Enlever les valeurs extrêmes de concentration: pour quoi faire ?

Jacques.Rivoirard@ensmp.fr
CG Fontainebleau

Ce court texte formalise certaines approches basiques au problème posé par l'élimination des valeurs extrêmes. Il développe, avec plus de détails, la présentation faite aux Journées de Géostatistique des 18-19 septembre 2003.

1. Introduction

Les valeurs extrêmes de concentration sont très répandues, qu'il s'agisse de ressources naturelles (métaux précieux, poisson, …) ou de polluant. Elles posent un sérieux problème, vu leur importance en abondance globale ou en dépassement de seuil. Elles ont en particulier l'inconvénient de rendre instables des statistiques, comme justement la moyenne (donnant l'abondance) ou le variogramme.

Les besoins de modèles aptes à tenir compte des valeurs extrêmes sont divers. Il existe une abondance littérature statistique sur les queues de distribution, utiles pour la prédiction d'évènements rares (crue centenaire, par exemple). Des modèles similaires peuvent trouver aussi leur utilité en géostatistique (cas des gisements de diamants, par exemple). Dans beaucoup de cas cependant, on n'a pas réellement besoin de modéliser la queue de distribution. Le mineur d'or, par exemple, a l'habitude de rabattre à un seuil donné (par exemple 50g/t) les valeurs d'échantillonnage qui la dépassent (technique de l'écrêtage), du moins pour l'estimation locale. La partie enlevée dépend de la moyenne des valeurs dépassant le seuil, et le mineur est beaucoup plus intéressé par cette valeur, que par les détails de la distribution aux très fortes valeurs.

Une technique très répandue, dans les premiers stades d'une étude géostatistique, est la suppression des valeurs extrêmes. Elle peut servir à stabiliser la structure variographique, laquelle pourra être utilisée pour un krigeage, excluant ou non les valeurs extrêmes. La question abordée ici est celle de savoir ce que l'on peut dire d'une telle technique de façon théorique. En particulier, est-il possible de construire des modèles géostatistiques qui la légitiment ? par exemple dans lesquels le variogrammes entier est identique (proportionnel) au variogramme sans les extrêmes, ou s'en déduit simplement (exemple : addition d'une pépite) ? ou dans lesquels l'estimation peut se baser sur un krigeage des seuls valeurs au-dessous du seuil ?

Nous adopterons un certain nombre d'hypothèses de base :

1. Seront considérées comme valeurs extrêmes, les valeurs au-dessus d'un seuil z donné, supposées peu fréquentes et beaucoup plus fortes que les autres.

2. Ces valeurs extrêmes ne sont pas des valeurs erronées.

3. Nous supposerons a priori que ces valeurs extrêmes peuvent se trouver n'importe où dans le champ étudié. Il est clair que si nous savons qu'elles sont absentes sur une partie connue du champ, celle-ci peut être écartée. Cependant il se peut que les valeurs extrêmes apparaissent

avec une probabilité (du moins une fréquence) non stationnaire à travers le champ, et soient donc plus rares dans certaines parties.

4. On ignorera les incertitudes sur le variogramme calculé sans les extrêmes: celui-ci est supposé parfaitement connu

2. Le variogramme conditionnel

Dans le modèle de Fonction Aléatoire, le variogramme sans les extrêmes est un variogramme conditionnel qui s'écrit :

$$\gamma_{-z}(h) = \frac{1}{2} E\left[\left(Z(x+h) - Z(x) \right)^2 \mid Z(x) < z, Z(x+h) < z \right]$$

Il ne s'agit d'ailleurs pas nécessairement d'une fonction variogramme (conditionnellement définie négative). Imaginons un processus de type mosaïque, où chaque compartiment à valeur faible ou moyenne est entourée de valeurs extrêmes, par exemple à 1D une alternance de plages de valeurs extrêmes de largeur L et de plages de valeurs modérées de plus grande largeur : le variogramme sans les extrêmes est identiquement nul sur [0, L], mais n'est pas nul sur toutes les distances. Un tel variogramme ne peut être le variogramme d'une FA intrinsèque : les variations, nulles jusqu'à L, seraient nulles de proche en proche pour toute distance, d'où une FA constante, de variogramme identiquement nulle.

