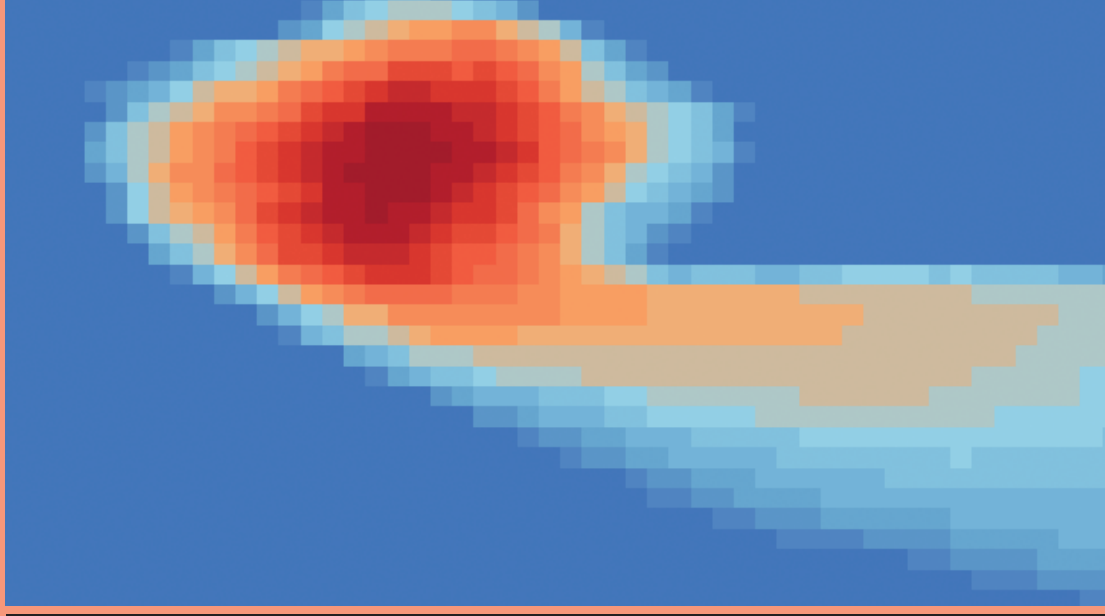


Les Cours

Chantal de Fouquet

Exercices corrigés de géostatistique

Presses des Mines

Chantal de Fouquet, *Exercices corrigés de géostatistique*, Paris : Presses des Mines, collection Les Cours, 2019.

60, boulevard Saint-Michel - 75272 Paris Cedex 06 - France
presses@mines-paristech.fr
www.pressesdesmines.com

ISBN : 978-2-35671-539-5

Dépôt légal : 2019
Achevé d'imprimer en 2019 (Paris)

Cette publication a bénéficié du soutien de l'Institut Carnot M.I.N.E.S

Exercices corrigés de géostatistique

Collection Les cours

Dans la même collection :

Renaud Gicquel, *Modéliser et simuler les technologies énergétiques*

Pascal Debu, *Noyaux et radioactivité*

Ali Azarian et Yann Pollet, *Analyse fonctionnelle des systèmes*

Francis Maisonneuve, *Mathématiques 2 et 3*

Bernard Wiesenfeld, *Une introduction à la neutronique*

Francis Maisonneuve, *Probabilités*

Francis Maisonneuve, *Mathématiques 2*

Renaud Gicquel, *Introduction aux problèmes énergétiques globaux*

Francis Maisonneuve, *Mathématiques 3*

Francis Maisonneuve, *Mathématiques 1*

J. Adnot, D. Marchio, Ph. Rivière, *Cycles de vie des systèmes énergétiques*

Brigitte d'Andréa-Novel, Benoît Fabre, Pierre Jouvelot, *Acoustique-Informatique-Musique*

Jean-Claude Moisdon, Michel Nakhla, *Recherche opérationnelle*

Anne-Françoise Gourgues-Lorenzen, Jean-Marc Haudin, Jacques Besson, *Matériaux pour l'ingénieur*

Renaud Gicquel, *Systèmes énergétiques T. 3*

Renaud Gicquel, *Systèmes énergétiques T. 2*

Renaud Gicquel, *Systèmes énergétiques T. 1*

Thierry Weil, *Stratégie d'entreprise*

François Cauneau, *Mécanique des fluides*

Pierre Chauvet, *Aide-mémoire de géostatistique linéaire*

Dominique Marchio, Paul Reboux, *Introduction aux transferts thermiques*

François Engel, Frédéric Kletz, *Cours de comptabilité analytique*

François Engel, Frédéric Kletz, *Cours de comptabilité générale*

Jacques Bouchard, Jean-Paul Deffain, Alain Gouchet, *Introduction au génie atomique*

