, l'interpolation faite par krigeage est convenable :

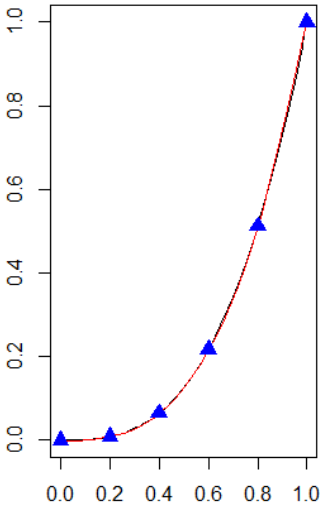


*Figure 2 : interpolation faite par le krigeage*

Pour un nombre suffisant de données, le krigeage a une meilleure connaissance de la variabilité. Il s'adapte alors mieux aux fortes variations, puisqu'il donne dans ce cas un poids plus fort aux données proches.

### Groupement des observations

Lorsque les données contiennent des observations groupées entre elles, l'EPH donne trop d'importance à ce groupe par rapport aux données isolées, ce qui fausse l'interpolation. Cela est un inconvénient majeur de l'EPH.

Le krigeage gère très bien cette situation puisqu'il donne un poids identique à un groupe de mesure et à une observation isolée s'ils sont à égale distance du point à reconstruire. Le krigeage donne un poids quasiment nul à toutes les observations du groupe excepté à une.

### Présence de données aberrantes

Lorsqu'il y a des valeurs aberrantes parmi les données, celles-ci entraînent une large modification de l'interpolation sur la quasi-totalité du domaine, aussi bien pour le krigeage que pour l'EPH. Cependant lorsque de nombreuses observations sont disponibles, l'EPH obtient de meilleures performances que le krigeage ; la valeur aberrante impacte l'interpolation sur une zone plus restreinte du domaine.

En présence d'une valeur aberrante, le krigeage ne parvient pas à estimer la dépendance spatiale des données et produit une surface constante tout à fait insatisfaisante.

Cependant la détection de données aberrante est une étape de préparation des données qui se fait avant l'utilisation de l'algorithme d'interpolation. Ce n'est pas le rôle d'une méthode de reconstruction de données que de détecter des données aberrantes. Il est donc normal que ces données soient intégrées dans l'interpolation au même titre que les autres observations.

Il existe toutefois une version du krigeage robuste aux données extrêmes nommée krigeage robuste, permettant de donner un faible poids aux données aberrantes.

## Lot 5. Qualité des procédures numériques

Pour réaliser le calcul des poids, le krigeage inverse une matrice de covariance. Lorsque le conditionnement de cette matrice est très mauvais, le krigeage peut présenter des instabilités numériques qui faussent l'estimation. Une simple perturbation sur la position des mesures entraîne une interpolation très différente. Le conditionnement est mauvais lorsque :

- des données sont très proches les unes des autres ;
- le paramètre du modèle de dépendance spatiale est très élevé.

De plus, lorsque le conditionnement est mauvais, des erreurs d'arrondis peuvent s'accumuler au cours de l'inversion de la matrice et causer une mauvaise interpolation. La matrice peut également ne pas être inversible et dans ce cas le logiciel R renvoie un message d'erreur.

Les poids dans le krigeage peuvent être inférieurs à zéro, contrairement à ceux utilisés par l'EPH. Cela peut engendrer une estimation supérieure à l'observation la plus grande. Ceci est étrange, puisqu'on ne peut pas inventer des données nouvelles, qui n'ont jamais été observées par le passé. De plus, l'estimation donnée par le krigeage peut être négative, ce qui n'a pas toujours de sens, par exemple quand on estime une concentration.

Prenons l'exemple de l'estimation d'une concentration en minerai (en gramme par mètre cube). On dispose de quelques mesures de concentration. Le krigeage peut attribuer un poids inférieur à zéro à une ou plusieurs observations. L'estimation étant une somme pondérée des mesures, le résultat obtenu peut être inférieur à zéro à cause de ces poids négatifs. Or une concentration ne peut jamais être négative, cela n'a pas de sens physique.

Ces problèmes numériques et conceptuels sont bien évidemment exclus dans l'EPH, puisqu'il n'y a pas de paramètres à estimer ni de matrice à inverser. Le faible nombre de données ne pose aucun problème à l'EPH grâce à sa construction grossière. La forte variabilité des données n'engendre pas une interpolation aberrante dans l'EPH.

### Hypothèses de modèle

Le krigeage fait plusieurs suppositions sur le processus à reconstruire :

- la dépendance entre les variables aléatoires n'est fonction que de la distance ;
- le processus à reconstruire est gaussien.

Ces hypothèses peuvent engendrer de mauvais résultats lorsqu'elles ne sont pas vérifiées ; c'est-à-dire dans les cas suivants :

- la fonction à reconstruire est composée d'une partie constante et d'une partie très variable ;
- les données ne sont pas la réalisation d'un processus gaussien.

La méthode EPH ne fait pas de telles hypothèses de modèle ; ces deux situations ne posent donc pas de problème pour l'EPH.

### Complément. Vitesse d'exécution de l'algorithme

Le temps de calcul pour l'EPH est largement supérieur à celui du krigeage. Cela est principalement dû à la discrétisation de la plage de valeurs de sortie qui n'est pas nécessaire dans le krigeage. Par exemple, on discrétise la variable de sortie en 5000 valeurs entre 0 et 500, et on veut reconstruire 1600 points de la fonction de Branin à partir de 16 mesures. Pour chaque point à reconstruire, on calcule la loi de probabilité associée à chaque observation. Ce sont des lois gaussiennes centrées sur la valeur de l'observation, et elles sont plus ou moins concentrées en fonction de la distance de cette mesure au point à reconstruire.

Les lois de probabilité sont quasiment nulles sur une grande partie de la plage de valeur de sortie. Il n'est donc pas nécessaire de calculer les lois sur tout l'intervalle [0,500]. Nous avons amélioré le code de l'EPH par rapport à la version implémentée dans le premier travail afin de calculer les lois de probabilité seulement là où elles prennent des valeurs significatives. L'EPH met alors 133 secondes contre 9 millisecondes pour le krigeage.

## Conclusion

L'EPH est un modèle à "information minimale" qui n'utilise rien d'autre que les données existantes. Le krigeage fait au contraire de nombreuses hypothèses de modèle. Ce modèle intègre le degré de variabilité des données et fait des suppositions sur la nature de cette variabilité.

L'EPH possède les avantages suivants :

- les suppositions faites par le krigeage sont rarement vérifiées ;
- la variabilité des données n'est pas forcément connue, notamment lorsque peu d'information est disponible. L'EPH est donc plus efficace lorsque l'information est pauvre ;
- l'EPH renvoie une incertitude en sortie sous la forme d'une loi de probabilité, ce que ne permet pas le krigeage. La prise en compte des incertitudes sur les paramètres d'entrée est plus aisée avec l'EPH ;
- l'EPH est très simple à mettre en oeuvre. Les procédures numériques du krigeage sont plus complexes et nécessitent notamment l'inversion d'une matrice. Parfois, cette matrice est mal conditionnée, ce qui entraîne une mauvaise interpolation, voire une erreur lorsque la matrice est non-inversible ;

A l'inverse l'EPH possède quelques inconvénients :

- lorsque des données sont groupées, elle donne trop d'importance à ce groupe par rapport à une donnée isolée ;
- lorsque les données varient fortement, l'EPH donne un mauvais résultat ;
- le temps de calcul du krigeage est inférieur à celui de l'EPH.

L'EPH est conçue pour ne pas ajouter d'information arbitraire. Dans certaines situations spécifiques, elle peut donner de mauvais résultats, mais c'est parce qu'il y a une information spécifique qui doit être incorporée dans le modèle.

# Sommaire

Résumé Opérationnel ..... 2
Sommaire ..... 9
I. Introduction ..... 11
II. Présentation du krigeage ..... 11
A. Principe général ..... 11
B. Hypothèses de modèle du krigeage ..... 13
1. Stationnarité ..... 14
2. Linéarité de l'estimateur ..... 16
C. Les principales méthodes de krigeage ..... 19
1. Krigeage simple ..... 19
2. Le krigeage ordinaire ..... 20
3. Le krigeage universel ..... 20
III. Présentation de l'EPH ..... 21
IV. Publication sur l'archive internet du langage R des précédents développements ..... 22
V. La prise en compte des incertitudes sur les données ..... **Erreur ! Signet non défini.**
A. La prise en compte des incertitudes dans l'EPH ..... **Erreur ! Signet non défini.**
B. Prise en compte des incertitudes dans la méthode de krigeage **Erreur ! Signet non défini.**
1. Sur la position des observations ..... **Erreur ! Signet non défini.**
2. Sur la dépendance spatiale ..... **Erreur ! Signet non défini.**
3. Sur la valeur des observations ..... **Erreur ! Signet non défini.**
VI. La convergence de l'algorithme lorsque le nombre de points de mesure augmente ..... 36
A. Pour le krigeage ..... 37
B. Pour l'EPH ..... 39
VII. L'écart entre valeur prédite et valeur réelle ..... 41
VIII. Qualité des procédures numériques ..... 61
A. Qualité des procédures numérique dans le krigeage ..... 61
1. Mauvais conditionnement de la matrice ..... 61
2. Non-inversibilité de la matrice ..... 66
3. Négativité des poids ..... 67
B. Qualité des procédures numérique dans l'EPH ..... 67
IX. Prise en compte du degré de variation des données ..... 41
A. Forte variabilité des données ..... 42
1. Pour le krigeage ..... 42

2. Pour l'EPH ..... 43
B. Faible nombre d'observations ..... 47
1. Pour le krigeage ..... 47
2. Pour l'EPH ..... 50
X. Gestion des données groupées ..... 51
A. Gestion des données groupées par l'EPH ..... 51
B. Gestion des données groupées par le krigeage ..... 56
XI. Robustesse aux données aberrantes ..... 58
XII. Temps de calcul ..... 68
A. Pour le krigeage ..... 68
B. Pour l'EPH ..... 68
C. Comparaison des temps de calcul ..... 69
XIII. Références ..... 70
XIV. Annexe ..... 71
A. Démonstration de la linéarité de l'estimateur dans le cas gaussien ..... 71
B. Méthode de l'estimation du maximum de vraisemblance ..... 71
C. Le krigeage est un interpolateur exact ..... 72
D. Situations menant à un mauvais conditionnement ..... 72
E. Compléments au premier rapport ..... 73
F. Comportement de l'EPH en cas de répartition inégale des données ..... 75
G. Archivage de l'implémentation en R ..... 76
1. Vérification de l'archive ..... 76
2. Description ..... 77
3. Predict.Rd ..... 77
4. Eph.Rd ..... 79

# I. Introduction

L'IRSN effectue des recherches destinées à améliorer la Sûreté Nucléaire. Le présent développement concerne les règles fondamentales de sûreté relatives aux installations nucléaires de base autres que les réacteurs. La question se pose de savoir si le "design" de certaines configurations est suffisamment sûr. La difficulté tient au fait que, de manière générale, tous les paramètres ne sont pas connus, et ceux qui le sont apparaissent entachés d'une incertitude mal quantifiée. Comme l'IRSN l'a observé, le fait de "mailler" l'espace des paramètres ne conduit pas nécessairement à de bonnes estimations, car les situations "à risque" peuvent se trouver en dehors du maillage.

Pour reconstituer les données manquantes, et évaluer les situations où aucune mesure n'a été faite, l'IRSN utilise actuellement un algorithme appelé EGO, qui fait appel à une technique standard appelée "krigeage". Lors d'un travail précédent (commande IRSN EX10 / 32001814 du 30.05.2014), la SCM a comparé la technique de krigeage et l'EPH (Experimental Probabilistic Hypersurface) sur un certain nombre de situations-types, fournies par l'IRSN. La conclusion obtenue a été très claire : le krigeage donne de meilleurs résultats lorsque l'information disponible est abondante ; l'EPH donne de meilleurs résultats lorsque cette information est pauvre.

Ces conclusions étant encourageantes pour la mise en œuvre de l'EPH (les situations d'information pauvre sont fréquentes), l'IRSN nous a demandé d'aborder un certain nombre de questions techniques plus approfondies, pour étudier les caractéristiques de l'EPH.

# II. Présentation du krigeage

## *A. Principe général*

Le krigeage est une méthode d'interpolation probabiliste. Cet outil a été développé en premier lieu pour l'étude des gisements miniers en géostatistique. Il n'est pas possible de forer un terrain sur toute sa surface, on relève la concentration en minerai seulement en certains points du domaine. Le krigeage permet d'estimer la concentration en minerai sur tout le terrain à partir de ces points d'observation.

Le krigeage intègre la distance ainsi que le degré de variation des données, afin d'estimer les valeurs en des points inconnus.

Chaque observation $y_i$ est interprétée comme une réalisation d'une variable aléatoire $Y_i$. A chaque point du domaine est associée une variable aléatoire. L'ensemble de ces variables aléatoires forme un processus stochastique $\{Y_x, x \in R\}$. La dépendance entre les variables aléatoires de ce processus est fonction de la distance. En effet, deux points proches prennent une valeur semblable, à l'inverse pour deux points éloignés, détenir de l'information sur la variable aléatoire associée à l'un ne donne aucune information sur la variable aléatoire

associée à l'autre. Ces variables aléatoires sont donc indépendantes. Le krigeage s'effectue en deux étapes :

- modéliser le processus qui décrit le mieux les données et leur variabilité ;
- reconstruire chaque point comme une combinaison linéaire des observations. Les poids sont choisis de manière à approcher au mieux le processus précédemment estimé.

### Première étape : modélisation du processus

En fonction du type de krigeage utilisé, la dépendance entre deux variables aléatoires est modélisée soit par une fonction de covariance soit par un variogramme. Ce dernier représente l'écart moyen entre les variables aléatoires éloignées d'une distance $h$. Evidemment, plus la distance est grande, plus l'écart est grand.

En pratique on construit d'abord un variogramme expérimental à partir des observations. Supposons par exemple que l'on dispose de 25 observations disposées aléatoirement sur un domaine de dimension 50 par 50. A chaque couple d'observations $(y_i, y_j)$, $i < j$, $(i, j) \in [1, 25]$, on associe un point qui représente l'écart $(y_i - y_j)^2$ par rapport à la distance entre les observations (points noirs sur la Figure 3). A partir de ce nuage de points, on définit l'écart moyen en fonction de la distance (points rouges sur la Figure 3).

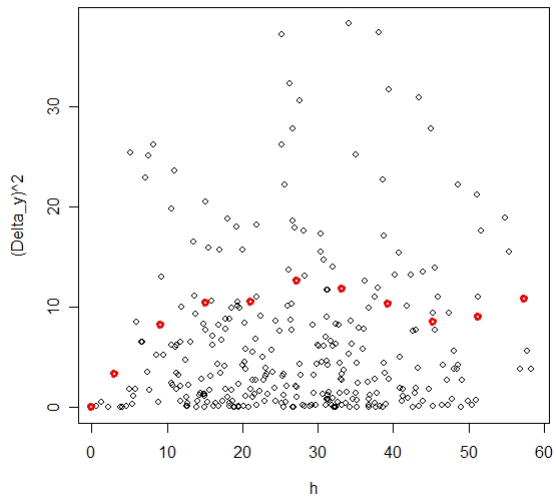


*Figure 3 : écart entre deux points en fonction de la distance : nuée variographique*

On ajuste ensuite un modèle (en bleu sur la Figure 4) sur ce variogramme expérimental. Cela permet d'obtenir une estimation de l'écart moyen quel que soit h.

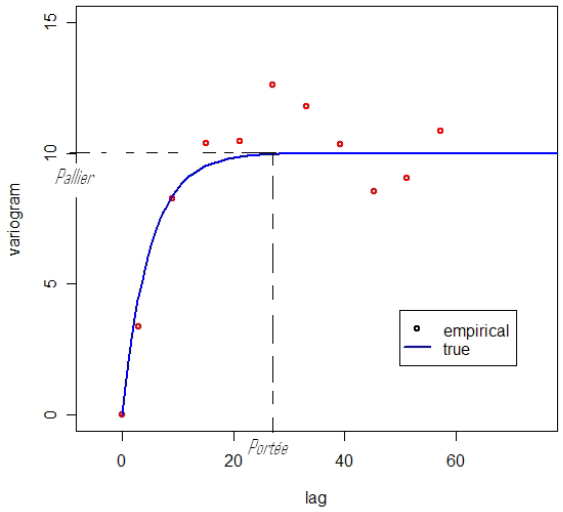


*Figure 4 : ajustement d'un modèle sur le variogramme expérimental*

Ce modèle comporte un paramètre important : la portée ; il représente une distance. Les variables aléatoires séparées d'au moins cette distance ont une dépendance minimale. Dans cet exemple, au-delà d'une distance de 25, les variables aléatoires sont considérées comme indépendantes.

On peut construire de la même manière un modèle de covariance afin d'estimer la covariance entre les variables aléatoires en fonction de leur distance. La covariance décroît avec la distance et tend vers 0. Une covariance nulle traduit une dépendance spatiale minimale.

**Deuxième étape : construction de l'estimateur**

Le modèle de dépendance spatiale estimé est utilisé pour définir un estimateur qui s'écrit comme une combinaison linéaire des observations. Les coefficients sont calculés de façon à minimiser la variance de l'erreur entre la variable aléatoire au point à reconstruire et la combinaison linéaire des variables aléatoires associées à chaque observation.

Les poids de cette combinaison linéaire sont propres à chaque point à reconstruire et dépendent :

- de la distance de ce point avec chaque observation ;
- de la distance des observations entre elles ;

## *B. Hypothèses de modèle du krigeage*

Le krigeage fait plusieurs hypothèses de modèle. Elles ne sont pas souvent vérifiées en réalité et nous montrons les problèmes qui se posent dans ce cas.

## 1. Stationnarité

Le krigeage suppose que la dépendance spatiale dépend uniquement de la distance séparant les variables aléatoires et pas de leur position. Pour tout couple de points distants d'une même longueur $h$, les covariances de ces couples de variables aléatoires sont les mêmes. Cette hypothèse est appelée "stationnarité de second ordre du processus".

Les principaux modèles de covariance sont :

- le modèle exponentiel $f(h) = \sigma^2 e^{-\frac{h}{\rho}}$ ;

- le modèle gaussien $f(h) = \sigma^2 e^{-\left(\frac{h}{\rho}\right)^2}$ ;

- le modèle de Matérn $f(h) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2}\nu h}{\rho}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2}\nu h}{\rho}\right)$.

où $K_{\nu}$ est une fonction spécifique appelée fonction modifiée de Bessel, $\sigma^2$ est la variance commune à chaque variable aléatoire à fixer, $h$ est la distance séparant deux variables aléatoires et $\rho$ est le paramètre de portée à fixer également :

- soit manuellement, après avoir analysé la variabilité des données ;

- soit par un algorithme à partir des données : par la méthode des moindres carrés ou par maximum de vraisemblance.

L'estimation du paramètre dépend fortement du jeu de données. Nous présenterons plus loin des situations-type où la mauvaise estimation de ce paramètre entraîne une mauvaise interpolation. Il faut donc avoir un avis critique sur l'estimation obtenue et régler manuellement les paramètres en fonction du jeu de données, ce qui rend l'utilisation du krigeage difficile.

Chacune de ces fonctions modélise une dépendance spécifique. Par exemple, dans le modèle exponentiel, lorsque les points sont suffisamment proches, la covariance décroît rapidement au fur et à mesure que la distance entre les points augmente. Cette croissance est plus douce dans le modèle gaussien. Les modèles de Matérn sont un compromis puisqu'ils permettent de régler cette pente avec le paramètre $\upsilon$. Dans le cadre de ce contrat nous utilisons ce modèle pour effectuer les tests.

L'hypothèse de stationnarité de second ordre n'est évidemment pas souvent vérifiée. Dans l'exemple du terrain de minerai, elle suppose une homogénéité de la variabilité des concentrations de minerai sur le terrain, ce qui n'est pas le cas en pratique. Il peut y avoir sur un même domaine deux régions dont l'une avec une concentration quasiment constante et l'autre avec une concentration très irrégulière. Illustrons cette situation par un exemple.

