Mise en oeuvre du krigeage sur arbre.

Edwige Polus, Chantal de Fouquet

▶ **To cite this version:**

Edwige Polus, Chantal de Fouquet. Mise en oeuvre du krigeage sur arbre.. 42èmes Journées de Statistique, 2010, Marseille, France, France. inria-00494804

HAL Id: inria-00494804
https://inria.hal.science/inria-00494804

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MISE EN ŒUVRE DU KRIGEAGE SUR ARBRE

Edwige Polus & Chantal de Fouquet

Mines-ParisTech/Centre de Géosciences/Géostatistique 35, rue Saint Honoré – 77305 Fontainebleau Cedex – France.

Résumé

Un modèle de fonctions aléatoires définies sur une topologie d'arbre (de Fouquet & Bernard-Michel, 2006) a été développé pour l'estimation de concentrations le long d'un réseau hydrographique. Le principe consiste à décomposer le réseau en filets élémentaires joignant chaque « source » à « l'exutoire ». La concentration Z au point x s'exprime comme une combinaison linéaire $Z(x)=\Sigma w_i Y_i(x)$ de variables aléatoires élémentaires Y_i définies sur ces filets, dont les coefficients w_i dépendent de la position sur l'arbre. Le krigeage des concentrations au point x revient donc à l'estimation d'une combinaison linéaire des $Y_i(x)$ à partir d'autres combinaisons linéaires, les concentrations mesurées aux points expérimentaux x_α : $Z(x_\alpha)= \Sigma w_i(x_\alpha)Y_i(x_\alpha)$.
Le krigeage dépend des hypothèses sur les concentrations, qui se ramènent à des hypothèses sur les variables élémentaires Y_i et sur les coefficients w_i. Ces hypothèses déterminent les conditions requises pour l'estimation : le nombre minimum de mesures et leur répartition par arête.
Ce modèle est appliqué aux concentrations en nitrates sur un réseau constitué d'une petite portion de la Seine (de l'amont de Paris à l'estuaire) et de la Marne. Tout d'abord, la simulation déterministe ProSe, disponible en tout point (Even et al., 1998), permet de tester les hypothèses. Ensuite, le krigeage est effectué à partir des mesures aux stations effectivement disponibles.

Abstract

A model of random functions defined on a tree (de Fouquet & Bernard-Michel, 2006) has been developed to estimate concentrations along a hydrographic network. The basic principle consists in considering elementary "streams" joining each "source" to the "outlet". The concentration Z at a point x is a linear combination $Z(x)=\Sigma w_i Y_i(x)$ of elementary random variables Y_i defined on these streams, and the coefficients w_i depends on the location on the tree. Kriging concentrations at the point x comes to the estimation of a linear combination of $Y_i(x)$ from other linear combinations: concentrations measured at the experimental points x_α : $Z(x_\alpha)= \Sigma w_i(x_\alpha)Y_i(x_\alpha)$.
Kriging depends on hypotheses on concentrations, which are hypotheses on elementary variables Y_i and on coefficients w_i. These hypotheses determine the required conditions for the estimation: minimum number of measurements and their distribution by edge.
This model is applied to nitrates concentrations on a network composed of a part of the Seine River (from upstream of Paris to the estuary) and of the Marne River. First the physically-based simulation ProSe (Even et al., 1998), available at any point, allows to test the hypotheses. Kriging is then made from measurements at sites which are actually available.

Mots-clés

Géostatistique, estimation, arbre, réseau hydrographique, concentrations, krigeage

Key-words

Geostatistics, estimation, tree, hydrographic network, concentrations, kriging

Introduction

L'estimation géostatistique des concentrations le long d'un réseau hydrographique amène des difficultés théoriques en raison du changement de topologie, mais également des problèmes pratiques à cause du nombre restreint de stations de mesures disponibles le long du réseau. Un

modèle de fonctions aléatoires adapté au support arborescent a ainsi été proposé par Ver Hoef (2006), généralisé par de Fouquet et Bernard-Michel (2006), et développé ici pour l'estimation de concentrations le long d'un réseau hydrographique. Ce modèle général est brièvement présenté, et différentes variantes sont examinées. Afin de pallier le manque de données, une simulation d'un modèle déterministe réaliste est d'abord exploitée pour tester les hypothèses d'une part et réaliser l'inférence de l'autre. Enfin, le krigeage est effectué, d'abord en validation croisée sur la simulation déterministe, puis à partir des données réelles.