Il est possible de faire le lien entre le variogramme sans les extrêmes et le variogramme entier :

$$\gamma(h) = \frac{1}{2} E\left[\left(Z(x+h) - Z(x) \right)^2 \right]$$

si l'on fait choix d'une hypothèse de lois bivariables [Z(x), Z(x+h)] : on rentre alors dans le domaine, assez exigeant en hypothèses, de la géostatistique non-linéaire. Les calculs de semblent pas déboucher en général sur des formules simples. Le prototype que constitue le modèle mosaïque à valuations indépendantes (partition de l'espace où chaque compartiment est valué indépendamment et selon la même loi) est assez intéressant. Dans ce modèle en effet, le variogramme entier de la variable brute, ou d'ailleurs de n'importe quelle transformée, est proportionnel à la probabilité $\gamma(h)$ pour que deux points (x, x+h) appartiennent à des compartiments différents. Pourtant le variogramme sans les extrêmes est différent (et ce n'est peut-être pas une fonction variogramme). Il s'écrit :

$$\gamma_{-z}(h) = \frac{\mathrm{var}(Z \mid Z < z)\ P(Z < z)\ \gamma(h)}{1 - \left[1 - P(Z < z)\right]\gamma(h)}$$

Il est un peu plus continu que $\gamma(h)$, ce qui est dû au fait qu'en retranchant une classe de valeurs, on supprime, dans le calcul du variogramme, relativement plus de paires de points à cheval sur deux compartiments aux petites distances. Le dénominateur varie linéairement en fonction de $\gamma(h)$, valant 1 pour $\gamma(h) = 0$ et P(Z<z) pour $\gamma(h) = 1$. La formule précédente peut d'ailleurs s'inverser, ce qui permet de corriger le biais (d'ailleurs faible, et d'autant plus que la probabilité des valeurs extrêmes l'est) :

$$\gamma(h) = \frac{\gamma_{-z}(h)}{\text{var}(Z \mid Z < z)\ P(Z < z) + \left[1 - P(Z < z)\right]\gamma_{-z}(h)}$$

3. Approche additive

3.1. Le modèle

L'élimination des valeurs fortes suggère le modèle additif : $Z(x) = Z_1(x) + Z_2(x)$ somme de deux composantes positives ou nulles :

- un fond Z_1 inférieur (ou égal) au seuil z ;
- une composante Z_2 responsable du dépassement de seuil.

Ainsi pour $Z(x) \leq z$, on a $Z(x) = Z_1(x)$, avec $Z_2(x) = 0$. De façon à accéder aux statistiques, éventuellement spatiales, de $Z_1(x)$ à partir de l'ensemble S_1 des seuls points connus où $Z(x) \leq z$, on fera l'hypothèse que les FA Z_1 et Z_2 sont indépendantes. On a :

$$Z_2(x) > 0 \Leftrightarrow Z(x) = Z_1(x)+Z_2(x) > z \Leftrightarrow Z_2(x) > z - Z_1(x)$$

Et comme $Z_1(x)$ et $Z_2(x)$ sont indépendants:

$$Z_2(x) > 0 \Rightarrow Z_2(x) > z\text{-}\min(Z_1) \geq \max(Z_1) - \min(Z_1)$$

(en supposant que la borne inférieure de la distribution de Z_1, $\min(Z_1)$, est atteinte: par exemple 0).