**Exemple**

On souhaite reconstruire la fonction $f(x)=1-10e^{-x^2/0.03}$ qui est tracée en noir sur le graphique ci-dessous. On reconstruit 20 points à partir de 7 mesures représentées par les triangles bleus. L'interpolation réalisée est affichée en rouge sur la figure suivante :

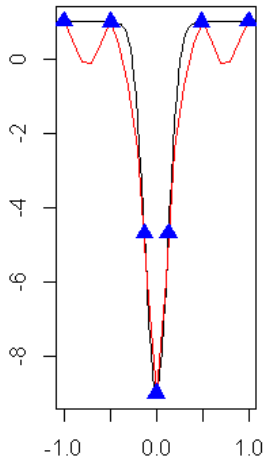


*Figure 5 : estimation de la portée par le logiciel R*

Dans cette première simulation, la portée est estimée par le logiciel R par la méthode du maximum de vraisemblance et elle est égale à 0,13. Ce coefficient est faible, car l'algorithme a détecté que les données proches les unes des autres varient fortement au milieu du domaine. Cependant, ce coefficient est trop faible pour pouvoir bien reproduire le comportement constant aux bords. En effet, lorsque la portée est faible, cela signifie que l'on estime un processus dont les variables aléatoires (y compris celles proches) sont indépendantes. On augmente alors manuellement la portée, l'interpolation est affichée sur la figure ci-dessous :

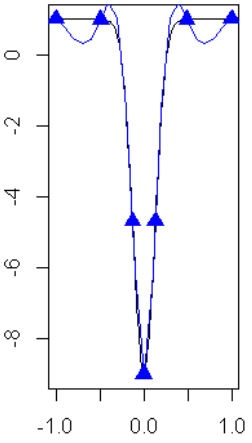


*Figure 6 : portée fixée manuellement à 0,4*

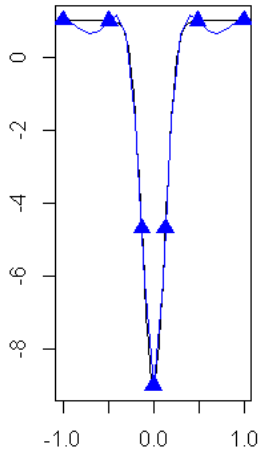


*Figure 7 : portée fixée manuellement à 0,3*

L'interpolation est meilleure. La portée est plus grande, ce qui signifie que les variables composant le processus à reconstruire sont dépendantes les unes des autres. Ainsi, le poids donné aux observations lointaines est plus fort. Lorsque l'on reconstruit un point situé sur les bords, les observations situées au milieu du domaine ont un poids fort. L'interpolation n'est donc pas constante sur les bords, comme elle le devrait. Il est donc impossible de reproduire un comportement constant aux bords lorsque la fonction est variable sur le milieu du domaine. Plus généralement, on ne peut pas reconstituer à la fois le comportement variable et constant des données avec la méthode du krigeage.

L'hypothèse de stationnarité de second ordre impose également que l'espérance de chaque variable aléatoire du processus est constante sur le domaine. Le variogramme n'impose pas cette contrainte, ce qui donne plus de liberté pour modéliser la dépendance spatiale. Le semi-variogramme est défini par $\gamma(h) = \frac{1}{2} Var(Y_{x+h} - Y_x)$.

L'avantage de cet outil est que, contrairement au cas stationnaire d'ordre 2, l'espérance de $Y_x$ n'a pas besoin d'être constante sur le domaine et peut être non finie. Dans ce cas le processus est dit stationnaire intrinsèque.

Les principaux modèles de variogramme sont de la forme $\gamma(h, \sigma, k, p) = \sigma^2(1 - e^{-\left(\frac{h}{p}\right)^k})$ où $k$ est un paramètre à définir lors du choix du variogramme.

### 2. Linéarité de l'estimateur

Pour reconstruire un point $x_0 \in D$ à partir des observations, le krigeage utilise l'estimateur linéaire :

$$Y(x_0) = m + \sum_i \lambda_i y_i$$

où $m$ est une constante, $y_i$ désigne la valeur de la i-ème observation et $\lambda_i(x_0)$ est le poids associé, ce dernier est propre au point $x_0$. On veut trouver les poids $\lambda_i$ minimisant la variance de l'erreur :

$$Var\left(Y_{x_0} - \left(a + \sum_i \lambda_i Y_i\right)\right)$$

où $Y_i$ est la variable aléatoire associée à la i-ème l'observation.

L'utilisation d'un estimateur linéaire n'est pas souvent appropriée. Elle est justifiée uniquement dans le cas où l'on souhaite reconstruire un processus aléatoire gaussien. En effet, considérons le problème général suivant.

### Problème général

On cherche un estimateur sans biais noté $\hat{x}(z)$ qui est fonction des observations $z$ et qui minimise l'erreur quadratique :

$$E\left(\left(x(z_0) - \hat{x}(z)\right)^2\right) = Var\left(x(z_0) - \hat{x}(z)\right)$$

où x est un processus aléatoire. La minimisation de cette quantité nous donne l'estimateur de moyenne quadratique $\hat{x}(z) = E(x(z_0) \mid z)$. Cet estimateur n'est une combinaison linéaire des observations que lorsque le processus étudié est gaussien (la démonstration, est donnée en Annexe) [CRE].

Le choix d'un estimateur linéaire implique donc de mauvaises performances lorsque les données ne sont pas normalement réparties. Nous montrons par un exemple que choisir un estimateur linéaire n'est pas toujours correct.

### Exemple

Cherchons à reconstruire la fonction $f(x) = 0.2 \times (\sin(5x) + \sin(\sqrt{3}x) + \tanh(20x))$. L'algorithme estime la portée à 0,1. Commençons d'abord par montrer que l'estimation linéaire réalisée par le krigeage correspond à la moyenne d'un processus gaussien de portée 0,1 conditionné aux observations. Pour cela on simule 100 processus gaussiens de tendance 0, de portée 0,1 et de variance 1,3, en leur imposant de passer par les points d'observation.

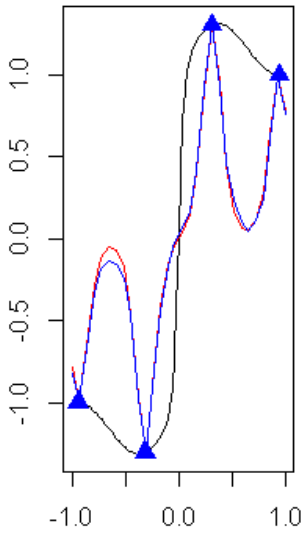


*Figure 8 : réalisations du processus gaussien et interpolation par krigeage pour une portée trop faible*

L'estimation linéaire réalisée par krigeage (courbe en rouge) correspond à la moyenne des processus gaussiens simulés (courbe en bleu).

On choisit maintenant une portée plus forte mieux adaptée ; cela permet de reproduire correctement le comportement général de la courbe.

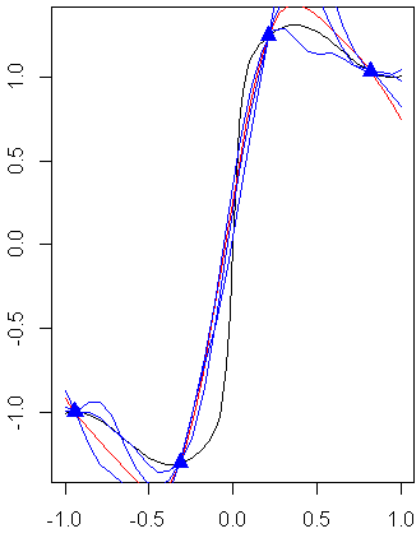


*Figure 9 : réalisations du processus gaussien et interpolation par krigeage pour une portée forte*

Les réalisations du processus sont affichées en bleu. Par endroits, la fonction à reconstruire (en noir) est une réalisation très peu probable du processus. Les données ne sont pas la réalisation d'un processus gaussien.

Afin de régler ce problème, la méthode couramment utilisée est la transformation des observations par la fonction logarithme, afin de rendre les données normalement distribuées. L'estimation linéaire est alors plus adaptée. La difficulté est alors la transformation inverse

qui peut introduire un biais d'estimation. Le krigeage lognormal se base sur ce principe. Il existe une autre variante non-linéaire du krigeage, il s'agit du krigeage disjonctif.

### *C. Les principales méthodes de krigeage*

Les méthodes de krigeage les plus couramment utilisées sont le krigeage simple, le krigeage ordinaire et le krigeage universel. Ces méthodes se différencient par l'hypothèse qu'elles font sur la nature du processus aléatoire.

#### 1. Krigeage simple

Le krigeage simple suppose que le processus est stationnaire d'ordre 2, c'est-à-dire que l'espérance de la variable aléatoire $Y_x$ est constante sur le domaine, et que la fonction de covariance $k(Y_i, Y_j)$ dépend seulement de la distance séparant deux variables aléatoires. Il est obligatoire de connaître l'espérance pour utiliser le krigeage simple, ce qui est rarement le cas en réalité : on ne connaît pas à l'avance la moyenne de la fonction à reconstruire sur le domaine.

On souhaite obtenir le meilleur estimateur linéaire sans biais $\hat{Y}_x$, c'est-à-dire tel que l'erreur est nulle en moyenne, soit $E(Y_x - \hat{Y}_x) = 0$, et minimisant la variance de l'erreur d'estimation $Var(Y_x - \hat{Y}_x)$. L'estimateur s'écrit comme :

$$\hat{Y}_x = \sum_i \lambda_i y_i$$

où $y_i$ est la i-ème observation, et la variance :

$$Var(Y_x - \hat{Y}_x) = Var\left(Y_x - \sum_i \lambda_i Y_i\right) = Var\left(Y_x\right) + Var\left(\sum_i \lambda_i Y_i\right) - 2cov(Y_x, \sum_i \lambda_i Y_i)$$

Dans la suite, Y représente le vecteur colonne des observations, λ est le vecteur des poids associés aux observations. On va chercher à minimiser l'expression de la variance :

$$Var\left(\sum_i \lambda_i Y_i\right) = Var\left(\lambda^t Y\right) = E\left(\lambda^t Y \left(\lambda^t Y\right)^t\right) = \lambda^t E\left(YY^t\right)\lambda = \lambda^t K \lambda$$

$$cov\left(Y_x, \sum_i \lambda_i Y_i\right) = E\left((Y_x - E(Y_x))(\lambda^t Y - E(\lambda^t Y))\right) = E\left(Y_x \lambda^t Y\right) = \lambda^t E\left(Y_x Y\right) = \lambda^t \gamma(x)$$

où K est la matrice de covariance des observations :

$$K(p)=\begin{pmatrix} k(Y_1,Y_1) & k(Y_1,Y_2) & \cdots & k(Y_1,Y_n) \\ k(Y_1,Y_2) & k(Y_2,Y_2) & \dots & k(Y_2,Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(Y_1,Y_n) & k(Y_2,Y_n) & \dots & k(Y_n,Y_n) \end{pmatrix},$$

et $\gamma(x)$ le vecteur de covariance entre $Y_x$ et les observations :

$$\gamma(x)=\begin{pmatrix} k(Y_x,Y_1) \\ k(Y_x,Y_2) \\ \vdots \\ k(Y_x,Y_{n-1}) \\ k(Y_x,Y_n) \end{pmatrix}$$

La variance d'estimation s'écrit :

$$Var\left(Y_x-\lambda^t Y\right)=\sigma^2+\lambda^t K\lambda-2\lambda^t\gamma(x)$$

Le vecteur poids minimisant cette quantité est unique et est donné par $\lambda = \mathrm{K}^{-1}\,\gamma(\mathrm{x})$. L'expression de l'estimateur dans le cas où les variables ne sont pas centrées et sont de moyenne $m$ est donnée par :

$$\hat{Y}_x = m+(\mathrm{K}^{-1}\,\gamma(\mathrm{x}))^t(\mathrm{Y}-\mathrm{m})$$

Cet estimateur est sans biais puisque $E\left(Y_x-\hat{Y}_x\right)=E\left(Y_x\right)-E\left(m+(\mathrm{K}^{-1}\,\gamma(\mathrm{x}))^t\left(\mathrm{Y}-\mathrm{m}\right)\right)=0$.

### 2. Le krigeage ordinaire

Le krigeage ordinaire ne suppose pas l'espérance connue. Chaque variable aléatoire du processus s'écrit $Y_x=\mu+\delta_x$, avec une partie déterministe $\mu$ constante inconnue appelée tendance et une variable aléatoire $\delta_x$ stationnaire intrinsèque et d'espérance nulle. Pour modéliser la dépendance, on utilise un semi-variogramme. Il permet de construire un estimateur en l'absence d'information sur la tendance. L'estimateur obtenu est indépendant de la tendance qui n'a pas besoin d'être calculée ou estimée. La somme des poids est contrainte à être égale à 1 afin que l'estimateur soit sans biais et afin que la variance et l'espérance de l'erreur d'estimation soient définies.

### 3. Le krigeage universel

Le krigeage universel reprend les mêmes hypothèses que le krigeage ordinaire. Cependant la tendance n'est plus supposée constante. L'utilisateur prédéfinit une famille finie de fonctions

de base notées $f_j$. La tendance $\mu(x)$ s'exprime comme une combinaison linéaire de ces fonctions $f_j$, c'est-à-dire $\mu(x) = \sum_j \beta_j f_j(x)$ avec $\beta_j$ des coefficients inconnus. Les contraintes d'autorisation et de non biais doivent également être respectées. La minimisation de l'erreur donne un estimateur indépendant des coefficients $\beta_j$ qui n'ont pas besoin d'être estimés.

Il existe d'autres versions de krigeage non-stationnaire, comme le krigeage bayesien ou le krigeage intrinsèque généralisé. Dans le krigeage bayesien, la tendance dépend de paramètres, traités comme des variables aléatoires et estimés par inférence bayesienne.

# III. Présentation de l'EPH

Contrairement au krigeage, la méthode de l'EPH n'utilise que les données elles-mêmes. Aucune supposition artificielle n'est faite. La construction de l'EPH peut être vue comme la propagation de l'information provenant des points d'observation vers des points inconnus. La propagation est basée sur le principe d'entropie, qui est une fonction croissante avec la distance au point de mesure. L'EPH nécessite plusieurs paramètres d'entrée :

- les bornes de chaque dimension du domaine ;
- les bornes des valeurs de sortie et le pas de discrétisation.

Ces bornes peuvent être obtenues à dire d'expert ou bien définies par un utilisateur.

Le résultat de l'EPH est donné sous la forme d'un ensemble de lois de probabilités dont la variance est maximale à entropie fixée. Les densités sont des fonctions de Dirac aux points de mesure, et deviennent de moins en moins concentrées lorsque l'on s'en éloigne. Chaque mesure donne sa propre contribution au résultat final sous la forme :

$$p_{n,j}(X) = \frac{\tau}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(j-C_n)^2}{2\sigma^2}\right\},$$

où

$$\sigma = \frac{\tau e^{\lambda d_n}}{\sqrt{2\pi e}},$$

où $X$ est le point à estimer, n est le numéro de la mesure $n = 1,...,N$, $C_n$ la valeur de la $n^{ème}$ mesure, j la discrétisation de la plage de valeurs de sortie avec le pas $\tau$, $d_n$ la distance entre le point inconnu et le $n^{ème}$ point, et $\lambda$ le paramètre calculé de sorte à maintenir l'entropie maximale (ou l'information minimale) en chaque point.

A la fin du processus, les lois individuelles sont recombinées en fonction de la distance du point à reconstruire avec chaque mesure. Le résultat final pour chaque point renvoyé par l'EPH est la loi probabilité donnée par la formule :

$$p_j(x) = \frac{1}{\sum_{i=1}^{N} 1/d_i}\left(\frac{1}{d_1}p_{1,j} + \ldots + \frac{1}{d_N}p_{N,j}\right)$$

Il est possible d'ajuster la formule lorsque l'on se trouve dans un espace de haute-dimension. La description détaillée de la construction de l'EPH est données dans le livre [PIT].

Dans le cadre du contrat précédent entre l'IRSN et la SCM, l'EPH a été implémentée dans sa forme générale en R. La fonction predict.eph prend en entrées les observations, les bornes des différents paramètres ainsi que les points à reconstruire et elle retourne à partir de la loi de probabilité obtenue par l'EPH en n'importe quel point du domaine, l'espérance, la valeur la plus probable, la médiane, ainsi que les 5e et 95e centiles.

Dans le cadre du présent contrat, nous avons modifié la fonction afin qu'elle renvoie également la loi de probabilité elle-même, permettant à l'utilisateur de calculer d'autres caractéristiques comme par exemple les quantiles 10 et 90.

# IV. Publication sur l'archive internet du langage R des précédents développements

Afin de valoriser les réalisations logicielles du contrat précédent, nous les publions sur l'archive centralisée du logiciel R : le CRAN. Nous publions l'implémentation de l'EPH dans sa version la plus générale : la dimension de l'espace est choisie par l'utilisateur, le nombre de points de mesure est également arbitraire.

Nous avons constitué un "package" regroupant les sources, la documentation du code, et les exemples d'utilisation. Le dossier du package est structuré comme ci-dessous :

- Un fichier "DESCRIPTION" indiquant la version du package, l'auteur du code, la licence, son titre et une description. Il est écrit au format Debian Control File (DCF).

- Un fichier "README.md" qui décrit ce pourquoi le package doit être utilisé, comment l'utiliser et ce qu'il contient.

- Un fichier "NAMESPACE" spécifiant quelles variables doivent être renvoyées à l'utilisateur du package et quelles variables ou fonctions doivent être importées depuis d'autres packages.

- Un dossier "R " contenant les codes source.

– Un dossier "man" contenant la description des fonctions et des classes. Chaque fonction est documentée dans un fichier séparé qui décrit les entrées et sorties de la fonction, son fonctionnement et les exemples d'utilisation. Le fichier est écrit au format R documentation (Rd), et respecte sa syntaxe.

Le contenu de ces fichiers est donné en Annexe.

Le logiciel R fournit un programme permettant de vérifier :

- que le package est correctement structuré,
- que le formalisme pour la documentation est respecté,
- et que les exemples fonctionnent.

Pour être accepté pour publication, l'exécution de ce programme ne doit produire aucune erreur ou avertissement. Nous l'avons lancé sur notre package EPH avec succès : aucune erreur ni avertissement n'est renvoyé. Le résultat de l'exécution et le détail des tests réalisés par cette commande sont donnés en Annexe.

# V. La prise en compte des incertitudes sur les données

Le travail fait précédemment supposait que les données étaient précises (coordonnées des points de travail, valeurs des mesures). Nous comparons les aptitudes du krigeage, d'une part, de l'EPH, d'autre part, à incorporer les incertitudes sur les différents paramètres : incertitudes sur les positions des points de mesure, incertitudes sur la valeur de la mesure.

## *A. La prise en compte des incertitudes dans l'EPH*

Elle est très facile à réaliser, parce que par définition l'EPH est de nature probabiliste.

Voyons-la de manière détaillée sur un exemple. Supposons qu'il y ait une incertitude sur la première coordonnée du premier point de mesure, c'est à dire sur le paramètre $\xi_1^{(1)}$. Supposons de plus, par simplicité, que ce paramètre puisse prendre seulement deux valeurs : $a_1$ avec probabilité $q_1$ et $a_2$ avec probabilité $q_2$, avec $q_1 + q_2 = 1.$

Nous supposons d'abord que $\xi_1^{(1)} = a_1$ et faisons la construction complète de l'EPH en ce cas. En un point $X$, nous obtenons la loi $p_1(j, X)$. Ensuite, nous faisons la même chose en supposant $\xi_1^{(1)} = a_2$ ; au même point $X$ nous obtenons la loi $p_2(j, X)$.

Alors, si nous prenons l'incertitude en compte, la valeur donnée par l'EPH au point $X$ sera :

$$p(j,X) = q_1 p_1(j,X) + q_2 p_2(j,X).$$

Cette procédure très simple s'étend au cas des incertitudes sur les coordonnées des points de mesure, aux incertitudes sur la valeur de la mesure, et sur la position des points tests $X$.

On fait la liste des mesures, avec leurs probabilités, et on réalise la construction de l'EPH séparément dans chaque cas. On construit à la fin une "EPH moyennée" en combinant les différentes constructions avec leurs probabilités respectives.

Les lois de probabilité relatives aux incertitudes peuvent être discrètes ou continues, le principe est le même.

Le résultat donné par l'EPH dans le cas d'une valeur précise est une combinaison de gaussiennes ; si nous prenons les incertitudes en compte, nous avons encore une combinaison de gaussiennes (mais plus complexe, bien sûr).

## Un exemple simple

Supposons que le second point, noté $B$, prenne seulement 4 positions, comme sur la figure ci-dessous :

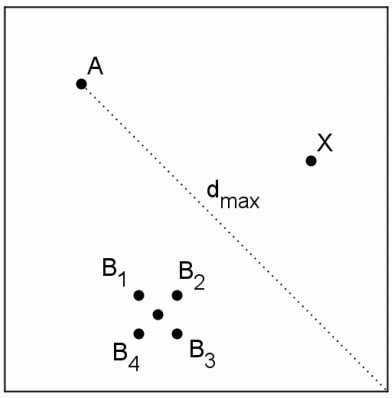


*Figure 1: positions possibles du point* $B$

Le carré est de côté 1 ; le point $A$ a pour coordonnées (1/5, 4/5) et le point $B$ (2/5, 1/5). Les points $B_i$ ont pour coordonnées $(\frac{2 \pm 0.25}{5}, \frac{1 \pm 0.25}{5})$. Nous nous intéressons au point $X$ de coordonnées (4/5, 3/5).