Principe du modèle

Le principe consiste à décomposer le réseau en chemins ou filets élémentaires joignant chaque « source » à « l'exutoire » (Figure 1). La concentration Z au point x s'exprime comme une combinaison linéaire $Z(x)=\Sigma w_i Y_i(x)$ de variables aléatoires élémentaires Y_i définies sur ces filets, dont les coefficients w_i dépendent de la position sur l'arbre.

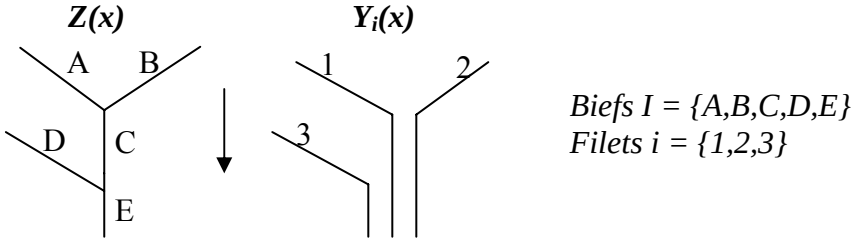

Figure 1 : Décomposition d'un réseau hydrographique à cinq bief en trois filets élémentaires.

L'estimation en un point x_0 de la concentration $Z(x_0)$ par une combinaison linéaire des mesures $Z(x_\alpha)$ revient donc à estimer une combinaison linéaire des $Y_i(x_0)$ par une autre combinaison linéaire des $Y_i(x_\alpha)$. Le système de krigeage dépend des hypothèses retenues pour les concentrations, qui se ramènent à des hypothèses concernant d'une part les composantes Y_i définies sur les filets et d'autre part les coefficients w_i.

Hypothèses et inférence

La première étape consiste à préciser les hypothèses :

- Indépendance des composantes Y_i et donc des concentrations Z(x) à l'amont d'une confluence ?
- Degré de stationnarité des composantes Y_i : fonctions aléatoires intrinsèques ou stationnaires d'ordre deux ?
- Espérance : modèle du type $Z = m + \Sigma w_i Y_i$ avec m inconnue et $E(Y_i) = 0$, ou $Z = \Sigma w_i Y_i$ avec $E(Y_i) = m_i$ inconnue ?
- les coefficients w_i sont-ils supposés connus ? Par exemple la conservation de la masse en chaque confluence fixe ces coefficients égaux aux débits relatifs des affluents, qui peuvent ou non être supposés connus. Si ces coefficients ne sont pas connus, ils peuvent être contraints, leur somme étant égale à l'unité en chaque confluence. Enfin, les coefficients peuvent être totalement libres, du fait des réactions entre les constituants des différents affluents en aval d'une confluence.

Suivant le modèle retenu, se pose alors la question de l'inférence. Le variogramme des composantes se déduit directement du variogramme des biefs. Le cas le plus simple est celui de composantes indépendantes et où les coefficients w_i, égaux aux débits relatifs des affluents, sont supposés connus. En admettant que tous les variogrammes des composantes sont proportionnels, la modélisation se réduit à la détermination des « paliers » ou d'un facteur par composante.

Ainsi, pour un réseau où tous les biefs sources sont informés, le système est entièrement déterminé pour des coefficients contraints, et donc a fortiori pour des coefficients connus où les hypothèses peuvent être vérifiées. Par contre si les coefficients sont libres, les différents paramètres devront être optimisés, via la méthode des moindres carrés par exemple.