Ainsi, les valeurs non nulles de $Z_2(x)$ excèdent l'intervalle de variation de Z_1, et à partir de n'importe quelle valeur de $Z_1(x)$, font passer Z(x) au-dessus du seuil z (ceci pourrait sans doute se rencontrer pour des phénomènes ou dans des circonstances très particulières). Pour résumer :

$$0 \leq \min(Z_1) \leq Z_1(x) \leq \max(Z_1) \leq z < \min(Z_1) + Z_2(x)\big|_{Z_2(x)>0}$$

On notera le rôle très particulier du seuil z dans la distribution de Z (bimodale, par exemple). Par ailleurs toute valeur de seuil entre $\max(Z_1)$ et $\min(Z_1) + \min(Z_2(x)|\ Z_2(x)>0)$ est équivalente.

3.2. La structure

Un avantage essentiel de ce modèle additif est la simplicité de sa structure :

$$\gamma_Z(h) = \gamma_1(h) + \gamma_2(h)$$

$$\gamma_{Z,Z_1}(h) = \gamma_1(h)$$

$$\gamma_{Z,Z_2}(h) = \gamma_2(h)$$

Ainsi, $\gamma_1(h)$ étant supposé connu, on peut faire l'hypothèse d'un $\gamma_Z(h)$ s'en déduisant simplement :

- soit par addition d'une composante supplémentaire $\gamma_2(h)$, pépitique par exemple ;
- soit par rescaling de $\gamma_1(h)$ (ce qui suppose $\gamma_2(h)$ identique à $\gamma_1(h)$).

3.3. L'estimation

L'estimation, ou le calcul de variances, peut s'envisager de diverses façons :

1. L'usage direct de $\gamma_Z(h)$ permet calcul de variance et krigeage de la variable brute Z(x).

2. Cependant le modèle bivariable précédent permet calcul de variances et cokrigeage à partir des données :

$Z_1(x)$ et $Z_2(x)=0$ connus sur S_1

$Z(x) = Z_1(x) + Z_2(x)$ connu sur les autres points S_2

(on notera que, du fait de ces dernières données, bien que $Z_1(x)$ et $Z_2(x)$ soient indépendantes, leur cokrigeage n'est pas leur krigeage).

3. Enfin, une alternative au cokrigeage (non optimale, mais privilégiant la structure de Z_1, la mieux connue) consiste à :

- kriger $Z_1(x)$, y compris sur S_2, à partir des seules valeurs non extrêmes $Z(x) = Z_1(x)$ sur S_1
- en déduire valeurs estimées de $Z_2(x) = Z(x) - Z_1(x)$ sur S_2
- estimer $Z_2(x)$ à partir de ces valeurs estimées de $Z_2(x)$ sur S_2 et des valeurs nulles de $Z_2(x)$ sur S_1 (saupoudrage uniforme de la moyenne m_2 de $Z_2(x)$ si celle-ci est pépitique).

3.4. Conclusion

L'intérêt de ce modèle est la simplicité de sa structure et de l'estimation. Ses inconvénients sont:

- une hypothèse très particulière sur la distribution de valeurs de Z ;
- Z_1 et Z_2 n'étant connus individuellement qu'aux points de données où $Z<z$, l'estimation ne tient pas explicitement compte du fait que Z_2 est > 0 aux points de données où $Z > z$ (pour tenir compte de telles inégalités, il faudrait recourir aux techniques, plus gourmandes en hypothèses, de simulation ; celles-ci auraient aussi l'avantage de maintenir Z_1 et Z_2 à l'intérieur de leur intervalle de variation).

4. Approche géométrique, ou ensembliste

4.1. Le modèle

On distingue ici les points de l'espace selon leur appartenance à l'un des deux ensembles : $A^c = A_z^c = \{x \mid Z(x) < z\}$ ou $A = A_z = \{x \mid Z(x) \geq z\}$, par exemple :

$$\begin{aligned} Z(x) &= Z(x)1_{Z(x)<z} + Z(x)1_{Z(x)\geq z} \\ &= Z(x)1_{x\in A^c} + Z(x)1_{x\in A} \end{aligned}$$

Les éléments à considérer pour une modélisation sont les structures et relations structurales de l'ensemble A des fortes valeurs, de Z(x) dans A, de Z(x) dans A^c, et de Z(x) au passage de A^c à A.