Tout d'abord, nous construisons l'EPH au point $X$, en considérant séparément les quatre situations $A, B_1$, $A, B_2$ , $A, B_3$ , $A, B_4$. Ensuite, nous calculons la moyenne des quatre lois obtenues.

Voici le graphe obtenu :

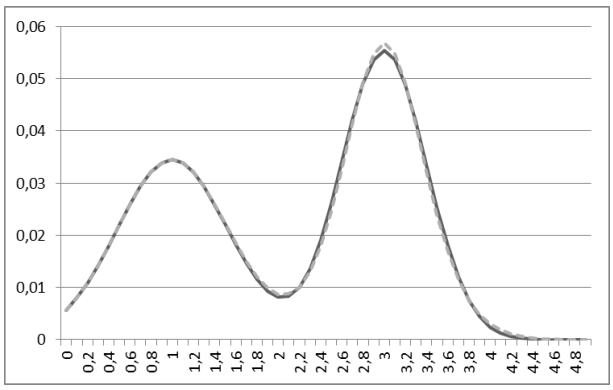


*Figure 2 : prise en compte des incertitudes sur* $B$

La courbe en pointillés correspond à l'EPH faite sur le point $B$ tout seul. La courbe en traits pleins correspond à la moyenne des courbes obtenues à partir des quatre points $B_i$. La seconde est légèrement moins concentrée : cette conclusion est générale ; lorsqu'on prend les incertitudes en compte, les lois deviennent moins concentrées.

### Précautions à prendre

La construction doit être reprise dès le début, en tenant compte des diverses possibilités séparément. Par exemple, dans le cas ci-dessus, si le point incertain était $A$ avec quatre positions, il nous faudrait calculer $d_{\max}$ séparément pour chacune de ces quatre positions, et pour chacune le coefficient de propagation $\lambda$. Il ne faut pas se contenter d'un $d_{\max}$ moyen et d'un $\lambda$ moyen. L'ensemble de la construction est non-linéaire.

### Prise en compte de plusieurs incertitudes

En pratique, toutes les données peuvent être affectées d'une incertitude :

- Coordonnées des points de mesure ;
- Coordonnées du point cible ;
- Valeurs observées pour la mesure ;

Si la dimension est $K$ et le nombre de points de mesure est $N$, nous avons $(N+1)K+N$ valeurs scalaires ; si chacune a deux valeurs possibles, nous avons $2^{(N+1)K+N}$ cas différents, ce qui est trop élevé pour un traitement exhaustif.

Nous utilisons donc une méthode de type Monte-Carlo, tout à fait appropriée dans ce cas, puisqu'on recherche une moyenne. Nous procédons comme suit :

On fixe un nombre total de runs $M$, par exemple $10^6$ ;

– Pour chaque $m = 1,...,M$, on tire une valeur aléatoire pour chaque paramètre affecté d'une incertitude ;

– On calcule l'EPH dans ce cas ;

– On fait la moyenne des $M$ lois obtenues séparément dans chaque cas.

Illustrons cette méthode sur un exemple simple.

**Exemple de prise en compte de plusieurs incertitudes**

On cherche à reconstruire la fonction Branin-Hoo à partir de 16 mesures. La valeur de la mesure et sa position suivent une loi uniforme de largeur d'intervalle 20 et 0,1 respectivement. On fixe le nombre total de runs à 200. La figure ci-dessous représente la surface attendue (les valeurs réelles) et la surface interpolée sans et avec gestion des incertitudes :

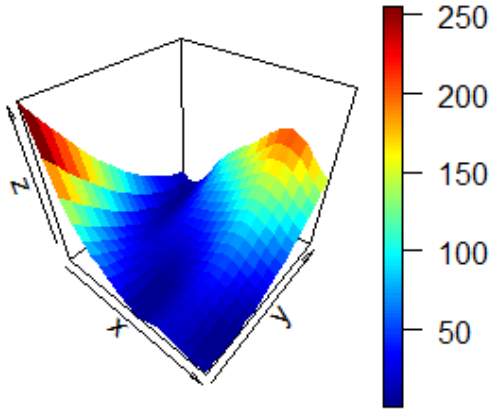


*Figure 10 : fonction de Branin*

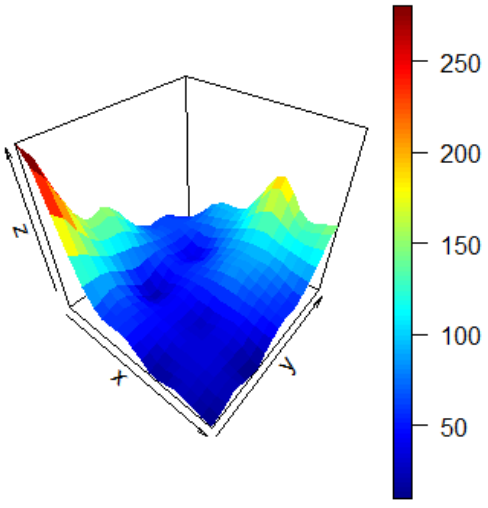


*Figure 11 : EPH sans incertitudes*

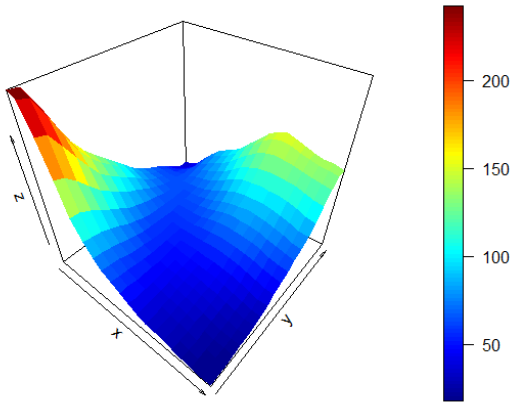


*Figure 12 : EPH avec incertitudes*

L'incertitude sur les paramètres d'entrée a été propagée. La surface avec prise en compte des incertitudes est plus lisse.

La Figure 13 montre la loi de probabilité renvoyée par l'EPH avec prise en compte des incertitudes pour le point de coordonnés (0, 4/5) :

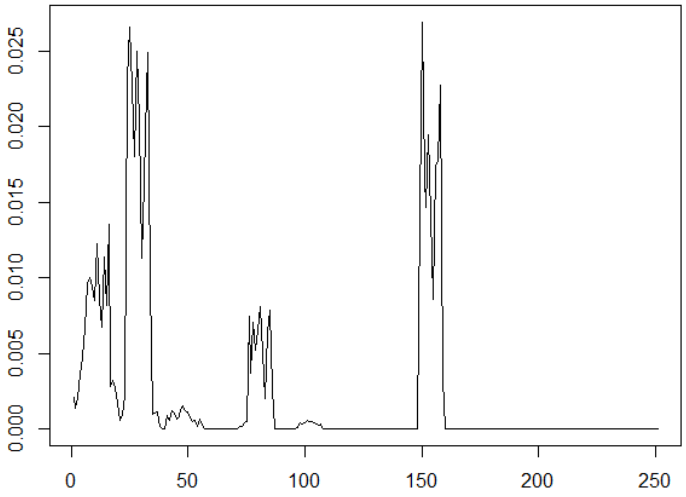


*Figure 13 : loi de probabilité dans le cas avec incertitudes*

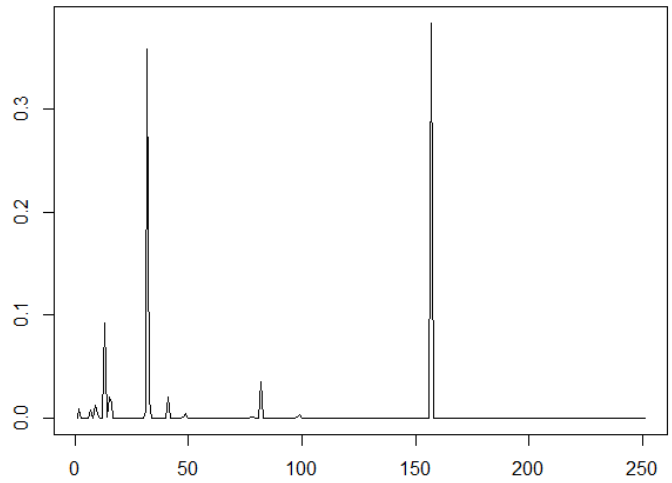


*Figure 14 : loi de probabilité dans le cas sans incertitudes*

Les lois de probabilité obtenues pour chaque point sont moins concentrées dans le cas avec incertitudes.

## *B. Prise en compte des incertitudes dans la méthode de krigeage*

Nous montrons dans un premier temps comment prendre en compte les incertitudes sur les données, puis exposons les risques d'instabilité survenant lorsqu'on traite ces incertitudes. En effet le calcul des poids $\lambda$ nécessite la résolution d'un système de la forme $K\lambda = K_0$. Lorsque les données présentent des incertitudes, cela se traduit par une perturbation sur le terme de droite et sur la matrice de ce système. Si la matrice est mal conditionnée, une telle perturbation, même minime, peut engendrer une variation considérable des poids et donc de l'interpolation réalisée.

Les incertitudes portent sur la valeur des observations, la localisation des observations et des points à estimer, ainsi que sur les paramètres du modèle utilisé (variogramme, fonction de covariance).

### 1. Sur la position des observations

#### Cas général

Soit une observation située au point $x_\alpha \in R^N$. On note $\varepsilon_\alpha$ le vecteur de composantes ${\varepsilon^i}_\alpha$ représentant l'erreur de localisation sur l'i-ème coordonnée de l'observation $x_\alpha$. On note $p(\varepsilon_\alpha)$ la loi de ce vecteur aléatoire et $p(\varepsilon_\alpha, \varepsilon_\beta)$ la loi conjointe des deux vecteurs aléatoires associés aux observations en $x_\alpha$ et $x_\beta$. Soit $k_{\alpha\beta}$ l'élément de la matrice $K$ situé en $(\alpha, \beta)$. $k_{\alpha\beta}$ est la

covariance (ou le variogramme) entre les variables aléatoires associées aux observations respectives. La fonction moyenne $\tilde{k}_{\alpha\beta}$ prenant compte des incertitudes s'écrit :

$$\tilde{k}_{\alpha\beta} = E_{\varepsilon_\alpha,\varepsilon_\beta}(f(x_\alpha+\varepsilon_\alpha, x_\beta+\varepsilon_\beta)) = \iint f(x_\alpha+\varepsilon_\alpha, x_\beta+\varepsilon_\beta)\, p(\varepsilon_\alpha,\varepsilon_\beta) d\varepsilon_\alpha d\varepsilon_\beta$$

où $f$ est la fonction de covariance ou le variogramme selon le type de krigeage choisi. De la même manière pour le variogramme, on étudie l'impact moyen des erreurs sur $\gamma_{x_0}$ (covariance entre le point x à estimer et l'observation $x_0$) et le variogramme moyen s'écrit :

$$\tilde{\gamma}_{x_0} = \iint f(x+\varepsilon_x, x_0+\varepsilon_0)\, p(\varepsilon_x,\varepsilon_0) d\varepsilon_0$$

Les poids sont maintenant calculés par $\lambda = \tilde{K}^{-1}\tilde{k}$

Cette méthode est introduite dans [CHI] et dans [DEV]. Appliquons-la maintenant sur un exemple simple.

### Un exemple simple

On dispose de deux mesures A, B et C. Supposons que le second point, noté $B$, prenne seulement 4 positions, comme sur la figure ci-dessous :

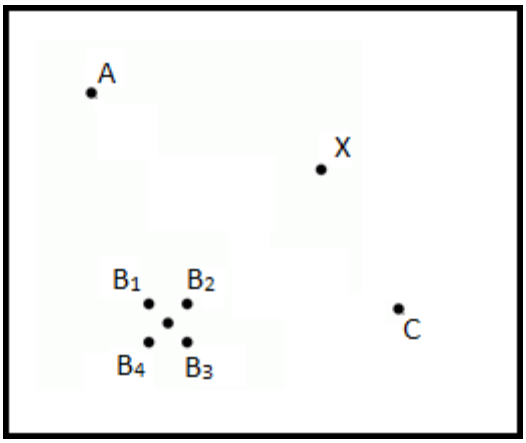


*Figure 15 : positions possibles du point* $B$

Nous nous intéressons au point $X$. On calcule d'abord la fonction $f$ qui est le modèle de covariance ou le variogramme. La fonction $f$ reste la même quelle que soit la position de $B$.

On cherche à calculer une matrice de covariance moyenne :

$$\tilde{K} = \begin{pmatrix} f(A,A) & f(A,B) & f(A,C) \\ f(B,A) & f(B,B) & f(B,C) \\ f(C,A) & f(C,B) & f(C,C) \end{pmatrix}$$

Pour calculer $f(A,B)$ par exemple, on effectue la moyenne de la covariance en supposant que B vaut tout à tour $B_1, B_2, B_3$ et $B_4$ :

$$f(A,B) = \frac{f(A,B_1) + f(A,B_2) + f(A,B_3) + f(A,B_4)}{4}$$

et de même pour $f(X,B)$.

On peut alors calculer les poids pour tout $X$ par $\lambda = \tilde{K}^{-1}k$, où k est le vecteur $k = (f(X,A), f(X,B), f(X,C))$, c'est-à-dire la covariance de $X$ avec chaque point d'observation.

### 2. Sur la dépendance spatiale

L'incertitude sur le degré de variabilité des données se traduit par une incertitude sur le paramètre de portée du modèle. Pour prendre en compte les incertitudes liées au coefficient de portée, on effectue le même raisonnement que plus haut, puisque les erreurs vont affecter la matrice $K$ et le vecteur $\gamma$. Le coefficient de portée est unique par hypothèse de stationnarité et il est estimé par EMV.

On note $\rho$ le coefficient de portée, $\delta$ la variable aléatoire représentant l'erreur commise sur l'estimation de la portée et $p(\delta)$ la loi de cette erreur. On a donc :

$$\tilde{k}_{\alpha\beta} = \int f(x_\alpha, x_\beta, \rho + \delta)\, p(\delta) d\delta ,$$

et :

$$\tilde{\gamma}_{xo} = \int f(x, x_o, \rho + \delta)\, p(\delta) d\delta$$

Les erreurs sur le paramètre $\rho$ affectent directement la matrice $K$ et le vecteur $\gamma$, d'où la nécessité d'étudier le conditionnement de la matrice $K$.

Deux variantes du krigeage permettent de prendre en compte l'incertitude sur les paramètres du modèle :

– le krigeage bayesien ;

– krigeage avec variogramme flou : il calcule une variance floue prenant en compte les incertitudes [FLO].

L'intérêt de ces méthodes est que la variance renvoyée prend en compte les incertitudes.

### 3. Sur la valeur des observations

Les erreurs sur les observations affectent indirectement la matrice K et le vecteur $\gamma$, puisque le modèle de dépendance est construit à partir des données. Pour simplifier l'étude, on suppose que les erreurs commises sur les observations n'ont pas d'influence sur la portée. Les poids de l'estimateur sont donc indépendants de la valeur des observations.

Les observations sont des réalisations des variables aléatoires $Y_i$ mais qui peuvent être entachées d'erreurs de mesure. Une manière de prendre en compte l'incertitude sur les observations est d'attribuer aux variables aléatoires $Y_i$ une variance plus élevée que celle associée aux autres variables du processus. Le krigeage donne alors une estimation qui tient compte de cette incertitude.

En pratique, dans le cas du krigeage ordinaire et universel, on rehausse le variogramme de la valeur de la variance associée aux erreurs de mesure qu'on suppose être la même pour chaque observation, notée $\varepsilon$ (Figure 16). Le variogramme $\gamma$ vérifie :

$$\lim_{h \to 0} \gamma(h) = \varepsilon ,$$

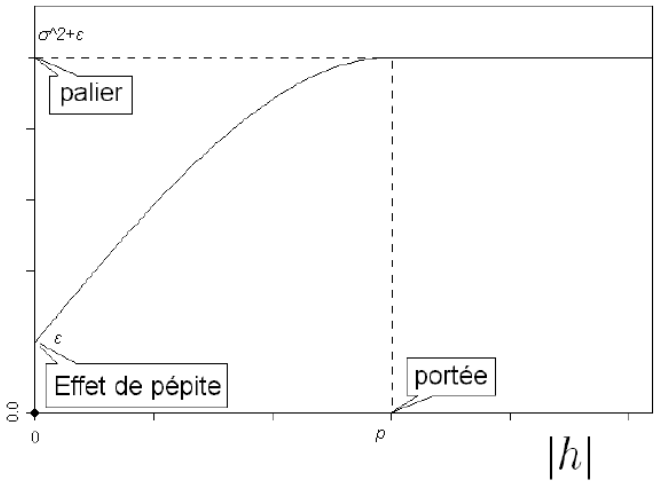


*Figure 16 : variogramme avec effet pépite*

Dans le cas du krigeage simple, on effectue la même manipulation pour la fonction de covariance. Il suffit d'ajouter le terme $\varepsilon$ sur la diagonale de la matrice de covariance K.

L'ajout de cette variance permet de prendre en compte les incertitudes. En effet, considérons que les observations sont les réalisations d'un processus dont les variables aléatoires s'écrivent comme $Y_i + E_i$ où $E_i$ est la variable aléatoire modélisant l'erreur de mesure associée à l'observation $i$ telle que $Var(E_i) = \varepsilon$, $cov(E_i, Y_x) = 0$ et $cov(E_i, E_j) = 0$. On cherche à minimiser :

$$Var(Y_x - \sum_i \lambda_i (Y_i + E_i))$$

Cette quantité est égale à :

$$Var(Y_x) + Var\left(\sum_i \lambda_i \ (Y_i + E_i)\right) - 2cov(Y_x, \sum_i \lambda_i (Y_i + E_i)).$$

Avec :

$$cov(Y_x, \sum_i \lambda_i (Y_i + E_i)) = \sum_i \lambda_i cov(Y_x, Y_i)$$

puisque $\text{cov}(Y_x, E_i) = 0$.

Et :

$$Var\left(\sum_i \lambda_i \ (Y_i + E_i)\right) = \lambda^t (K + \varepsilon Id)\lambda$$

car $\text{cov}(Y_i + E_i, Y_j + E_j) = \text{cov}(Y_i, Y_j)$ pour $i \neq j$, et $\text{cov}(Y_i + E_i, Y_i + E_i) = Var(Y_i + E_i) = \sigma^2 + \varepsilon$

La quantité à minimiser s'écrit donc :

$$Var(Y_x) + \lambda^t (K + \varepsilon Id)\lambda - 2cov(Y_x, \sum_i \lambda_i Y_i)$$

Le poids minimisant cette expression est :

$$\lambda = (K + \varepsilon Id)^{-1} k$$

Ecrire les variables aléatoires comme $Y_i + E_i$ revient donc bien à ajouter la variance sur la diagonale de la matrice K. Une démonstration similaire peut être faite pour le variogramme.

**Exemple**

Dans la simulation ci-dessous, on reconstruit la fonction :

$$f(x) = 0.2 \times (\sin(5x) + \sin(\sqrt{3}x) + \tanh(20x))$$

en noir à partir de 5 points de mesures représentés par les triangles bleus :

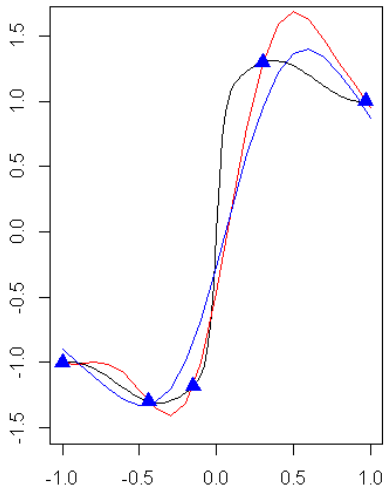


*Figure 17 : gestion des incertitudes par effet nugget*

Les courbes bleue et rouge représentent l'interpolation faite par krigeage avec et sans effet pépite respectivement. La variance associée aux erreurs de mesure est fixée à 0.2.

L'estimation obtenue avec effet pépite est plus régulière que l'estimation obtenue sans effet pépite. Elle atténue les extrema et ne passe pas forcément par les points d'observation.

# VI. Capacité à renvoyer une incertitude sur le résultat

## *A. Pour le krigeage*

Le résultat renvoyé par le krigeage est déterministe. Il fournit une variance d'estimation qui dépend de la position des mesures, mais pas de leurs valeurs. Elle reflète surtout la densité des points de mesure environnants et non la variabilité des données. De plus, pour obtenir un intervalle de confiance à partir de cette variance, il faut faire des hypothèses, comme supposer une distribution gaussienne, ce qui n'a pas lieu d'être vérifié en réalité.