Conditions de non biais

Les conditions de non biais ou d'autorisation dépendent du degré de stationnarité des composantes Y_i ainsi que des hypothèses sur les moyennes.
Dans un cadre intrinsèque, les filets étant supposés indépendants, les conditions d'autorisation, s'écrivent $\Sigma\lambda_\alpha w_{i\alpha} = w_{i0}$ pour tout filet i. L'estimation n'est donc possible que sur des biefs dont tous les filets sont échantillonnés, et requiert un nombre de données strictement supérieur au nombre de filets. Le système de krigeage s'en déduit par minimisation sous contrainte de la variance d'estimation.

Application aux concentrations en nitrates

Ce modèle a été testé sur les concentrations annuelles en nitrates le long d'un réseau restreint constitué d'une portion de la Seine (de l'amont de Paris à l'estuaire) et de la Marne, informé en tout point par le modèle déterministe ProSe (Even et al., 1998) pour l'année 2003. La figure 2 présente ce réseau ainsi que sa décomposition en filets.

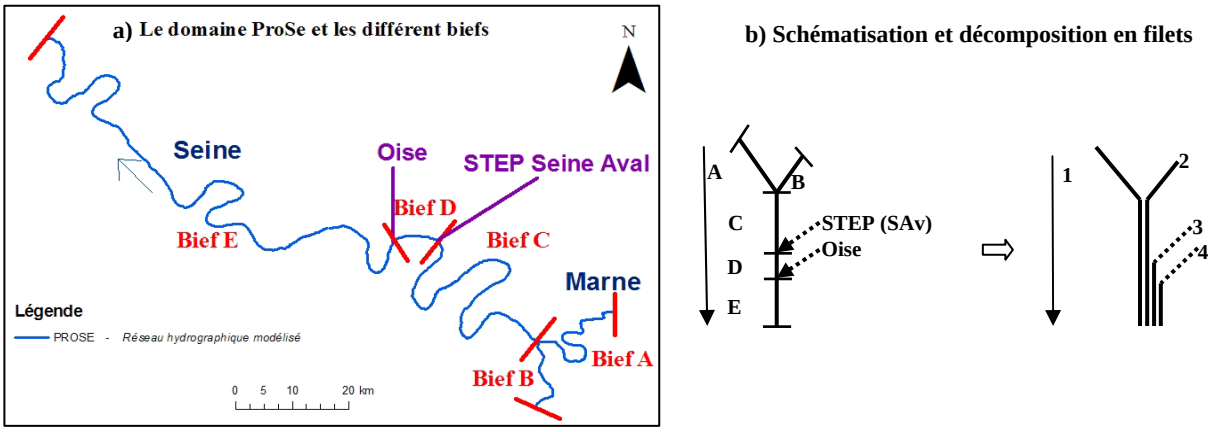

Figure 2 : a) Le réseau et les biefs le constituant. b) Schématisation du réseau et décomposition en filets.

L'hypothèse d'indépendance des filets, testée à l'amont de la confluence Seine/Marne sur les biefs A et B pour les filets 1 et 2, paraît tout à fait plausible. La conservation de la masse à la confluence a été vérifiée et semble valide sur une trentaine de kilomètres : les débits relatifs paraissent donc bien indiqués pour pondérer les filets. L'inférence des variogrammes des composantes conduit à un système de cinq équations à quatre inconnues.

En pratique, l'inférence n'est pas toujours aussi aisée, en raison du faible nombre de mesures disponibles par bief. Pour le réseau considéré, le Réseau National de Bassin (RNB) possède 25 stations réparties environ tous les 10 km, ce qui amène des difficultés pratiques. Par exemple, l'ajustement devient impossible avec seulement une ou deux stations par bief.
Lorsque les variogrammes expérimentaux ne peuvent être ajustés directement, des solutions sont envisageables, par exemple en adaptant la méthode d'identification automatique des covariances généralisées dans un cadre non stationnaire présentée par Chilès & Delfiner (1999). Cela évite le regroupement de biefs, au prix d'une autre approximation consistant à approcher une variance de combinaison linéaire par un écart quadratique.