On obtient un modèle très simple en écrivant :

$$Z(x) = Y_1(x)1_{x\in A^c} + Y_2(x)1_{x\in A}$$

avec $1_{x\in A}$, $Y_1(x)$ (variant entre 0 et z), et $Y_2(x)$ ($\geq$ z) supposés indépendants. Ceux-ci peuvent faire l'objet d'estimations séparées (mais hétérotopiques, car là où on connaît $Y_1(x)$ par exemple, on ignore $Y_2(x)$). Ceci légitime donc au passage une estimation séparée des valeurs de Z(x) inférieures au seuil.

Dans ce modèle la structure de Z(x) est une combinaison linéaire des structures des 3 variables et de produits de ces structures. Un cas intéressant est celui où $1_{x\in A}$ et $Y_2(x)$ sont supposés (quasi) pépitiques. Le variogramme de la variable brute s'écrit alors :

$$\begin{aligned} \gamma_Z(h) &= \gamma_{-z}(h)\left[P(Z<z)\right]^2 + \textit{pépite} \\ \text{avec } \textit{pépite} &= \operatorname{var} Z - \operatorname{var}(Z \mid Z<z)\left[P(Z<z)\right]^2 \end{aligned}$$

(soit, à peu de choses près si la probabilité des valeurs fortes est faible, le variogramme sans les extrêmes avec un pépite additionnel égal à la chute de variance). L'estimation de Y_1(x), soit Z(x) | Z(x)<z, se complète par saupoudrage de $m_2 = E[Y_2(x)] = E[Z(x) \mid Z(x) \geq z]$ en proportion $P(Z \geq z)$.

Il est possible de réduire légèrement l'hypothèse d'indépendance en se plaçant dans un modèle factorisé par :

$1_{Z(x)\geq z}$ (éventuellement centrée), $[Z(x) - m_1]1_{Z(x)<z}$, et $[Z(x) - m_2]1_{Z(x)\geq z}$, avec :

$m_1 = E[Z(x) \mid Z(x) < z]$ et $m_2 = E[Z(x) \mid Z(x) \geq z]$. Les hypothèses correspondent à la non corrélation spatiale entre ces facteurs, ce qui donne :

- $E[Z(x) \mid Z(x) < z, Z(x+h) \geq z] = m_1 = E[Z(x) \mid Z(x) < z]$, soit absence d'effets de bord de Z(x) dans l'ensemble des Z(x) < z.

- $E[Z(x) \mid Z(x) \geq z, Z(x+h) < z] = m_2 = E[Z(x) \mid Z(x) \geq z]$, soit absence d'effets de bord de Z(x) dans l'ensemble des Z(x) ≥ z.
- Et enfin $E[Z(x)Z(x+h) \mid Z(x) < z, Z(x+h) \geq z] = m_1 m_2$, soit absence d'effets de bord couplés entre les deux ensembles.

Dans ce modèle, l'estimation par cokrigeage s'obtient par krigeage séparé (et isotopique) des facteurs.

5. Extension à plusieurs seuils

On a vu que les modèles obtenus faisaient jouer un rôle très particulier au seuil, correspondant à un certain type d'indépendance entre ce qui se passe au-dessous et ce qui se passe au-dessus. Les approches peuvent-elles rester cohérentes, lorsqu'on considère plus d'un seul seuil ?

5.1. Le modèle additif

Sauf à considérer le cas trivial de seuils équivalents, l'approche multi-seuil revient à enchaîner les opérations. De même que l'on dépasse le seuil z par ajout de $Z_2(x)$ (indépendant) à $Z_1(x)$, on passe le seuil z' ≥ z par ajout de $Z_3(x)$ (indépendant) à $[Z_1(x) + Z_2(x)]$. A partir de $Z_1(x) + Z_2(x)$ bimodal par exemple, l'ajout de la troisième composante pourra donner une distribution de $Z(x) = Z_1(x) + Z_2(x) + Z_3(x)$ à quatre modes… Les seuils jouent donc un rôle extrêmement particulier.