Le krigeage ordinaire est conçu pour effectuer une prédiction ponctuelle sans reproduire la variabilité locale du phénomène sur le domaine, ce qui rend impossible le calcul d'une probabilité de dépassement de seuil. Le krigeage disjonctif ou par indicatrices rend cependant possible le calcul d'une telle probabilité.

Le krigeage par indicatrice consiste à reconstruire la loi de probabilité conditionnée aux observations pour chaque point à reconstruire. On veut effectuer l'estimation en un point $x_0$.

On discrétise la variable de sortie en $N$ valeurs possibles $c_i$, avec $1 \le i \le N$. L'indicatrice associée à $c_i$ pour chaque observation $x_j$ est :

– $I(x_j, c_i) = 1$ lorsque $Y_j \le c_i$ ;

– $I(x_j, c_i) = 0$ lorsque $Y_j > c_i$

Pour chaque valeur $c_i$, on effectue la combinaison linéaire des indicatrices associées à chaque observation en utilisant les poids du krigeage ordinaire $I(x_0, c_i) = \sum_{j=1}^{M} \lambda_j I(x_j, c_i)$. Cela donne une estimation de la fonction de répartition pour le point $x_0$, $F(x_0, c_i) = P(Y_0 < c_i) = I(x_0, c_i)$. La démonstration est présentée en détail dans [KI].

Le krigeage donne toujours une interpolation moins variable que le phénomène réel, c'est-à-dire plus lisse. Afin de prendre en compte les variabilités locales, on peut utiliser les simulations séquentielles gaussiennes, qui permettent de générer des représentations plus proches de la réalité.

### *B. Pour l'EPH*

Il est possible avec l'EPH d'obtenir une incertitude sur le résultat fourni. En effet l'EPH renvoie une loi de probabilité en tout point, à partir de laquelle on peut calculer un intervalle de confiance ainsi qu'une probabilité de dépassement de seuil. On peut également calculer une probabilité de dépassement de seuil globale à partir des lois de probabilité de chaque point.

L'intervalle de confiance calculé avec l'EPH est très large, bien plus que pour le krigeage. Pour l'EPH, les bornes de cet intervalle sont les quantiles de la loi de probabilité.

Dans l'exemple ci-dessous, nous calculons les quantiles 20 et 80 des lois de probabilités en chaque point. Les intervalles de confiance et la courbe à reconstruire (en noir) sont affichés ci-dessous :

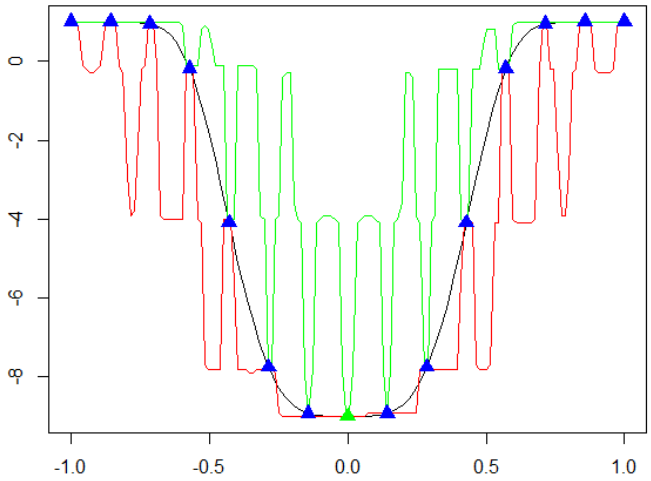


*Figure 18 : intervalle de confiance à 80% pour l'EPH*

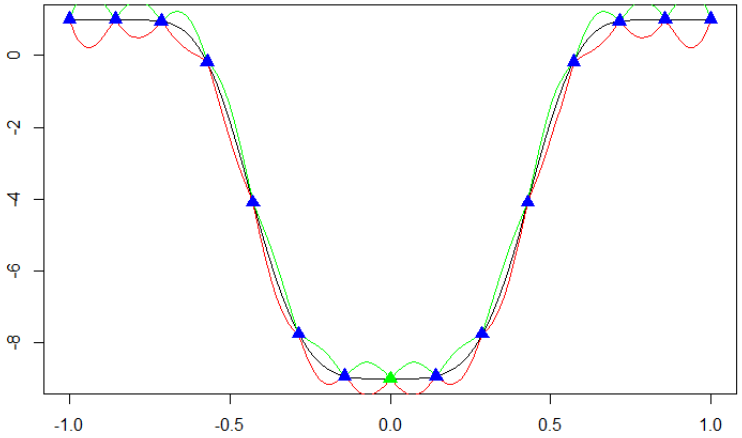


*Figure 19 : intervalle de confiance à 80% pour le krigeage*

L'intervalle de confiance de l'EPH est bien plus large que celui du krigeage.

Cherchons à reconstruire 400 points de la fonction de Branin à partir de 81 points de mesure choisis régulièrement dans [0,1]x[0,1] en utilisant la méthode de l'EPH.

La figure ci-dessous montre la surface à reconstruire, et la suivante montre la largeur de l'intervalle de confiance à 80% renvoyé par l'EPH en chaque point. Plus la couleur est rouge plus l'intervalle est fin :

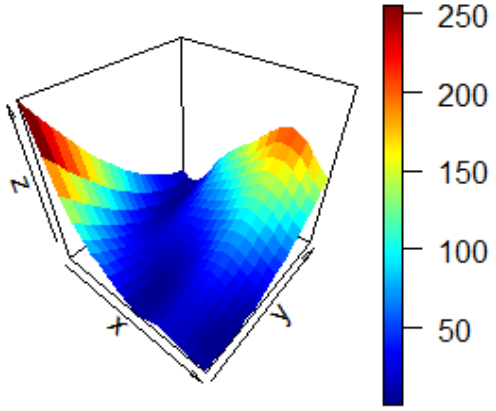


*Figure 20 : fonction de Branin*

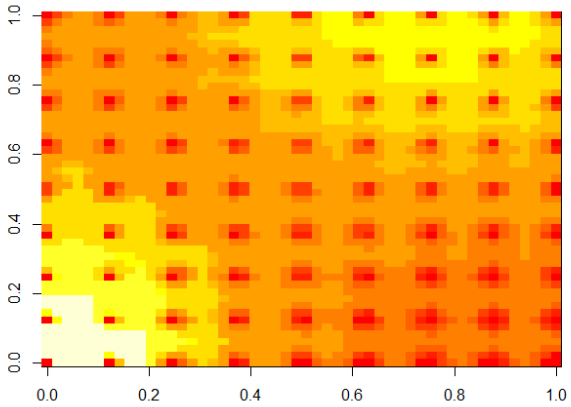


*Figure 21 : largeur de l'intervalle de confiance 80%*

Les intervalles de confiance sont plus larges entre les mesures. Evidemment, aux points d'observation ils sont de largeur nulle. Etonnamment, dans cet exemple les intervalles de confiance sont plus larges aux points où la fonction à reconstruire est la plus grande, et plus resserrés aux points où elle prend des valeurs plus faibles. L'explication est que le quantile 20 est quasiment constant sur le domaine.

Il est naturel que l'intervalle de confiance soit plus large pour l'EPH, puisque celle-ci ne fait pas d'hypothèse de modèle, à la différence du krigeage.

# VII. La convergence de l'algorithme lorsque le nombre de points de mesure augmente

On peut s'attendre, lorsque le nombre de points de mesure augmente (et, à la limite, lorsque tout point est mesuré) à ce qu'un algorithme d'extrapolation converge vers la valeur vraie en tout point ; c'est une question tout à fait naturelle.

*A. Pour le krigeage*

Lorsque le processus que l'on cherche à reconstruire est continu, le krigeage converge vers la valeur vraie en tout point :

**Théorème.** – *Soit* $Y_{N,x} = \sum_{i=1}^{N} \lambda_i(x) Y_i$ *l'estimation de* $Y_x$ *à partir de N observations. Soit* $\lambda$ *le vecteur colonne des poids, alors :*

$$\min_{\lambda} \left( Var\left( Y_x - Y_{N,x} \right) \right) \underset{N \to \infty}{\to} 0$$

*La variance de l'erreur d'estimation tend vers 0 quand le nombre d'observations augmente.*

**Démonstration du théorème (inspirée de [SMO])**

La variance de l'erreur d'estimation s'écrit :

$$Var\left( Y_x - Y_{N,x} \right) = E\left( \left( Y_x - Y_{N,x} - E(Y_x - Y_{N,x}) \right)^2 \right) = E\left( \left( Y_x - Y_{N,x} \right)^2 \right) ,$$

puisque l'estimateur $Y_{N,x}$ est sans biais. On majore l'erreur d'estimation :

$$\min_{\lambda} \left( E\left( \left( Y_x - Y_{N,x} \right)^2 \right) \right) \le E\left( \left( Y_x - Y_i \right)^2 \right), \forall i \in \{1, ..., N\}$$

Ce qui s'écrit, dans le cas du krigeage simple :

$$\min_{\lambda} \left( E\left( \left( Y_x - Y_{N,x} \right)^2 \right) \right) \le E(Y_x)^2 + E(Y_i)^2 - 2cov\left( Y_x, Y_i \right) = 2\sigma^2 - 2k\left( Y_x, Y_i \right) \quad (1)$$

et dans le cas du krigeage ordinaire et universel :

$$\min_{\lambda} \left( E\left( \left( Y_x - Y_{N,x} \right)^2 \right) \right) \le Var\left( Y_x - Y_i \right) = \gamma(Y_x, Y_i) \quad (2)$$

Le processus est continu, donc la fonction $k$ vérifie $\lim_{h \to 0} k\left( Y_{x+h}, Y_x \right) = \sigma^2$. Le semi-variogramme vérifie $\lim_{h \to 0} \gamma\left( Y_{x+h}, Y_x \right) = 0$

Le krigeage est un interpolateur exact. Donc si le point d'estimation correspond à un point de mesure, la valeur de l'estimation est égale à la valeur au point d'observation. En effet, on annule la quantité $E\left( \left( Y_x - Y_{N,x} \right)^2 \right)$ en choisissant le vecteur poids donnant l'observation concernée.

Supposons que l'on augmente le nombre d'observation. Le point à estimer sera alors entouré par un nombre d'observations de plus en plus grand. On suppose que les observations forment un ensemble A qui est inclus dans le domaine d'étude D et tel que A est dense dans D. Cela veut dire que les observations peuvent être aussi proches du point à estimer que l'on veut sans y être égales. Donc $\forall x \in D$ il existe une suite $\left(x_n\right)_n \in A$ tel que $x_n \underset{n\to\infty}{\to} x$.

Soit N le nombre d'observations. On note $I = \{1,\ldots,N\}$ et $\left(x_i\right)_{i\in I} \in A^{\mathbb{N}}$ la suite de ces N éléments. On fait tendre N vers l'infini. Cette suite fait partie d'un ensemble dense, il existe donc une sous suite $\left(x_{i_k}\right)_{i_k\in I}$ telle que $x_{i_k} \underset{k\to\infty}{\to} x$. Dans les inégalités définies ci-dessus (1) et (2) on choisit l'observation $Y_{i_k}$ pour la majoration. Quand l'indice $k$ tend vers l'infini $d\left(Y_x, Y_{x_{i_k}}\right) \to 0$ et $k\left(Y_x, Y_{x_{i_k}}\right) \to \sigma^2$ donc dans le cas du krigeage simple en passant l'inégalité (1) à la limite :

$$\min_{\lambda}\left(Var\left(Y_x - Y_{N,x}\right)\right) \underset{N\to\infty}{\to} 0$$

De même dans le cas du krigeage ordinaire et universel, quand l'indice k tend vers l'infini $d\left(Y_x, Y_{i_k}\right) \to 0$ donc $\gamma(Y_x, Y_{i_k}) \to 0$ donc $\min_{\lambda}\left(Var\left(Y_x - Y_{N,x}\right)\right) \underset{N\to\infty}{\to} 0$

L'estimateur est sans biais et la variance de l'erreur d'estimation tend vers 0 quand le nombre d'observations augmente. L'estimation est donc de plus en plus proche de la réalisation du processus en tout point du domaine. L'algorithme converge vers la valeur vraie en tout point.

La démonstration n'est pas valable lorsque la fonction à reconstruire n'est pas continue. L'exemple suivant de reconstruction de la fonction d'Heaviside illustre cette situation :

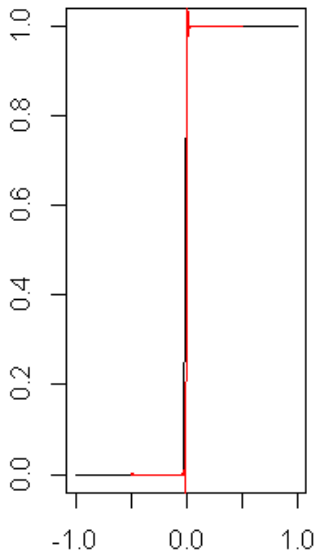


*Figure 22 : reconstruction de 131 points à partir de 150 mesures*

L'estimation échoue au voisinage de zéro. Les phénomènes naturels sont caractérisés par une grande variabilité et les variables physiques les décrivant (teneurs, cote topographiques, pressions…) sont souvent très discontinues. Il est donc important que l'algorithme fonctionne raisonnablement bien lorsqu'il y a une discontinuité ; ce qui n'est pas le cas du krigeage.

### *B. Pour l'EPH*

Lorsque la fonction que l'on cherche à reconstruire est continue, le résultat concernant l'EPH est très satisfaisant, et il est simple à établir :

**Théorème.** – *Soit* $f$ *une fonction continue sur* $\mathbb{R}^K$ *(la dimension* $K$ *étant quelconque). Soit* $X$ *un point de* $\mathbb{R}^K$ *et soit* $A_n$ *une suite quelconque de points de* $\mathbb{R}^K$ *convergeant vers* $X$. *Soient* $\vartheta_n$ *les valeurs mesurées aux points* $A_n$, *c'est-à-dire* $\vartheta_n = f(A_n)$ *; soit encore* $e_n$ *la valeur extrapolée prédite par l'EPH au point* $X$ *à partir des points* $A_1,...,A_n$ *(c'est-à-dire, par définition, l'espérance de la loi de probabilité construite en* $X$ *à partir des informations envoyées par* $A_1,...,A_n$*). Alors la suite* $(e_n)$ *converge vers* $e = f(X)$ *lorsque* $n \to +\infty$.

**Démonstration du théorème**

Rappelons que le point $A_n$ envoie en $X$ une information sous la forme d'une loi de probabilité gaussienne discrétisée :

$$p(j, A_n, X) = c \exp\left\{ -\frac{(t_j - \vartheta_n)^2}{2\sigma_n^2} \right\}$$

où $c$ est une constante de normalisation et où la variance $\sigma_n$ vaut :

$$\sigma_n = \frac{\tau}{\sqrt{2\pi}} \exp\{\lambda d(A_n, X)\}$$

Lorsque $d(A_n, X) \to 0$, $\sigma_n \to \frac{\tau}{\sqrt{2\pi}}$, ce qui nous donne une gaussienne, d'espérance $\vartheta_n$, où $\vartheta_n \to f(X)$.

Remarquons que la loi ainsi obtenue à la limite n'est pas une masse de Dirac ; la gaussienne tend vers une Dirac si le pas de la discrétisation en $t$, c'est à dire $\tau$, tend vers zéro.

Maintenant, si on combine les sources $A_1,...,A_n$, le résultat est une loi de probabilité, mélange de gaussiennes, définie par :

$$P_n(X) = \gamma_1 p_1(X) + \ldots + \gamma_n p_n(X)$$

où $p_j$ est la contribution du point $A_j$, définie ci-dessus, et les coefficients $\gamma_j$ sont définis par :

$$\gamma_j = \frac{d_j^{-K}}{\sum_{i=1}^{n} d_i^{-K}}$$

(voir [PIT], Part III, Lemma 4, p. 166)

Comme $d_n \to 0$, $\frac{1}{d_n^K} \to +\infty$ et, pour chaque $j$, $\gamma_j \to 0$ lorsque $n \to +\infty$. Autrement dit, les premières contributions des points $A_j$ deviennent négligeables lorsque $n$ augmente.

Fixons $\varepsilon > 0$. Il existe, puisque la fonction est continue et puisque $A_n \to X$, un entier $n_0$ tel que si $n \geq n_0$ $\left|\vartheta_n - f(X)\right| < \varepsilon$. Mais, comme nous l'avons vu, les contributions des points $A_1, ..., A_{n_0}$ tendent vers 0 lorsque $n \to +\infty$. Ceci prouve le théorème.

Le théorème sera évidemment en défaut si le processus à estimer présente une discontinuité. Dans l'exemple très simple ci-dessous, on ne trouvera pas la même valeur selon que l'on s'approche de 1 par valeurs inférieures et par valeurs supérieures :

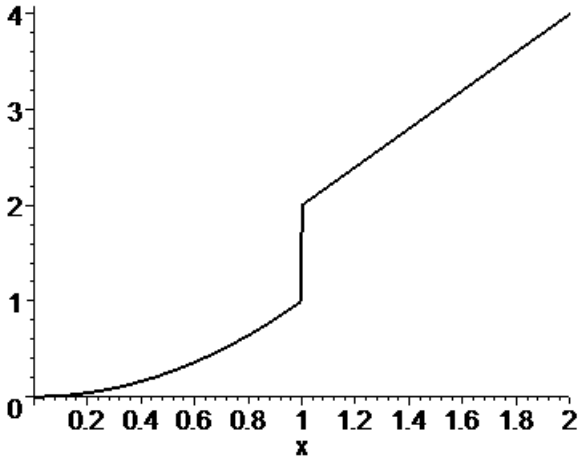


*Figure 23 : cas d'un processus discontinu*

Si on utilise l'EPH à partir de points situés des deux côtés de la discontinuité, elle donnera une prédiction sous la forme d'une loi de probabilité avec deux "bosses" équiprobables, et la valeur moyenne sera le milieu.

# VIII. L'écart entre valeur prédite et valeur réelle

Nous avons défini deux mesures pour comparer les performances du krigeage et de l'EPH :

Soit $N$ le nombre de points reconstruits.
Soit $X$ le vecteur de $N$ termes des points estimés par EPH.
Soit $Y$ le vecteur de $N$ termes des points estimés par krigeage.
Soit $R$ le vecteur de $N$ termes des valeurs réelles de la fonction.

On note respectivement $X_n$, $Y_n$ et $R_n$ le nième coefficient des vecteurs $X, Y$ et $R$.

**Dist_1 :**

Elle représente l'écart entre la valeur moyenne (valeur assignée par la reconstruction) et la valeur réelle :

$Dist_1 = \frac{1}{N}\sum_{n=1}^{N} | X_n - R_n |$ pour l'EPH.

$Dist_1 = \frac{1}{N}\sum_{n=1}^{N} | Y_n - R_n |$ pour le krigeage.

**EQ_2 :**

Elle représente la moyenne des erreurs quadratiques, aussi appelée moyenne des risques quadratiques :

$EQ_2 = \frac{1}{N}\sum_{n=1}^{N} (X_n - R_n)^2$ pour l'EPH.

$EQ_2 = \frac{1}{N}\sum_{n=1}^{N} (Y_n - R_n)^2$ pour le krigeage.

Nous avons comparé ces mesures pour le krigeage et pour l'EPH dans différentes situations-types :

# IX. Prise en compte du degré de variation des données

L'algorithme du krigeage présente l'avantage d'intégrer la structure de dépendance spatiale. Cependant, lorsque l'information en pauvre, ou que les données sont très variables, il ne parvient pas à estimer la variabilité correctement et donne de mauvais résultats. Nous voyons si l'EPH souffre des mêmes problèmes.

## *A. Forte variabilité des données*

### 1. Pour le krigeage

Lorsque les données varient fortement il peut y avoir des difficultés d'estimation. Voici trois exemples illustrant cette difficulté :

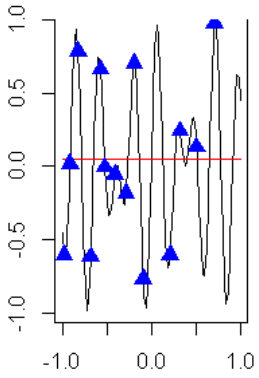


*Figure 24 : Krigeage avec 14 points de mesures*

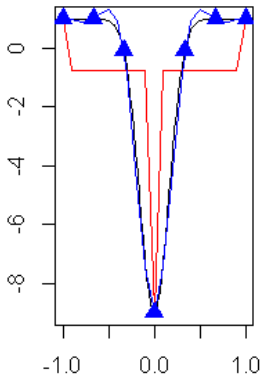


*Figure 25 : Krigeage avec 6 points de mesures*

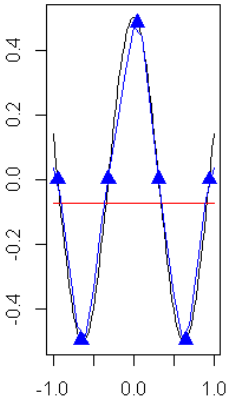

*Figure 26 : Krigeage de la fonction sinus avec 7 points de mesures*

Sur les trois graphiques ci-dessus, la courbe bleue est l'estimation obtenue avec modification manuelle de la portée, la courbe rouge est l'estimation obtenue avec la portée calculée par l'algorithme par EMV (Estimateur du Maximum de Vraisemblance). Dans les deux premières simulations, les données varient fortement, et l'algorithme considère les données comme la réalisation d'un processus gaussien dont les variables aléatoires sont indépendantes. La portée estimée par EMV vaut 0, donc l'estimation donnée par le krigeage correspond à la tendance du processus.