Résultats

La figure 4 présente les résultats de la validation croisée effectuée sur les valeurs ProSe (une valeur sur trois est conservée pour l'inférence et les deux autres sont ensuite estimées) ainsi qu'une estimation de la concentration annuelle en nitrates effectuée à partir des mesures du RNB. Les débits relatifs ont été retenus comme coefficients, et pour le second exemple, le bief D a été regroupé avec le bief E. Dans les deux cas, le résultat est très convaincant, particulièrement à proximité des confluences.

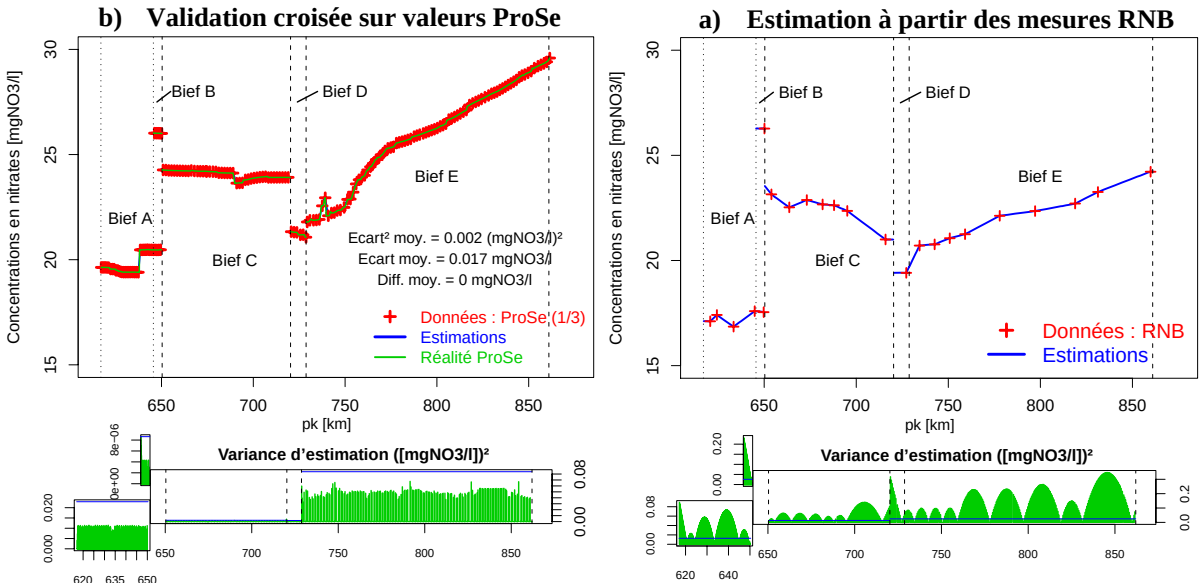

Figure 4 : a) Validation croisée du modèle à partir d'un tiers des valeurs ProSe. b) Estimation à partir des seules mesures du RNB.

Perspectives

L'extension à d'autres variables et à un réseau plus étendu constituent les principales perspectives de ce travail, afin de valider le modèle et les méthodes d'inférence proposées. Le recalage du modèle aux mesures par un krigeage avec dérive externe est en cours.

Bibliographie

[1] Ver Hoef, J. M., Peterson, E. et Theobald, D. (2006). Spatial Statistical Models that Use Flow and Stream Distance. *Environmental and Ecological statistics* (in press).

[2] de Fouquet, C and Bernard-Michel, C. (2006). Modèles géostatistiques de concentrations ou de débits le long des cours d'eau. *Comptes rendus Géoscience*, **338**, 5, 307-318.

[3] Bernard-Michel, C. (2006) Indicateurs géostatistiques de la pollution dans les cours d'eau. Thèse de doctorat, Ecole des Mines de Paris.

[4] Even, S., Poulin, M., Garnier, J. [et al.] (1998) River ecosystem modelling. Application of the PROSE model to the Seine river (France). *Hydrobiologia;* 373/374:27–45.

[6] Chilès, J-P. et Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York.