5.2. L'approche géométrique

Commençons par le second modèle, postulant la seule absence d'effets de bord simples et croisés entre les ensembles A et A^c. Il se généralise à plusieurs coupures par le modèle mosaïque à valuations indépendantes. La factorisation proposée pour un seuil z est parfaitement compatible avec une factorisation similaire pour le seuil z'. Cependant une telle factorisation perd son intérêt, dans la mesure où, dans ce modèle, l'estimation linéaire (isotopique) de Z(x) (ou d'une transformée de Z(x)) à l'aide de valeurs de celle-ci et d'autres transformées se réduit à son krigeage. On notera que dans ce modèle, le comportement de Z(x) dans A n'est pas indépendant de sa géométrie, par exemple les frontières de A sont aussi des bordures des compartiments valués.

Venons en maintenant au premier modèle, dans lequel les valeurs de Z(x) en dessous et au dessus du seuil sont des FA indépendantes. Celui-ci ne se généralise pas à des seuils différents (et non équivalents). Soit en effet z' > z. Du fait du comportement indépendant de Z(x) dans $A = A_z = \{x \mid Z(x) \geq z\}$, les valeurs ≥ z' se rencontrent n'importe où dans A, mais inversement, quittant les valeurs ≥ z', on rencontrera préférentiellement à petite distance des valeurs ≥ z (existence d'effets de bord descendants) : le comportement de Z(x) au-dessous de z' n'est donc pas indépendant de A_z^c. De la même façon, supposer un comportement indépendant de Z(x) dans $A_{z'}^{\ c} = \{x \mid Z(x) < z'\}$ donne des effets de bord en montant au-dessus de z.

On sait construire des modèles sans effet de bord, montant ou bien descendant, en superposant des ensembles aléatoires indépendants représentant des barrières de concentration, des plus hautes au plus basses pour l'absence d'effets de bord montant, et des plus basses aux plus

hautes pour l'absence d'effets de bord descendant. Ce dernier cas justifie une estimation séparée des valeurs en dessous d'un seuil quelconque. Mais il est plus pratique de recourir à la factorisation par les résidus d'indicatrices de tels modèles.

6. Conclusion et perspectives

Finalement, le variogramme calculé sans les valeurs extrêmes peut être rescalé (à un faible biais près) pour donner la structure unique d'un modèle mosaïque à valuations indépendantes, ce qui permet alors d'estimer par krigeage la variable brute ou toute transformée.

Il peut être également rescalé dans un modèle additif très particulier, où la seconde composante aurait même structure que la première composante.

Plus généralement dans le modèle additif, on peut rajouter la structure (pépitique) par exemple de la seconde composante. Ceci permet diverses variantes d'estimation comme krigeage et cokrigeage.

Enfin le variogramme calculé sans les valeurs extrêmes peut être utilisé pour estimer les valeurs inférieures au seuil, en supposant indépendance entre la distribution spatiale de ces valeurs, et la géométrie de leur champ, que l'on peut estimer par exemple par krigeage d'indicatrices. Mais il est alors nécessaire de faire des hypothèses sur ce qui se passe au-dessus de la coupure. Une hypothèse d'indépendance vis-à-vis de la géométrie fait jouer un rôle très particulier au seuil. A ce stade il paraît plus intéressant d'utiliser une factorisation, par exemple par résidus d'indicatrices, moyennant absence d'effets de bord montant ou descendant.

Plus généralement, on peut se placer dans le domaine de la géostatistique non-linéaire. Plutôt que de supprimer les valeurs extrêmes, il paraît alors plus judicieux d'utiliser des anamorphoses améliorant la robustesse.