Une mauvaise estimation du processus peut être évitée si les conditions suivantes sont respectées :

- Les observations sont suffisamment proches les unes des autres. Cela permettra à l'EMV de pouvoir capter une dépendance.

- On choisit certains points d'observation comme points à estimer. Le krigeage est un interpolateur exact (démonstration en Annexe), c'est-à-dire que l'estimation faite à un point de mesure correspond la valeur de ce point de mesure. De cette façon, on est sûr que l'interpolation va passer par ces points. La Figure 35 illustre ce résultat. Les observations 1, 4 et 7 correspondent à des points à estimer et ce choix permet d'avoir une estimation grossière malgré une portée nulle.

### 2. Pour l'EPH

Dans l'exemple utilisé plus haut pour le krigeage, l'EPH donne de meilleurs résultats que le krigeage, car il n'y a pas eu de problème d'estimation de paramètre : il n'y a pas de paramètre à estimer dans l'EPH.

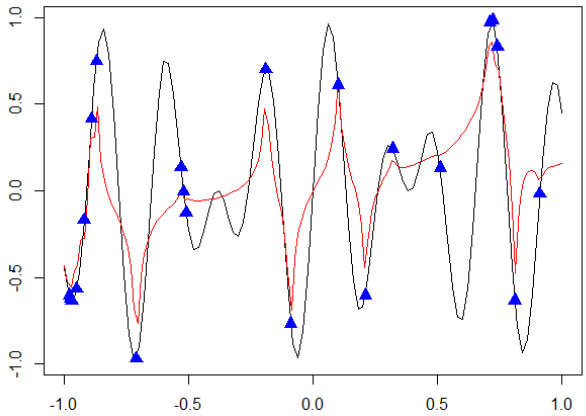


*Figure 27 : EPH sur 20 points très variables*

Cependant l'EPH donne souvent de mauvais résultats lorsque les données présentent de fortes variabilités.

Cherchons à reconstruire la fonction $y = x^2$ sur l'intervalle [-1, 1] à partir de 6 mesures réparties de façon régulière.

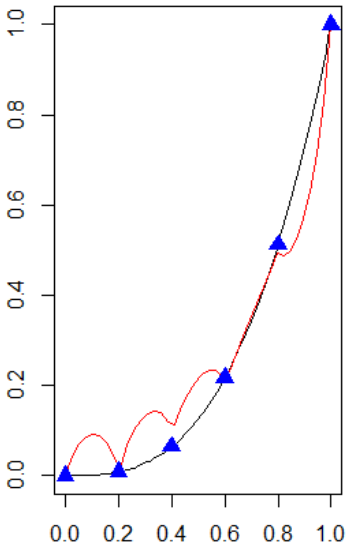


*Figure 28 : reconstruction de 150 points à partir de 6 mesures par l'EPH*

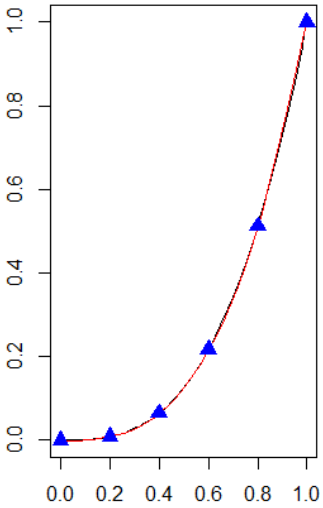


*Figure 29 : reconstruction de 150 points à partir de 6 mesures par krigeage*

L'interpolation devrait normalement être comprise entre les valeurs des points d'observation adjacents, ce qui n'est pas le cas ici. La courbe reconstruite est très irrégulière. L'interpolation faite par krigeage est satisfaisante.

Les mesures de dispersions autour des vraies valeurs sont données dans le tableau suivant ($\times 10^{-6}$) :

| | Dist_1 | RMSE |
|---|---|---|
| Krigeage | 1333 | 5 |
| EPH | 65580 | 6365 |

La distance moyenne à la valeur réelle et l'erreur quadratique moyenne sont meilleures pour le krigeage que pour l'EPH.

Le problème est toujours présent lorsque l'on dispose de plus de mesures. Par exemple pour 20 mesures :

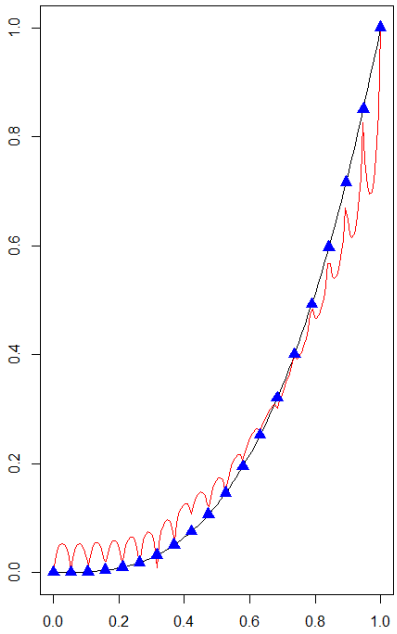


*Figure 30 : reconstruction de 150 points à partir de 20 mesures par l'EPH*

Il s'agit là du problème majeur de l'EPH. La variabilité des données n'a pas été prise en compte ; par conséquent les données situées loin d'un point d'observation sont intégrées dans l'interpolation dans une trop grande proportion. Cela est dû :

– aux poids utilisés pour combiner les lois de probabilité associées à chaque mesures ;

– au principe de transfert de l'information utilisé dans l'EPH.

Toutes les mesures ont un poids non-nul, ce qui est rarement souhaitable. On peut choisir un voisinage de points pour effectuer la reconstruction. Dans l'exemple ci-dessus, cela permet d'obtenir une estimation correcte. De plus cela réduit le temps de calcul.

La construction des poids est la suivante. Supposons qu'il y ait deux observations. L'influence de chacune sur un point à reconstruire est proportionnelle au ratio de la distance à la première mesure divisé par la distance à la deuxième mesure. Ce ratio permet d'introduire le minimum d'information possible (voir [PIT] part III p.167). Cependant les poids ainsi construits donnent en pratique de mauvais résultats.

Dans l'exemple ci-dessus, l'exposant $K=1$ utilisé dans le calcul du poids $\gamma_j = \frac{d_j^{-K}}{\sum_i d_i^{-K}}$ n'est pas convenable car il pondère trop faiblement les données proches par rapport aux données lointaines. Evidemment, en augmentant $K$, nous avons obtenu une interpolation satisfaisante.

La principale différence entre l'EPH et le krigeage réside ici :

- L'EPH ne fait aucune hypothèse a priori sur le degré de variation des données et la pondération des observations est la même quelle que soit la variabilité des données.

- Le krigeage donne à l'utilisateur la responsabilité d'étudier la variabilité des données. C'est l'analyse variographique. L'algorithme intègre alors la structure de dépendance spatiale estimée à l'issue de cette étude pour calculer les poids.

Le modèle de propagation de l'information pose également problème. Rappelons-en d'abord le principe. L'information apportée par une mesure devient de moins en moins précise avec la distance et l'entropie croît linéairement (voir [PIT] part III). L'entropie relie l'information à la loi de probabilité, la distribution associée est donc de plus en plus dispersée. La méthode de l'EPH définit alors l'entropie comme étant maximale en le point le plus éloigné. Cette entropie maximale est celle d'une loi uniforme.

Cependant, en réalité, les mesures qui sont loin du point à estimer peuvent encore apporter une information précise, et à l'inverse, les observations peu éloignées du point à estimer peuvent être totalement indépendantes, et n'apporter aucune information. Et une nouvelle mesure ne doit pas nécessairement ajouter de l'information partout sur le domaine ; cela dépend complètement du degré de variabilité des données.

L'entropie maximale n'est donc pas nécessairement atteinte en le point le plus éloigné de la mesure ; elle peut l'être bien avant, ou bien après.

On peut imaginer un modèle différent où le maximum de l'entropie serait atteint à une distance donnée, définie sur la base de l'étude de la variabilité des données dans le cas où l'on dispose de suffisamment d'observations. Ainsi, la pente de croissance de l'entropie serait modifiée comme sur la Figure 39.

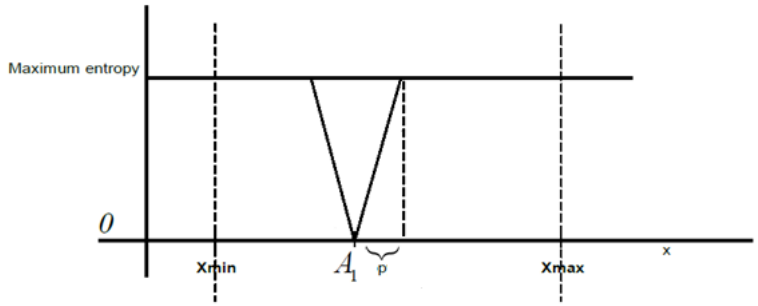


*Figure 31 : croissance modifiée de l'entropie avec la distance*

Dans ce modèle, toutes les mesures situées au-delà d'une distance $p$ à estimer à partir de l'étude des données n'apportent aucune information. Si $p = d_{\max}$ alors c'est le modèle de l'EPH.

Supposer que les mesures suffisamment lointaines n'apportent aucune information est une hypothèse de modèle que l'on s'interdit dans l'EPH

Dans certaines situations spécifiques comme celle présentée ici (forte variabilité), l'EPH peut donner de mauvais résultats ; mais c'est parce qu'il y a une information spécifique qui doit être incorporée dans l'EPH. En effet, l'EPH est un modèle à information minimale, il ne fait pas d'hypothèse a priori sur les données.

### *B. Faible nombre d'observations*

Comme nous l'avions montré dans le premier rapport (IRSN EX10 / 32001814 du 30.05.2014), l'EPH donne de meilleurs résultats que le krigeage lorsque l'information est pauvre. Nous expliquons dans le présent rapport pourquoi le krigeage donne de mauvaises performances lorsque peu d'observations sont disponibles.

#### 1. Pour le krigeage

Prenons un exemple en dimension 1. Soit la fonction :

$$f(x) = 0.2 \times (\sin(5x) + \sin(\sqrt{3}x) + \tanh(20x))$$

On reconstruit 20 points sur $[-1,1]$ à partir de 4 mesures. Pour une première simulation, on choisit deux couples de points de coordonnées symétriques (-0.94,-0.314) et (0.314,0.94). Dans la deuxième simulation, les couples d'observations (-0.94,-0.314), (0.21, 0.82) ne sont pas exactement symétriques. La courbe à reconstruire est en noir, la courbe rouge est l'estimation par krigeage.

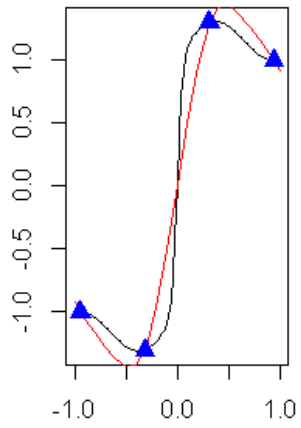


*Figure 32 : krigeage dans le cas où les points sont symétriques*

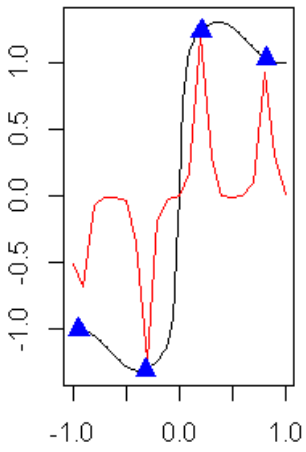


*Figure 33 : krigeage dans le cas où les points ne sont pas symétriques*

La seconde simulation donne de mauvais résultats, car le paramètre de portée a été largement sous-estimé. La répartition des observations influe sur cette estimation. Dans la première simulation, les variables éloignées sont considérées comme très dépendantes, puisque la portée est de 0,43. Dans la deuxième simulation, la portée est presque nulle, ce qui signifie que les variables même proches n'ont aucune influence entre elles. L'algorithme considère donc les observations comme la réalisation d'un processus formé par des variables aléatoires indépendantes.

Dans ce cas, la courbe reconstruite est quasiment constante. En effet, la formule de l'estimateur est $\hat{Y}_x = m + \lambda^t (\mathrm{Y} - \mathrm{m})$, avec $\lambda = \mathrm{K}^{-1}\,\mathrm{k}(\mathrm{x})$. La covariance entre le point à estimer et les observations est nulle car la portée est très faible et elle est très largement dépassée, donc $k(x) = 0$ . Les poids sont alors tous nuls et l'estimateur est finalement égal à la tendance qui dans cet exemple vaut 0.

Nous montrons un cas où le coefficient de portée est trop haut cette fois, en dimension 2. Reprenons l'exemple utilisé dans le premier rapport (IRSN EX10 / 32001814 du 30.05.2014).

On souhaite reconstruire la fonction de Branin sur 400 points à l'aide de 3 observations situées en (0, 0), (1/2, 1), (1, 0) :

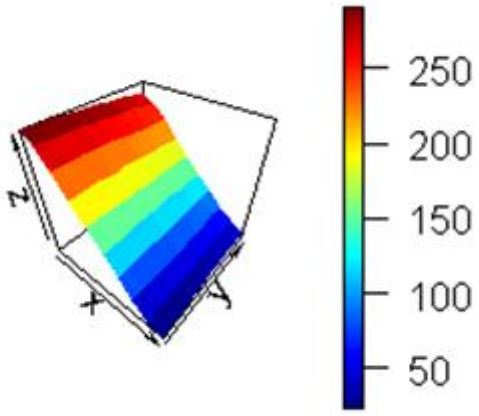


*Figure 34 : Krigeage à partir de 3 observations*

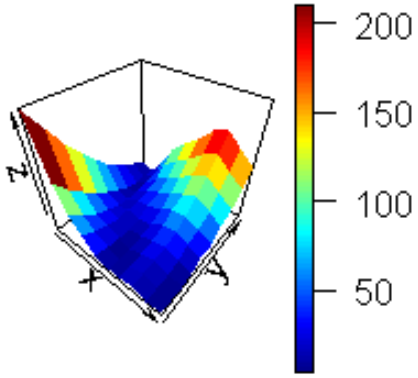


*Figure 35 : Surface à reconstruire, fonction de Branin*

La simulation donne une mauvaise estimation. Les paramètres de portée estimés sont : 0.359 pour la dimension X et 2 pour la dimension Y. Ici le modèle est anisotrope, c'est-à-dire qu'il y a un coefficient de portée différent pour chaque direction de l'espace. Pour le paramètre Y, le coefficient est fort. Etant donné que le domaine de reconstruction est un carré de côté 1, la portée n'est jamais atteinte. La forte dépendance se traduit par une surface constante dans la dimension Y.

Dans une deuxième simulation, le coefficient de portée pour le paramètre Y est volontairement abaissé à 0,6. Ce paramètre aurait pu être fixé correctement à la main dès le départ, grâce à une étude de la variabilité des données.

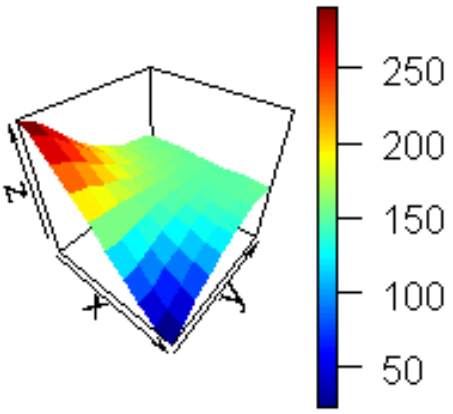


*Figure 36 : Krigeage avec paramètre de portée modifié*

La dépendance spatiale est plus faible, et l'estimation n'est plus constante dans la dimension Y. La surface obtenue est plus fidèle à celle de référence.

Remarquons que, lorsque le modèle ne présente pas de dépendance directionnelle, la portée estimée est nulle et le résultat est un plan horizontal correspondant à la tendance constante du processus.

Nous avons vu deux situations opposées : l'une où l'estimation du coefficient de portée est nulle, l'autre où elle est trop élevée. Ces exemples sont construits à partir de peu d'observations (4 pour le premier, 3 pour le second). Un faible nombre de mesures est insuffisant pour avoir une estimation correcte du paramètre de covariance.

### 2. Pour l'EPH

L'EPH s'accommode très bien du faible nombre de données puisque, contrairement au krigeage, elle ne fait pas de supposition sur la structure de dépendance spatiale. Il n'y a donc pas de risque de mal estimer les paramètres d'un quelconque modèle. Reprenons l'exemple utilisé pour le krigeage.

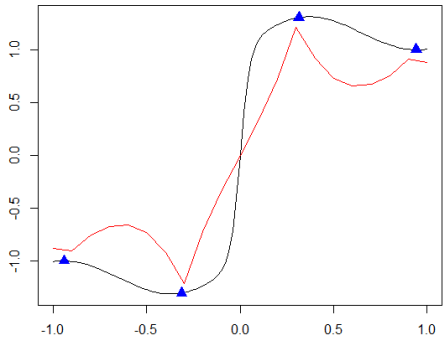


*Figure 37 : EPH avec 4 points de mesures*

La courbe calculée par l'EPH correspond relativement bien à la fonction à reconstruire.

# X. Gestion des données groupées

Les mesures sont parfois regroupées sur une petite zone de l'espace. Elles peuvent être surreprésentées dans l'interpolation par rapport aux observations isolées. Nous comparons la capacité de l'EPH d'une part et du krigeage d'autre part à prendre en compte la redondance des données.

## *A. Gestion des données groupées par l'EPH*

Un désavantage majeur de l'EPH est qu'elle ne gère pas correctement les regroupements de mesures. Illustrons ce point sur un exemple simple. Soient $B_1, B_2, B_3, B_4, B_5$ les mesures telles que $B_1, B_2, B_3, B_4$ sont regroupées, et $X$ le point à reconstruire.

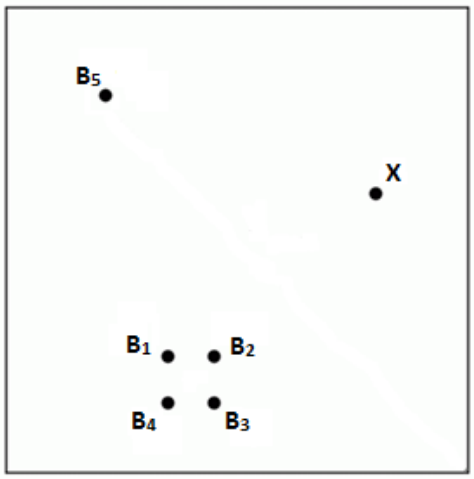


*Figure 38 : mesures groupées*

Les distances entre $X$ et les mesures $B_1, B_2, B_3, B_4, B_5$ sont les suivantes :

$$d(B_1, X) = 3,$$
$$d(B_2, X) = 2.9,$$
$$d(B_3, X) = 3,$$
$$d(B_4, X) = 3.1,$$
$$d(B_5, X) = 3,$$

Chaque observation envoie en $X$ une information sous la forme d'une loi de probabilité gaussienne discrétisée :

$$p_{B_n}(X) = c \exp\left\{ -\frac{\left(t_j - \vartheta_n\right)^2}{2\sigma_n^2} \right\},$$

où $c$ est une constante de normalisation et $\sigma_n$ est la variance.

On combine les sources individuelles $B_1, B_2, B_3, B_4, B_5$, le résultat est une loi de probabilité définie par :

$$P_n(X) = \gamma_{B_1} p_{B_1}(X) + \gamma_{B_2} p_{B_2}(X) + \gamma_{B_3} p_{B_3}(X) + \gamma_{B_4} p_{B_4}(X) + \gamma_{B_5} p_{B_5}(X),$$

où les poids $\gamma_{B_i}$ sont définis par :

$$\gamma_j = \frac{{d_{B_5}}^{-1}}{{d_{B_1}}^{-1} + {d_{B_2}}^{-1} + {d_{B_3}}^{-1} + {d_{B_4}}^{-1} + {d_{B_5}}^{-1}}$$

Ce qui donne dans notre exemple :

$$\begin{aligned}
\gamma_{B_1} &= 0.1675, \\
\gamma_{B_2} &= 0.1733, \\
\gamma_{B_3} &= 0.1675, \\
\gamma_{B_4} &= 0.1621, \\
\gamma_{B_5} &= 0.1675,
\end{aligned}$$

Le poids de l'observation $B_5$ est très faible par rapport au poids cumulé des 4 autres mesures, qui est de 0,83. La densité de probabilité résultante $P_n(X)$ va être composée de deux bosses, une première associée à la mesure $B_5$, et une seconde bien plus haute associée aux quatre mesures groupées. Le point X étant quasiment à égale distance des 5 mesures, les deux bosses auraient dû être de même taille.

Illustrons le problème que cela engendre sur l'interpolation créée. On répartit régulièrement sur le domaine 9 points d'observation et on ajoute 8 mesures au voisinage du point de coordonnées (0,0).

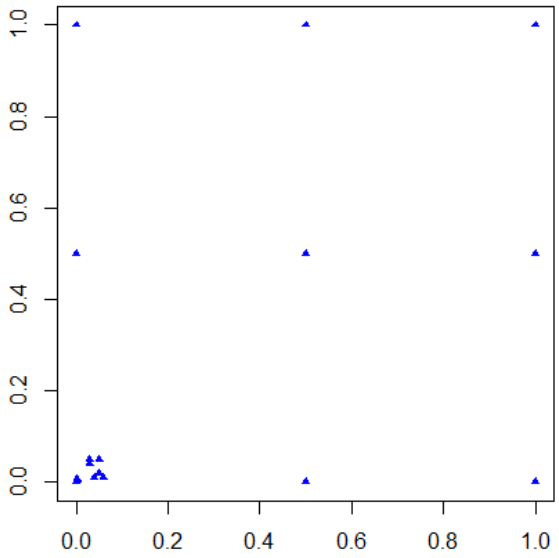


*Figure 39 : répartition des mesures sur le domaine*

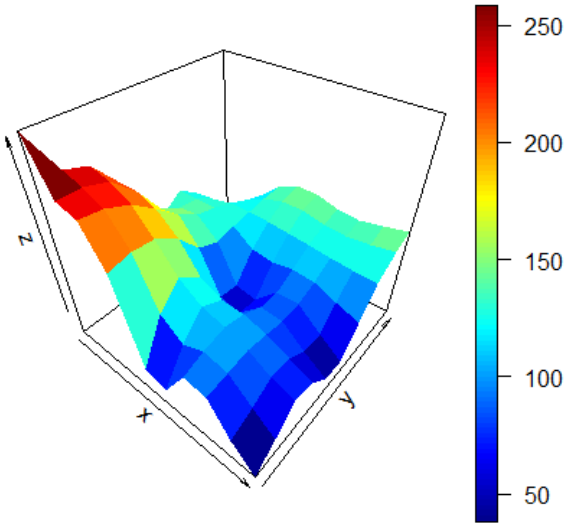


Figure 40*: EPH sur des données groupées*

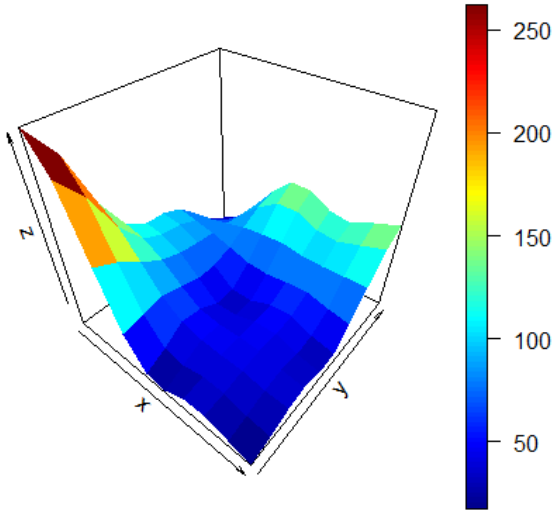


*Figure 41 : EPH sur des données non-groupées*

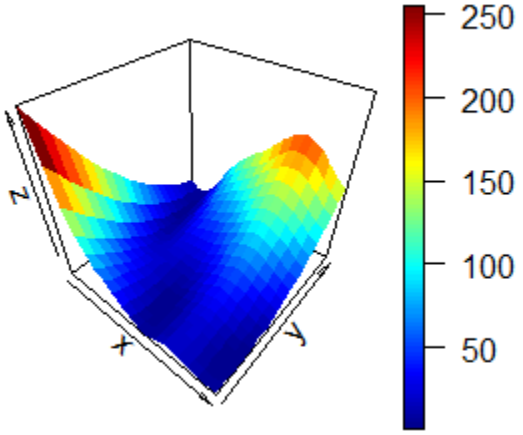


*Figure 42 : fonction de Branin*

L'aspect de la surface reconstruite par EPH est bien différent lorsqu'il y a des données groupées. La surface reconstruite prend de fortes valeurs sur toute la zone proche du point (0,0) dans un rayon de 0,5 ce qui est très large. Les données groupées ont été surreprésentées.

La technique généralement utilisée pour diminuer le poids des données groupées est la méthode des cellules. Elle consiste à diviser l'espace en cellules de tailles identiques. En dimension 1, ce sont des intervalles, en dimension deux des rectangles, en dimension 3 des pavés, etc. Pour chaque mesure $i$, on compte le nombre $n_i$ d'observations présentes dans la même cellule. On note $d$ le nombre total de cases contenant au moins une mesure. Le poids initial $\gamma_i$ de la mesure $i$ est alors remplacé par :

$$\omega_i = \frac{1}{n_i d}\gamma_i$$

Plus généralement, si l'on dispose d'un plus grand nombre de mesures d'un côté du maillage que de l'autre, cela se verra dans l'interpolation qui accordera plus d'importance à ce côté. Il ne s'agit pas d'une erreur : il y a plus d'information d'un côté. S'il y a d'un côté du domaine un grand nombre d'observations, chacune égale à 10, et de l'autre côté seulement quelques observations égales à 5, alors l'interpolation au milieu du domaine doit être plus proche de 10 que de 5. Nous expliquons cela en Annexe F. Rappelons que l'EPH prend en compte uniquement la distance aux points de mesure, et non la position de ces mesures : qu'elles soient groupées ou non n'intervient pas.

Montrons le comportement de l'EPH en cas de répartition inégale des données. Nous réalisons le test avec la fonction $f(x) = 1 - 10 \cdot \exp(-20 \cdot x^4)$ en ajoutant 40 mesures sur la partie constante de la fonction à gauche et à droite de l'intervalle, et en gardant très peu de mesures au milieu.

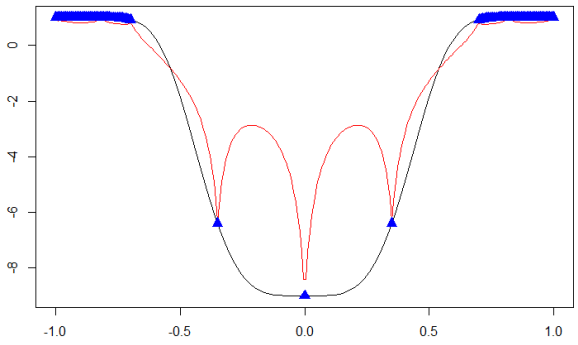


*Figure 43 : reconstruction de 150 points à partir de 85 mesures*

L'interpolation effectuée au milieu du domaine est biaisée par le grand nombre d'observations sur les bords du maillage. Lorsqu'il y a seulement 2 mesures au lieu de 40 sur les bords, l'interpolation est améliorée :

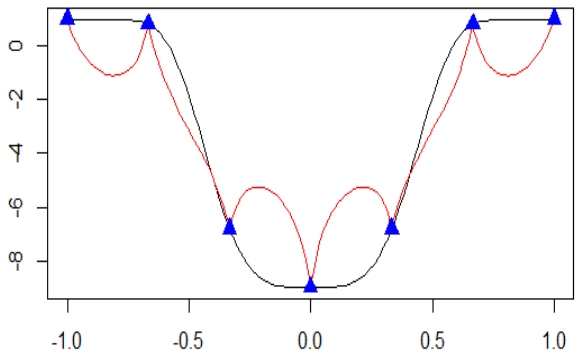


*Figure 44 : reconstruction de 150 points à partir de 7 mesures par l'EPH*

Notons que, dans cet exemple, le krigeage ne fonctionne pas, car les mesures sont trop rapprochées et il y a un problème lors de l'inversion de la matrice de covariance ; nous y reviendrons dans la section VII traitant de la qualité des procédures numériques.

### *B. Gestion des données groupées par le krigeage*

Un des avantages du krigeage est de ne pas être biaisé par la non-uniformité de la répartition des mesures. On reprend l'exemple précédent en utilisant le krigeage cette fois.

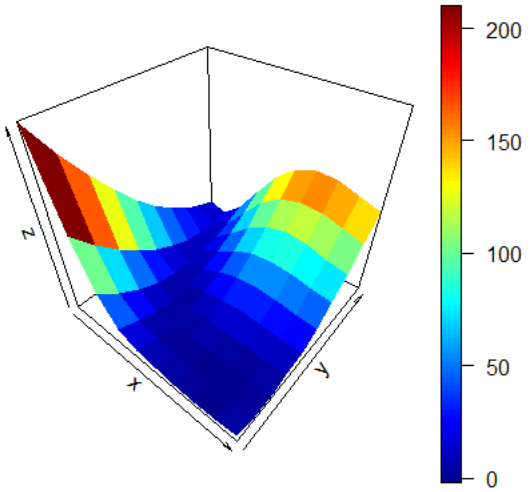


*Figure 45 : krigeage sur des données groupées*

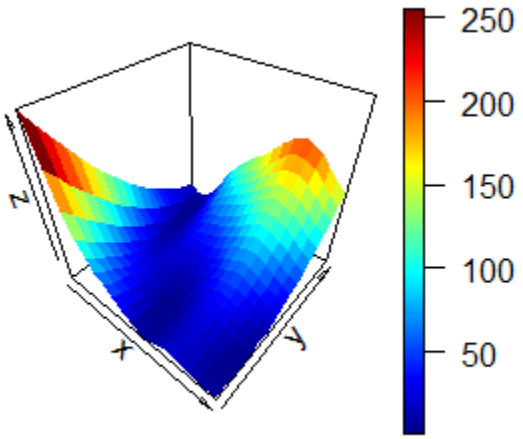


*Figure 46 : fonction de Branin*

L'interpolation réalisée est tout à fait convenable, contrairement à celle de l'EPH.

Les mesures de dispersion autour des vraies valeurs pour l'EPH et le krigeage sont donc reportées dans le tableau suivant :

| | Dist_1 | RMSE |
|---|---|---|
| Krigeage | 10 | 14 |
| EPH | 63 | 89 |

Nous montrons maintenant comment le krigeage gère cette situation. Soit l'exemple suivant, dans lequel il y a un groupe d'observations : les points 1 jusqu'à 7. On calcule l'estimation au point P. On ajoute une huitième observation, de façon à ce que les deux groupes d'observation soient environ à égale distance du point P. La figure ci-dessous représente la situation décrite.

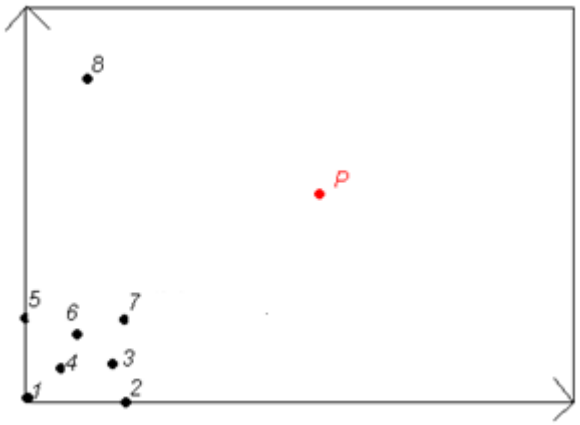


*Figure 47 : répartition des mesures sur le domaine*

La somme des poids des mesures appartenant au groupe d'observations vaut 0,35 et le poids de l'observation 8 est de 0,34. Le poids du groupe est donc le même que celui de l'observation isolée, ce qui est souhaitable puisqu'ils sont chacun à la même distance du point P.

Avec un modèle de covariance de type exponentiel, il se produit un "effet écran". Cela signifie qu'une observation dans le groupe a un poids très élevé alors que les observations aux alentours ont un poids quasiment nul. Ici l'observation 7 à un poids de 0,47 alors que les observations 1, 2, 3, 4, 5, 6 ont un poids négatif très faible. Les données autour de 7 sont négligées par l'algorithme car redondantes.

En revanche, avec un noyau de type Matérn, l'observation isolée a un poids de 0,22 et la somme des poids du groupe est de 0,76. Le krigeage attribue des poids négatifs à certaines observations du groupe, ce qui permet d'éviter une surestimation du groupe.

# XI. Robustesse aux données aberrantes

Les données aberrantes impactent l'interpolation sur une large partie du domaine, aussi bien pour le krigeage que pour l'EPH. Nous incorporons une valeur aberrante de 2500 au point de coordonnées (1/3, 1/3) parmi les 16 mesures. Les 15 autres mesures ont des valeurs inférieures à 400.

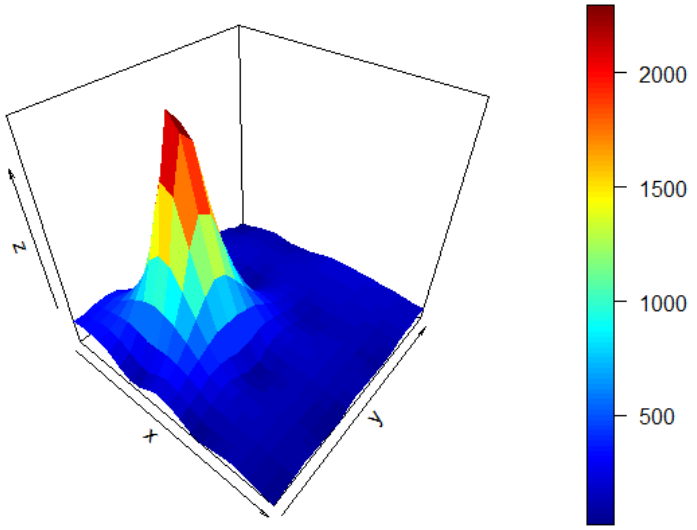


*Figure 48 : EPH avec présence d'une valeur aberrante*

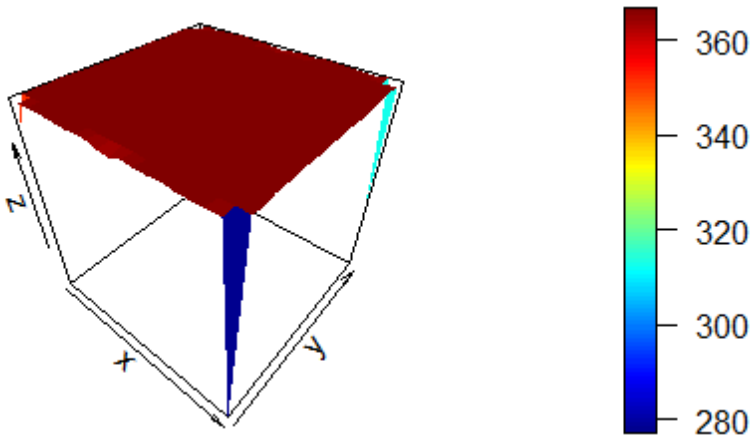


*Figure 49 : krigeage avec présence d'une valeur aberrante*

Pour l'EPH, la valeur aberrante a une influence sur l'interpolation sur la totalité du maillage. Le krigeage produit une surface plate correspondant à la moyenne du processus. La portée n'est pas correctement estimée à cause de la forte variabilité des données due au point aberrant. Il est nécessaire de modifier manuellement le paramètre de portée, ce qui donne l'interpolation présentée ci-dessous :

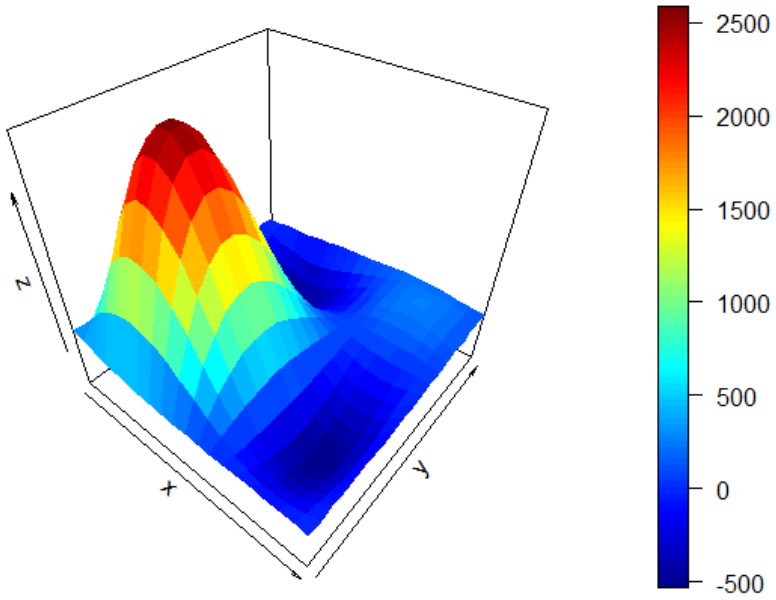


*Figure 50 : krigeage avec paramètre de portée ajusté manuellement*

La valeur aberrante entraîne une large modification de l'interpolation sur l'ensemble du maillage, faussant l'intégralité de l'estimation.

L'EPH obtient de meilleurs résultats que le krigeage, car l'influence du point aberrant est plus restreinte. Les performances obtenues par le krigeage et l'EPH sont reportées dans le tableau suivant :

| | Dist_1 | MSE |
|---|---|---|
| Krigeage | 495 | 664661 |
| EPH | 244 | 212149 |

Les performances de l'EPH sont nettement supérieures à celles du krigeage.

Lorsque le nombre de points de mesures est plus grand, l'EPH est nettement plus robuste aux données aberrantes que le krigeage. Nous effectuons le test avec 121 mesures réparties régulièrement sur le domaine.

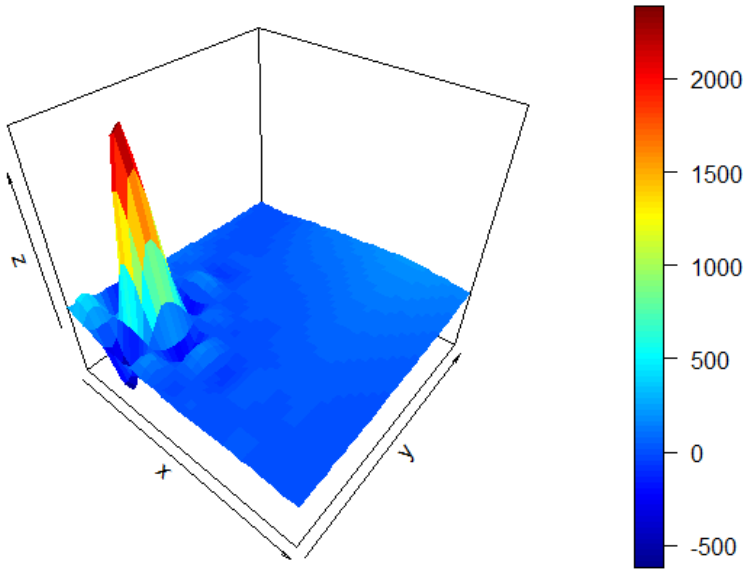


*Figure 51 : krigeage avec 121 points*

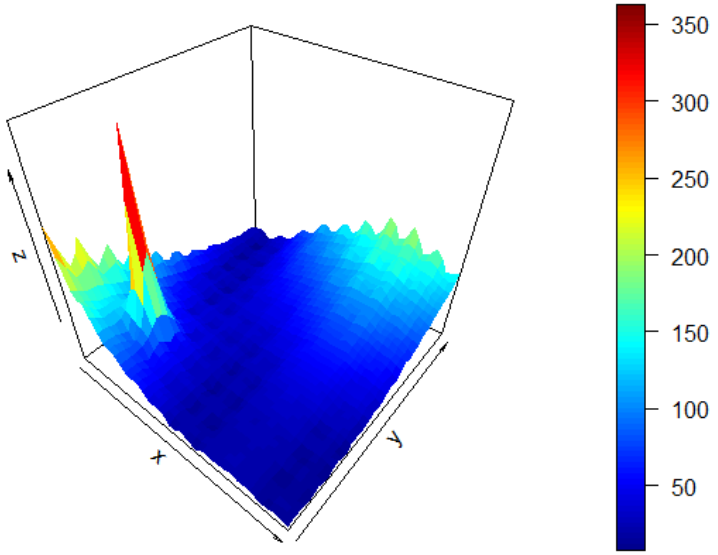


*Figure 52 : EPH avec 121 points*

L'EPH reconstruit correctement la fonction, excepté au voisinage proche du point aberrant. Bien évidemment, on peut encore réduire la zone d'influence du point aberrant si on prend la médiane de la loi de probabilité fournie par l'EPH au lieu de son espérance. L'interpolation réalisée par le krigeage reste en revanche totalement faussée par le point aberrant sur un large périmètre.

La détection de valeurs aberrantes est une étape de préparation des données qui se fait avant l'utilisation de l'algorithme d'interpolation. Ce n'est pas le rôle d'une méthode de reconstitution de données que de détecter des données aberrantes. Il est donc normal que les données extrêmes soient intégrées dans l'interpolation au même titre que les autres observations, et que cela influe sur la surface reconstruite.

Il existe toutefois une version du krigeage robuste aux données extrêmes nommée krigeage robuste, permettant de donner un faible poids aux données aberrantes ; elle est présentée dans [CRE], p. 144.

# XII. Qualité des procédures numériques

Dans certaines situations, le krigeage donne un résultat aberrant. Cela est parfois dû à une mauvaise estimation du coefficient de portée. De plus la matrice utilisée dans le krigeage peut être mal conditionnée et causer des instabilités numériques. Nous examinons les situations entraînant de telles instabilités et voyons si l'EPH souffre des mêmes problèmes dans ces situations.

## *A. Qualité des procédures numérique dans le krigeage*

### 1. Mauvais conditionnement de la matrice

Le conditionnement permet de mesurer la dépendance de la solution d'un système numérique par rapport aux données. Soit un système linéaire $Ax = b$. Si la matrice A est mal conditionnée, alors une petite variation sur b entraîne une grande variation sur la solution $x$. Le nombre de conditionnement d'une matrice est défini par la formule :

$$cond(A) = \| A \|_p \| A^{-1} \|_p$$

Dans le cas du krigeage, le système à résoudre est $K\lambda = k(X)$ où $K$ est la matrice de covariance et $k(X)$ est la covariance entre le point à reconstruire $X$ et chaque observation. Si la matrice $K$ est mal conditionnée, une faible variation du vecteur $k(X)$ peut entraîner une forte variation des poids de krigeage.

#### Situations menant à un mauvais conditionnement

Trois situations mènent à un mauvais conditionnement :

- lorsque la densité des points d'observation est grande ;
- lorsque des points sont trop rapprochés ;
- lorsque la portée est grande par rapport aux écarts entre les observations.

La preuve de ces trois affirmations est donnée en annexe.

Nous montrons avec trois exemples les problèmes qui surviennent lorsque le conditionnement est mauvais. D'une part, lorsque l'on perturbe la localisation des points à estimer ou des observations, l'interpolation réalisée est instable. De plus, il y a des erreurs d'arrondis qui causent une interpolation aberrante.

### Exemple 1 : erreur sur la localisation des points à estimer

Soient $x_1$, $x_2$, $x_3$, $x_4$, quatre observations situées aux coordonnées -1, -0.16, -0.15, -0.97 respectivement. On applique le krigeage sur ces données. La matrice de covariance calculée par l'algorithme est mal conditionnée :

$$K = \begin{pmatrix} 1,49 & 1,48 & 1.47 & 0,9 \\ 1,48 & 1,49 & 1,48 & 0,99 \\ 1,47 & 1,48 & 1,49 & 1,06 \\ 0,9 & 0,99 & 1,06 & 1,49 \end{pmatrix}$$

Le nombre de conditionnement de cette matrice est de $6\times10^5$. La reconstruction d'un point X se fait en effectuant la combinaison linéaire des observations. Les poids de cette combinaison s'obtiennent en résolvant le système linéaire suivant :

$$K\lambda = k(X)$$

où K est la matrice de covariance, et $k(X)$ le vecteur des covariances entre X un point à reconstruire et chaque observation.

Considérons le vecteur de covariance $k(X)$ associé au point $X = -0.25$. Le vecteur poids calculé est $\lambda = (0,0,0,1)$. Supposons que $k(X)$ ne soit connu qu'avec trois chiffres significatifs, il y a une perturbation du vecteur de l'ordre de $10^{-3}$. Le nouveau vecteur poids vaut alors $\tilde{\lambda} = (-172,371, -203, 4.5)$. La variation du vecteur poids est très grande, ce qui fausse l'estimation au point $X$ : la valeur retournée est 6 alors que la valeur exacte est -1.3.

Lorsque le conditionnement est mauvais, une imprécision sur les coefficients de la matrice ou du vecteur $k$ peut perturber significativement l'estimation.

### Exemple 2 : erreur sur la localisation des observations

Les observations sont réparties aux points -1, -0.45, -0.05, 0.05, 0.45, 1. Les observations 3 et 4 sont très rapprochées. Ensuite on perturbe légèrement la localisation des mesures, la valeur des observations est inchangée. Les nouvelles positions sont -0.98, -0.47, -0.02, 0.02, 0.42, 1.01. Le résultat du krigeage avant et après perturbation est affiché en rouge et bleu respectivement, sur la figure ci-dessous :

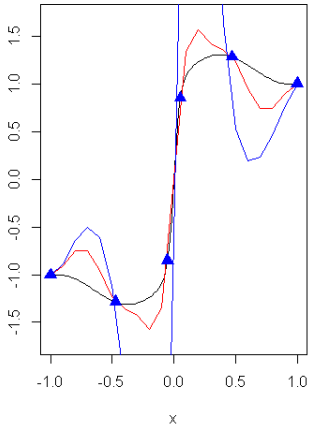


*Figure 53 : instabilité du krigeage suite à une erreur sur la position des points*

Les deux interpolations sont très différentes. Le krigeage n'est donc pas stable pour cette configuration.

Les observations 3 et 4 sont très rapprochées, et le conditionnement de la matrice de covariance est mauvais. Pour la deuxième simulation, on perturbe légèrement les positions des mesures, ce qui entraîne une perturbation sur la matrice de covariance. Les poids ont donc été fortement modifiés. Ils sont reportés dans le tableau ci-dessous pour le point x=-0.1 :

| | Poids 1 | Poids 2 | Poids 3 | Poids 4 | Poids 5 | Poids 6 |
|---|---|---|---|---|---|---|
| Simulation 1 | 0 | 0.04 | 1.2 | -0.32 | 0.01 | 0 |
| Simulation 2 | 0 | 0.06 | 4.2 | -3.7 | 0.02 | 0 |

Considérons maintenant la reconstruction de la fonction de Branin à partir de 16 mesures dont 8 sont concentrées autour du point (0,0). Dans la deuxième simulation, on introduit une légère perturbation sur certaines positions du groupe. Cette perturbation engendre une interpolation aberrante. La surface reconstruite avec et sans perturbation est affichée ci-dessous :

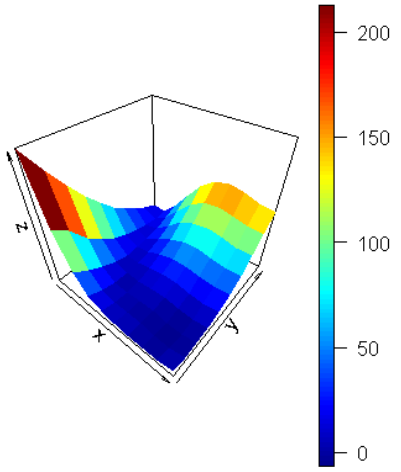


*Figure 54 : surface reconstruite par krigeage sans perturbations*

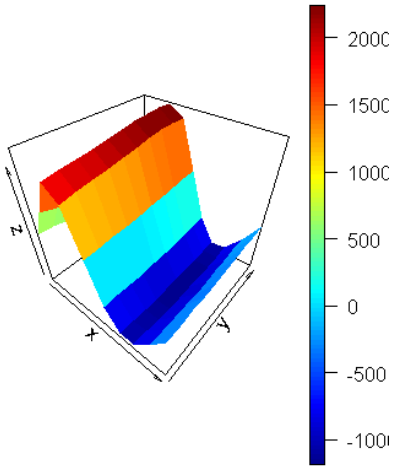


*Figure 55 : surface reconstruite par krigeage avec perturbations*

Cette instabilité s'explique par une forte variation des poids associés au groupe d'observations. Les poids négatifs entraînent une surestimation et une sous-estimation extrêmement exagérées.

Une telle instabilité n'a pas été observée lorsque les données ne sont pas groupées.

### Exemple 3 : erreurs d'arrondis

Plus le nombre de conditionnement de la matrice est élevé, plus les ordres de grandeur des nombres intervenant dans l'inversion de cette matrice sont importants. Or un nombre ne possède qu'un nombre fini de chiffres significatifs dans la mémoire d'un ordinateur. La manipulation d'ordres de grandeurs différents peut amener à des erreurs d'arrondis qui vont se propager au cours de l'inversion de K, et donc fausser le calcul du vecteur poids. Voici un exemple illustrant cette difficulté :

On dispose de 9 observations et on utilise un modèle de covariance gaussien, de portée 2,89. Dans la première simulation, les observations sont réparties uniformément. Le conditionnement vaut $8\times10^{16}$ pour cette répartition. Dans la deuxième simulation, on décale légèrement les points de mesures. Le conditionnement est alors de $2\times10^{17}$ . La portée est trop importante par rapport à l'écart entre les observations ce qui entraîne un fort conditionnement. La matrice est mal conditionnée car ses colonnes sont presque identiques (voir démonstration en Annexe).

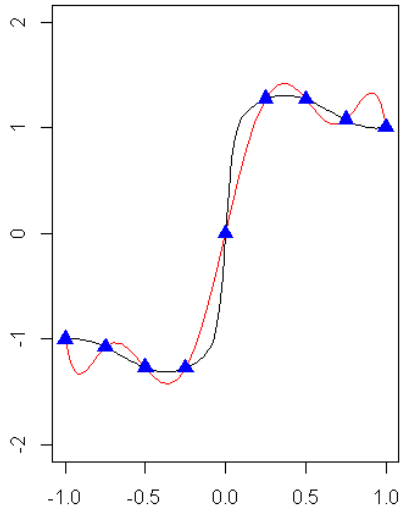


*Figure 56 : simulation avec répartition uniforme des observations*

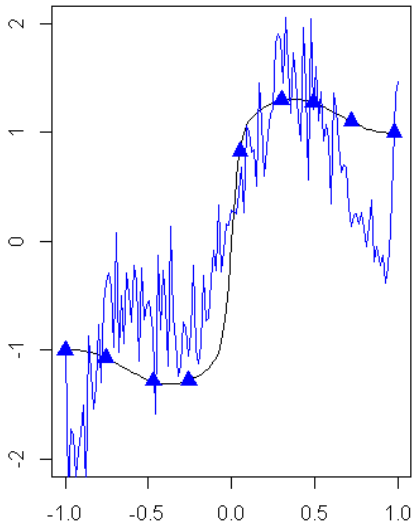


*Figure 57 : simulation avec décalage des points de mesure*

La deuxième simulation est très irrégulière à cause d'instabilités numériques. Prenons par exemple l'estimation du premier point de mesure. Il doit être interpolé exactement, puisqu'on se trouve en un point d'observation, pourtant l'estimation est différente. Les poids n'ont pas été calculés correctement. Pour la première simulation, l'erreur est beaucoup plus faible.

Pour limiter ces instabilités numériques, on peut introduire une petite perturbation sur la diagonale de la matrice de covariance. Dans les deux situations précédentes, si on apporte un effet pépite de $10^{-9}$ le conditionnement passe de $10^{16}$ à $10^{10}$.

Les résultats obtenus sont très différents des résultats sans effet pépite.

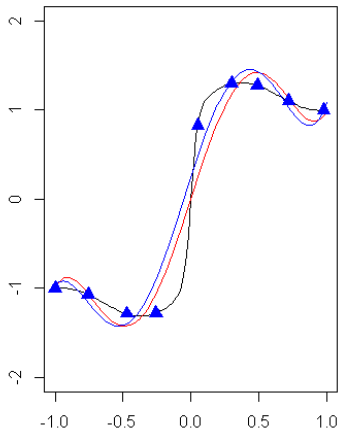


Figure 58: simulation avec effet pépite de e-9

L'estimation du premier point est exacte. La courbe rouge est l'estimation obtenue pour la répartition uniforme et la courbe bleue est celle pour les points décalés. Les deux résultats sont très proches et ne présentent pas d'irrégularités.

### 2. Non-inversibilité de la matrice

Lorsque les mesures sont très rapprochées, comme sur la figure ci-dessous, la matrice peut ne pas être inversible.

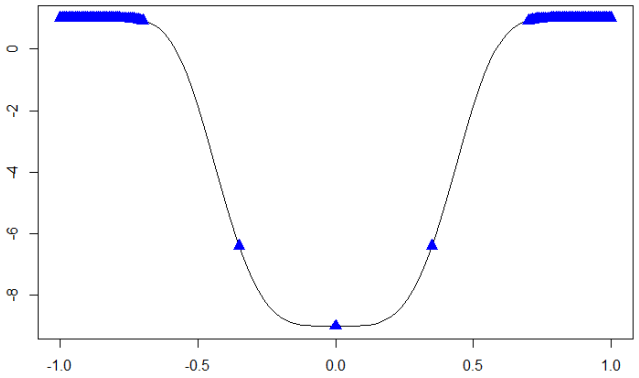


*Figure 59 : répartition des mesures sur le domaine*

Certaines colonnes de la matrice de covariance sont quasiment identiques et linéairement dépendantes, ce qui a entraîné un bug lors de l'inversion de cette matrice.

Avec seulement 6 mesures l'algorithme du krigeage donne une bonne interpolation.

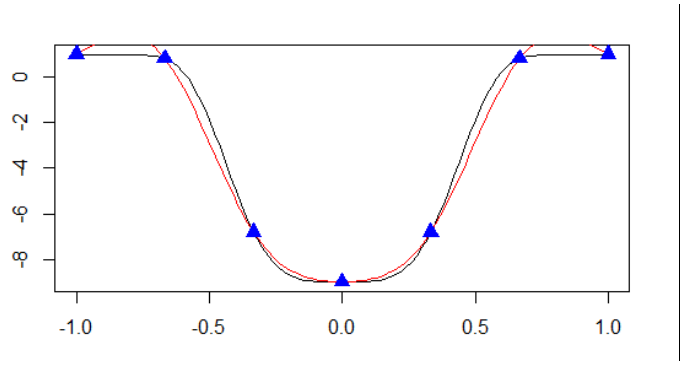


*Figure 60 : reconstruction de 150 points à partir de 7 mesures par krigeage*

Il y a également un bug numérique lorsque l'on fixe une portée trop forte. Par exemple, nous avons tenté de reconstruire la fonction de Branin à partir de 64 observations réparties régulièrement sur le domaine. Nous avons forcé la portée à 5. La covariance entre les observations devient très forte et les colonnes de la matrice sont presque identiques et linéairement dépendantes à nouveau. La matrice est presque non-inversible (singulière), et le programme ne parvient pas à effectuer la décomposition de Cholesky nécessaire à l'inversion de la matrice. Dans le logiciel R, nous obtenons le message d'erreur suivant : "Error in chol.default(C) : the leading minor of order 47 is not positive definite"

### 3. Négativité des poids

Les poids de l'estimateur et les valeurs ne sont pas contraints à être positifs dans le krigeage. Lorsque des poids sont négatifs, cela peut entraîner :

- Une interpolation négative, ce qui n'a pas de sens pour certaines applications, par exemple une concentration ne peut jamais être négative ;

- Des valeurs de sorties supérieures à la mesure la plus grande ou inférieures à la plus faible. Cela est surprenant car on ne peut pas inventer des données qui n'ont jamais été observées par le passé. Cependant, cela peut être considéré comme un avantage pour les utilisateurs qui souhaitent une interpolation très douce pour laquelle les extrema ne sont pas atteints en les points d'observation ;

- Des instabilités : une faible modification de la position des mesures engendre une interpolation très différente [CHA].

## *B. Qualité des procédures numérique dans l'EPH*

La méthode EPH étant fondamentalement différente de celle du krigeage, elle ne souffre pas des instabilités numériques que connaît le krigeage. Les problèmes de type non-inversibilité de la matrice sont bien évidemment exclus puisqu'il n'y a pas de matrice à inverser. Nous

examinons le comportement de l'EPH lorsqu'il y a un faible nombre d'observations ou une grande variabilité des données.

# XIII. Temps de calcul

Souvent la justification principale de l'utilisation d'un algorithme d'interpolation est de faire gagner du temps. En effet, il peut être très coûteux en temps d'effectuer le calcul d'une variable de sortie pour chaque combinaison de paramètres d'entrée possible. On utilise donc un algorithme d'interpolation afin de déduire à partir de quelques calculs la valeur de la variable de sortie sur l'intégralité du maillage. Il est donc indispensable dans ce cas que la reconstruction soit rapide et au minimum qu'elle soit plus rapide que le calcul lui-même.

## *A. Pour le krigeage*

L'opération la plus coûteuse en temps pour le krigeage est le calcul de l'inverse de la matrice de covariance. Dans le cas où on dispose de 9 mesures, il faut inverser une matrice de dimension $9\times9$. Cette opération est très rapide. Il suffit ensuite de multiplier cette matrice avec les vecteurs $K_0$ de taille $9\times1$pour chaque point à reconstruire. Il faut donc effectuer 400 produits matrice-vecteur. Aucune discrétisation de la plage de valeurs de sortie n'est faite dans le krigeage. Si le nombre de valeurs que peut prendre un paramètre est grand, le temps de calcul sera plus rapide pour le krigeage que pour l'EPH.

## *B. Pour l'EPH*

Le temps de calcul de l'EPH augmente avec le nombre de points à reconstruire et la précision de la discrétisation de la variable de sortie.

Prenons un exemple où il faut reconstruire 400 points. On dispose de 16 mesures et la variable de sortie varie entre 0 et 500 avec un pas de 0,1. Pour chaque point à reconstruire il est nécessaire de calculer la distribution générée par chaque observation. La loi de probabilité est une gaussienne centrée en la valeur de l'observation. Cette densité de probabilité prend des valeurs quasi-nulles sur une large plage de valeurs, il n'est donc pas nécessaire de la calculer sur tout l'intervalle.

La taille de cette plage dépend de la distance de l'observation avec le point à estimer : lorsque l'observation est proche du point à reconstruire, la loi de probabilité est très concentrée et est presque nulle sur une large plage de valeurs, à l'inverse si elle est éloignée, la loi est très peu concentrée.

Dans l'exemple qui suit, nous discrétisons la variable de sortie en 5 000 valeurs. Dans le programme de l'EPH en R, nous avons calculé la taille de la plage de valeurs sur laquelle la loi de probabilité prend une valeur significative, et elle est en moyenne de 150.

Il faut donc réaliser au total $400\times16\times150=9\times10^5$ opérations. Nous avons mesuré un temps de calcul pour effectuer ces $9\times10^5$ opérations de 33 secondes. Le temps d'exécution du krigeage pour réaliser la même tâche est de quelques millisecondes, avec un processeur Intel® Core™ i7-4770, 3.40 GHz.

Le temps de calcul augmente avec le nombre de paramètres à prendre en compte. Ajoutons à l'exemple ci-dessus un paramètre, compris entre 0 et 500 avec un pas de 0,1. Il faut alors multiplier le nombre de calculs à nouveau par 150, ce qui représente plus d'une heure de calculs.

Le temps de calcul explose également lorsqu'on incorpore les incertitudes. Choisissons un nombre de runs $M=10^6$. Pour chaque run il faut reprendre la construction de l'EPH depuis le début ce qui fait grossièrement $10^{10}$ calculs.

Si l'on ne souhaite obtenir qu'une simple interpolation, sans incertitude ni loi de probabilité associée, alors appliquer la méthode de l'EPH revient (une fois le coefficient $\lambda$ déterminé) à faire une pondération inverse de la distance. Cette méthode est extrêmement simple à mettre en œuvre et nécessite un temps de calcul très court.

### *C. Comparaison des temps de calcul*

Les temps de calculs en millisecondes pour reconstruire la fonction de Branin sont reportés dans le tableau suivant :

| | 16 mesures 400 points | 16 mesures 1600 points | 64 mesures 400 points |
|---|---|---|---|
| Krigeage | 3 | 9 | 7 |
| EPH | 33772 | 133140 | 122360 |

Le krigeage est plus rapide que l'EPH.

# XIV. Références


[PIT] Olga Zeydina et Bernard Beauzamy : Probabilistic Information Transfer. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA. ISBN: 978-2-9521458-6-2, ISSN : 1767-1175. Relié, 208 pages, mai 2013.

[ICAPP] Bernard Beauzamy, Hélène Bickert, Olga Zeydina (SCM), Giovanni Bruna (IRSN) : Probabilistic Safety Assessment and Reliability Engineering: Reactor Safety and Incomplete Information; Proceedings of ICAPP 2011 Nice, France, May 2-5, 2011 Paper 11399

[CRE] Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

[SMO] Sándor Molnár, *On the convergence of the Kriging method*, 1983

[FAZ] Istivaan Fazkas, Alexander Kukush, *Kriging and measurement errors*, University of Debrecen hongrie, Université de Kiev, 2000

[CHA] Pierre Chauvet, *Réflexions sur les pondérateurs négatifs du krigeage*, 1987, Disponible sur : <http://cg.ensmp.fr/bibliotheque/public/CHAUVET_Publication_00583.pdf> [consulté le 16/07/2015]

[KI] Denis Marcote ; *Krigeage par indicatrices*, Polytechnique Montréal, Disponible sur : <http://www.groupes.polymtl.ca/geo/marcotte/glq3401geo/chapitre7.pdf> [consulté le 16/07/2015]

[CHI] Jean-Paul Chiles, *Géostatistique des phénomènes non stationnaires*, 1977 Disponible sur : <http://cg.ensmp.fr/bibliotheque/public/CHILES_These_00475.pdf> [consulté le 11/08/2015]

[DEV] Fabrice Deverly, *Echantillonnage et géostatistique* 1984, Disponible sur : <http://cg.ensmp.fr/bibliotheque/public/DEVERLY_These_00463.pdf> [consulté le 11/08/2015]

[FLO] J. N. Paoli, *Utilisation d'un variogramme flou dans une méthode d'agrégation sémantique*, 2006, Disponible sur : <https://www.lirmm.fr/~strauss/Publications/VariogrammeLFA06.pdf> [consulté le 11/08/2015]

# XV. Annexe

## *A. Démonstration de la linéarité de l'estimateur dans le cas gaussien*

**Théorème.** – *L'estimateur* $\hat{x}(z)$ *fonction des observations* $z$ *qui minimise l'erreur quadratique* $E\left(\left(x(z_0)-\hat{x}(z)\right)^2\right)$ *est :*

$$\hat{x}(z)=E(x(z_0)\mid z)$$

**Démonstration du théorème (inspirée de [CRE] page 110)**

L'erreur quadratique s'écrit :

$$E\left(\left(x-\hat{x}(z)\right)^2\right)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(x-\hat{x}(z)\right)^2 f_{x,z}(x,z)dxdz,$$

où $f_{x,z}(x,z)$ est la densité de probabilité du couple$(x,z)$.

$$E\left(\left(x-\hat{x}(z)\right)^2\right)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(x-\hat{x}(z)\right)^2 f_{x|z}(x\mid z)\,f_z(z)dxdz=\int_{-\infty}^{+\infty}E\left(\left(x-\hat{x}(z)\right)^2\mid z\right)f_z(z)dz$$

On cherche à minimiser $E\left(\left(x-\hat{x}(z)\right)^2\mid z\right)$ :

$$E\left(\left(x-\hat{x}(z)\right)^2\mid z\right)=\int_{-\infty}^{+\infty}\left(x-\hat{x}(z)\right)^2 f_{x|z}(x|z)\,dx$$

$$=\int_{-\infty}^{+\infty}x^2 f_{x|z}(x|z)\,dx-2\hat{x}(z)\int_{-\infty}^{+\infty}x f_{x|z}(x|z)\,dx+\hat{x}(z)^2\int_{-\infty}^{+\infty}f_{x|z}(x|z)\,dx$$

$$=\int_{-\infty}^{+\infty}x^2 f_{x|z}(x|z)\,dx-2\hat{x}(z)E(x\mid z)+\hat{x}(z)^2$$

Le $\hat{x}$ annulant la dérivée de l'expression ci-dessus est : $\hat{x}(z)=E(x(z_0)\mid z)$

## *B. Méthode de l'estimation du maximum de vraisemblance*

La méthode du maximum de vraisemblance fait la supposition que le processus aléatoire est gaussien, ce qui permet d'avoir une expression de la densité de probabilité des observations sachant le paramètre de portée. L'algorithme détermine alors le paramètre de portée ayant le plus vraisemblablement généré les observations.

Notons $Y=(Y_1,Y_2,\ldots,Y_n)$ le vecteur des observations et $m=(E(Y_1),...,E(Y_n))$ le vecteur des espérances associées aux variables aléatoires. On note $K(p)$ la matrice de covariance associée aux observations c'est-à-dire :

$$K(p)=\begin{pmatrix} k(Y_1,Y_1) & k(Y_1,Y_2) & \cdots & k(Y_1,Y_n) \\ k(Y_1,Y_2) & k(Y_2,Y_2) & \ldots & k(Y_2,Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(Y_1,Y_n) & k(Y_2,Y_n) & \ldots & k(Y_n,Y_n) \end{pmatrix}$$

où $k(Y_i,Y_j)$ est la fonction de covariance appliquée aux observations $(Y_i,Y_j)$, c'est-à-dire tel que $h=d(Y_i,Y_j)$, la distance entre $Y_i et Y_j$. La densité de probabilité des observations sachant le paramètre de portée est :

$$f(Y|p)=(2\pi)^{-\frac{n}{2}}\det(K(p))^{-1/2}e^{-\frac{1}{2}Y^tK(p)^{-1}Y}$$

C'est une fonction dépendant uniquement du paramètre de portée, les observations sont connues. Le principe de l'EMV est de trouver p tel que l'observation des données Y soit la plus vraisemblable. Il suffit donc de trouver p qui maximise la densité de probabilité $f(Y\mid p)$.

### *C. Le krigeage est un interpolateur exact*

Dans le cas du krigeage simple, l'estimateur est $\hat{Y}_x=m+\left(\mathrm{K}^{-1}\ \mathrm{k(x)}\right)^t(\mathrm{Y}-\mathrm{m})$. Or si le point « x » correspond à l'i-ème point de mesure, on en déduit que $k(x)$ est la colonne $i$ de la matrice K, matrice de covariance des observations. Donc $\mathrm{K}^{-1}\ \mathrm{k(x)}=e_i$ c'est-à-dire l'i-ème vecteur de la base canonique. On a donc bien comme estimateur l'observation $Y_i$.

### *D. Situations menant à un mauvais conditionnement*

**Le nombre de conditionnement augmente avec la densité des observations**

Considérons le conditionnement en norme 2. La matrice de covariance $K$ est symétrique, définie positive, donc inversible. Le nombre de conditionnement correspond dans ce cas à $\frac{\lambda_{\max}}{\lambda_{\min}}$ où $\lambda_{\max}$ est la plus grande valeur propre de $K$ et $\lambda_{\min}$ est la plus petite valeur propre. Le déterminant de cette matrice est liée à ses valeurs propres par :

$$\det(K)=\prod_i\lambda_i\ .$$

Une propriété du déterminant est qu'il correspond au volume du parallélotope défini par les vecteurs de la matrice K. Une face de ce parallélotope est la surface délimitée par deux vecteurs de K. Quand on rapproche ces deux vecteurs, la face devient de plus en plus petite donc le volume du parallélotope diminue, le déterminant de K tend alors vers 0.

Lorsque deux observations sont suffisamment proches, leur vecteur respectif intervenant dans K sont très proches également, donc la surface délimitée par ces vecteurs est petite. Quand les observations sont de plus en plus proches, plusieurs faces du parallélotope diminuent et tendent vers un segment. Le déterminant tend vers zéro donc au moins une valeur propre tend vers 0. Le dénominateur correspondant à $\lambda_{\min}$ est proche de 0, le nombre de conditionnement est alors grand.

**Le nombre de conditionnement augmente avec le coefficient de portée**

Plaçons-nous en dimension quelconque et prenons un modèle de covariance noté $c(h, p)$ de type exponentiel et de portée $p$ . Multiplions la portée par 10. La covariance associée à deux observations séparées d'une distance $h$ est $c(h,10\cdot p) = e^{-\frac{h}{10\cdot p}} = e^{-\frac{h/10}{p}} = c\left(\frac{h}{10}, p\right)$. Multiplier la portée par 10 revient à appliquer un modèle de covariance de paramètre $p$ sur des observations séparées d'une distance $\frac{h}{10}$, donc à considérer un domaine où l'écart entre les points est plus petit, donc de densité plus grande. Le nombre de conditionnement de la matrice associée à $c(h,10\cdot p)$ est plus grand que le nombre de conditionnement de la matrice associée à $c(h, p)$ .

## *E. Compléments au premier rapport*

Le travail précédent (commande IRSN EX10 / 32001814 du 30.05.2014), a mis en évidence une mauvaise performance du krigeage lorsque l'information est pauvre. Nous avions reconstruit la fonction de Branin à partir de 3 et 9 points d'observation successivement, par la méthode du krigeage et par l'EPH. La surface reconstituée par l'EPH était plus proche de la surface de référence que celle du krigeage. L'interpolation faite par krigage était insatisfaisante car elle était constante pour les points éloignés des observations.

La présente étude permet d'expliquer cette mauvaise performance : le coefficient de portée dans le krigeage n'a pas été estimé correctement en raison du faible nombre d'observations. Le paramètre peut être estimé de deux façons différentes dans le logiciel R :

- manuellement, en ayant étudié au préalable la variabilité des données ;
- automatiquement : le paramètre est estimé par l'algorithme du logiciel R. Cet algorithme utilise la méthode du maximum de vraisemblance ;

Dans notre première étude, le paramètre a été estimé automatiquement par l'algorithme, par la méthode du maximum de vraisemblance, et il a été mal estimé.

Reprenons l'exemple utilisé dans le premier rapport. On souhaite reconstruire la fonction de Branin à partir de 9 points de mesure réparties régulièrement sur le domaine [0,1]x[0,1]. La reconstruction obtenue avec le krigeage est affichée ci-dessous :

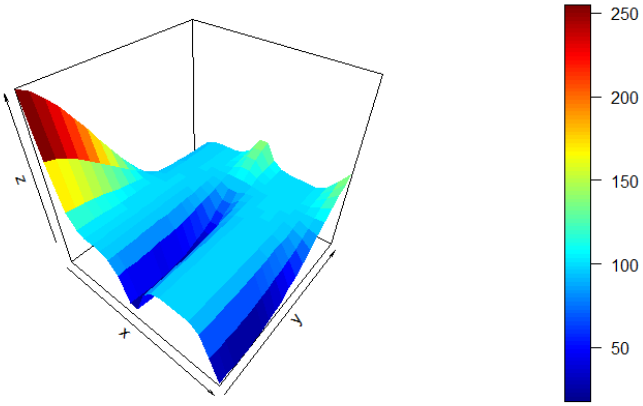


*Figure 61 : surface obtenue par krigeage dans le premier rapport*

Il y a un coefficient de portée pour chaque axe. Dans la direction $x$, le coefficient de portée estimé par l'algorithme est quasiment nul, d'où la tendance constante quand les points sont éloignés des observations. Ce coefficient n'a aucune raison d'être nul : les données proches les unes des autres ne sont pas indépendantes. Nous le fixons donc à 0.2. L'interpolation obtenue est bien meilleure.

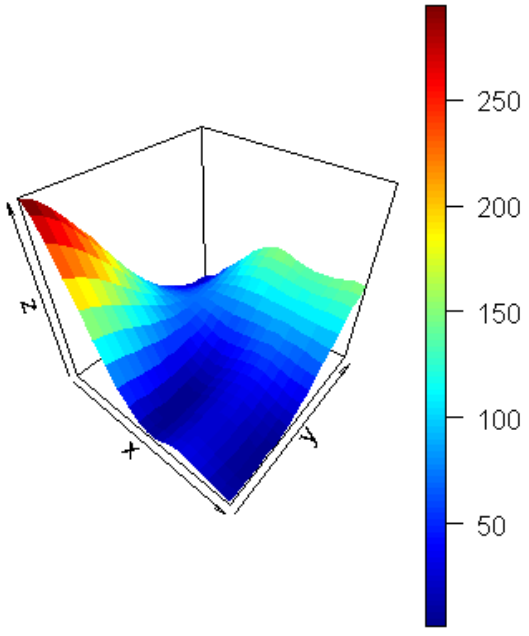


*Figure 62 : krigeage avec initialisation correcte de la portée dans la direction x*

Les performances des deux méthodes sont données dans le tableau suivant :

| | Dist_1 | EQ_2 |
|---|---|---|
| Krigeage | 45 | 36 129 |
| Krigeage avec correction | 27 | 1508 |
| EPH | 32 | 4880 |

Le krigeage avec correction de la portée donne une meilleure estimation que l'EPH.

### *F. Comportement de l'EPH en cas de répartition inégale des données*

Supposons qu'il y ait d'un côté du domaine un grand nombre d'observations, prenant chacune la valeur 10, et de l'autre côté seulement quelques observations prenant la valeur 7. Nous souhaitons reconstruire la valeur au milieu du domaine avec l'EPH.

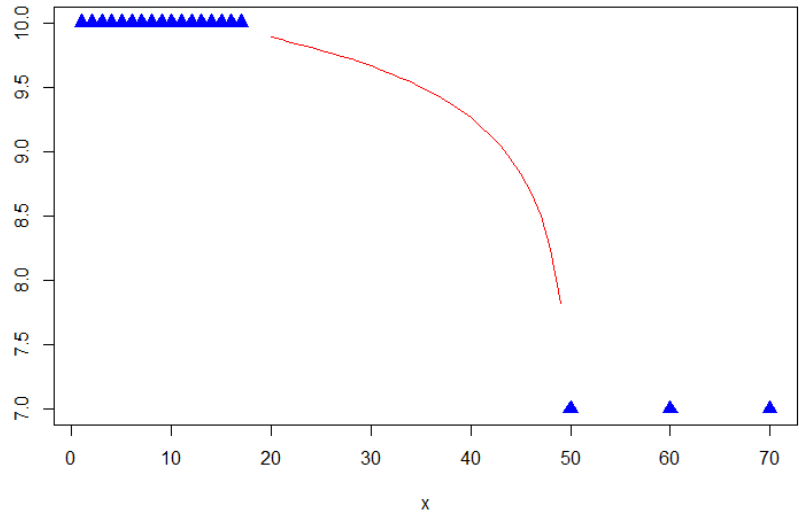


*Figure 63 : interpolation par EPH avec une répartition inégale des mesures*

Le comportement de l'EPH peut paraître surprenant, mais il n'est pas le fruit d'une erreur. L'EPH pondère chaque observation uniquement en fonction de sa distance avec le point à reconstruire. La localisation de la mesure n'intervient pas : si une observation se trouve à une distance de 20 du point à reconstruire, son poids est le même qu'elle se trouve à gauche ou à droite. Dans cet exemple, il y a beaucoup plus d'observations donnant la valeur 10 que de mesures donnant la valeur 7, par conséquent il est plus probable d'obtenir la valeur 10 que 7.

Ainsi lorsque l'on se trouve au milieu du domaine, il est normal d'obtenir une valeur plus proche de 10 que de 7.

Il existe cependant des versions non-isotropes de l'EPH. Le domaine est divisé en plusieurs parties, et la propagation de l'information est différente pour chacune. Ceci doit être justifié par des raisons physiques.

## *G. Archivage de l'implémentation en R*

### 1. Vérification de l'archive

Le programme fourni par le logiciel R permet de vérifier la validité d'un package. Le résultat de l'exécution est donné ci-dessous :

```
* using log directory 'C:/Users/***************/Code/EP
H_package_CRAN/mypkg.Rcheck'
* using R version 3.1.2 (2014-10-31)
* using platform: x86_64-w64-mingw32 (64-bit)
* using session charset: ISO8859-1
* checking for file 'mypkg/DESCRIPTION' ... OK
* this is package 'eph' version '0.1'
* checking CRAN incoming feasibility ... NOTE
Maintainer: 'Gottfried Berton <Gottfried.Berton@scmsa.eu>'
New submission
* checking package namespace information ... OK
* checking package dependencies ... OK
* checking if this is a source package ... OK
* checking if there is a namespace ... OK
* checking for .dll and .exe files ... OK
* checking for hidden files and directories ... OK
* checking for portable file names ... OK
* checking whether package 'eph' can be installed ... OK
* checking package directory ... OK
* checking DESCRIPTION meta-information ... OK
* checking top-level files ... OK
* checking for left-over files ... OK
* checking index information ... OK
* checking package subdirectories ... OK
* checking R files for non-ASCII characters ... OK
* checking R files for syntax errors ... OK
* checking whether the package can be loaded ... OK
* checking whether the package can be loaded with stated dependencies ...
OK
* checking whether the package can be unloaded cleanly ... OK
* checking whether the namespace can be loaded with stated dependencies
... OK
* checking whether the namespace can be unloaded cleanly ... OK
```

```
* checking loading without being on the library search path ... OK
* checking dependencies in R code ... OK
* checking S3 generic/method consistency ... OK
* checking replacement functions ... OK
* checking foreign function calls ... OK
* checking R code for possible problems ... OK
* checking Rd files ... OK
* checking Rd metadata ... OK
* checking Rd line widths ... OK
* checking Rd cross-references ... OK
* checking for missing documentation entries ... OK
* checking for code/documentation mismatches ... OK
* checking Rd \usage sections ... OK
* checking Rd contents ... OK
* checking for unstated dependencies in examples ... OK
* checking examples ... OK
* DONE

NOTE: There was 1 note.
See
  'C:/Users/*******/EPH_package_CRAN/mypkg
.Rcheck/00check.log'
for details.
```

### 2. Description

```
Package: eph
Version: 0.1
Date: 2015-01-01
Title: Experimental Probabilistic Hypersurface
Author: Gottfried Berton <Gottfried.Berton@scmsa.eu>
Maintainer: Gottfried Berton <Gottfried.Berton@scmsa.eu>
Depends: R (>= 3.1.0)
Imports: RODBC, tcltk, XLConnect, XLConnectJars, DiceKriging
Suggests: MASS
Description: This package aims at reconstruction information. The EPH
methods propagate the available information toward unknown points
License: GPL (>= 2)
Packaged: 2012-10-29 13:13:01 UTC; ripley
Repository: CRAN
```

### 3. Predict.Rd

```
% File man/predict_eph.Rd
\name{predict_eph.eph}
\alias{predict_eph.eph}
\title{run EPH}
\description{
```

```
this function estimates the expected value, the mean and mediane value of
the
probability law returned by the EPH in each point of the domain.
At the observation points, the algorithm returns the real value.
}
\arguments{
     \item{object}{object of class EPH}
     \item{newdata}{coordinantes on which performing the estimation}
}
\values{
     A list of vectors, the components are :
     \item{mean}{quantile 50 of the probability law}
     \item{upper95}{quantile 95 of the probability law}
     \item{probable}{most probable value}
     \item{esperance}{expected value of the probabilty law}
     \item{lower95}{quantile 5 of the probability law}
     \item{proba_density}{raw probabilty distribution computed by the
EPH}
}
\references{Olga Zeydina et Bernard Beauzamy : Probabilistic Information
Transfer.
Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA.
ISBN: 978-2-9521458-6-2, ISSN : 1767-1175. Relié, 208 pages, mai 2013.
}
\examples{

## a 2D example : reconstruct the branin function
library(DiceKriging)

d <- 2; n <- 9
design.fact <- expand.grid(x1=seq(0,1,length=3), x2=seq(0,1,length=3))
y <- apply(design.fact, 1, branin)

# bound output variable (e.g temperature)
tmax<- (500)
tmin<- (0)
response.fact<-cbind(y)

# bound of the parameter
mat=matrix(data=c(tmin,0,0,tmax,1,1), ncol=2)
boundariesP=rbind(c(tmin,tmax))

pas=cbind(c(0.1,0.1,0.1))
eph1 <- eph(design.fact, response.fact, mat, boundariesP, pas)
n.grid <- 10
x.grid <- y.grid <- seq(0,1,length=n.grid)
design.grid <- expand.grid(x1=x.grid, x2=y.grid)
response.grid <- apply(design.grid, 1, branin)
nb_to_estimate<-nrow(design.grid)
```

```
m<-1
eph1 <- eph(design.fact, response.fact, mat, boundariesP, pas)
system.time(carac <- predict_eph.eph(eph1, design.grid))
loi <- carac$proba_density
esp <- carac$esperance

}
```

### 4. Eph.Rd

```
% File man/eph.Rd
\name{eph}
\alias{eph}
\title{builds eph object}
\description{
builds a EPH object
}
\arguments{
      \item{Design}{value of the measure at each observation point}
      \item{Response}{value of the measure at each observation point}
      \item{Boundaries}{value of the measure at each observation point}
      \item{BoundariesP}{value of the measure at each observation point}
      \item{Step}{value of the measure at each observation point}
}
\values{
      \item{x}{coordinantes of each observation point}
      \item{y}{value of the measure at each observation point}
      \item{NofParam}{number of input parameter}
      \item{NofMeas}{number of observation points}
      \item{NofInk2}{number of output variables}
      \item{Boundariesin}{value of min and max for each input parameter}
      \item{Boundariesout}{value of min and max for each output
variables}
      \item{Step}{discretisation step for each input and output
parameter}
}
```