Daniel Fargue, *Abrégé de thermodynamique : principes et applications*

Georges Pierron, *Introduction au traitement de l'énergie électrique*

Bernard Degrange, *Introduction à la physique quantique*

Michel Cohen de Lara, Brigitte d'Andréa-Novel, *Cours d'automatique*

Fixari Daniel, *Les Imperfections des marchés*

Jacques Lévy, *Introduction à la métallurgie générale*

Hugues Molet, *Comment maîtriser sa productivité industrielle ?*

Margaret Armstrong, Jacques Carignan, *Géostatistique linéaire*

Chantal de Fouquet

Exercices corrigés de géostatistique

MINES ParisTech | PSL

Avant-propos

Entre le cours de géostatistique et sa mise en œuvre, la modélisation explicite semble trop souvent une étape oubliée des praticiens, du fait de la disponibilité de « solutions logicielles » pas toujours facilement modifiables. Or les applications de la géostatistique dépassent largement l'usage « standardisé » auquel, faute de revenir à la mise en équations, les utilisateurs se restreignent.

Ce recueil d'exercices de géostatistique essentiellement linéaire se veut une aide pour approfondir la modélisation probabiliste, afin de décrire des phénomènes variés comme les concentrations (nutriments, substances polluantes…) dans les différents milieux (air, cours d'eau et nappes, sols et sédiments), les teneurs des gisements miniers, ou les strates d'une formation géologique. Les exercices sont réalisables sur table, avec calculette mais sans programmation.

La faible part de la géostatistique multivariable, omniprésente dans les applications, s'explique par le fait que de nombreux exercices ont été initialement rédigés comme sujets d'examens, ce qui implique des calculs modérés. À l'inverse, des thématiques d'une grande importance pratique apparaissent récurrentes, comme la comparaison entre échantillons simples ou composites, la prise en compte des « doublons », ou l'adaptation des conditions de non biais du krigeage. Des redondances sont ainsi présentes entre quelques exercices.

Les solutions paraîtront trop détaillées à certains, insuffisamment aux autres, ce recueil étant destiné aussi bien aux étudiants d'école d'ingénieurs ou de divers masters qu'aux praticiens. Enfin, les notations n'ont pas été homogénéisées ; ainsi, la variance est notée *Var* ou D^2. J'espère que ces imperfections ne rebuteront pas les lecteurs, à qui je souhaite d'apprécier la variété de la modélisation géostatistique.

Je remercie les étudiants des différentes formations, contributeurs involontaires à ce recueil, ainsi que les relecteurs de plusieurs chapitres.

Partie I

Étude exploratoire et variographie

Exercice 1

L'utilisation des statistiques est-elle toujours sensée ?

La figure 1 présente la concentration journalière en Ozone (un polluant atmosphérique), exprimée en µg/m^3, mesurée en une station durant l'année 2002. Les mesures sont effectuées en continu, les concentrations étant déduites du flux cumulé à travers le capteur durant 15 mn. Les concentrations horaires sont calculées comme les moyennes de quatre mesures quart horaires, et les concentrations journalières comme les moyennes de 24 concentrations horaires. La mesure du 31 décembre manquant, nous considérons dans la suite une « année » de 364 jours.

L'intervalle de temps auquel une donnée se rapporte (le quart d'heure, l'heure, le jour) est son « support temporel ».

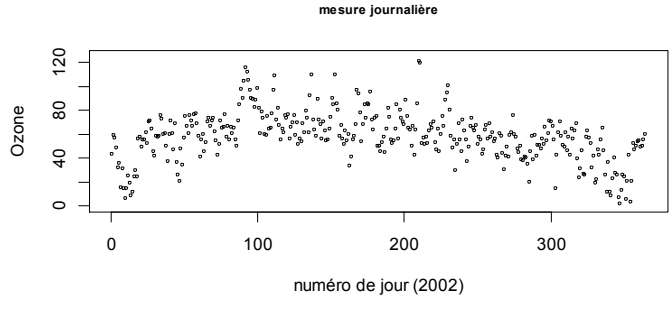


Figure 1. Concentrations journalières en Ozone, exprimées en µg/m^3, mesurées en 2002 en une station fixe. En abscisse, le temps exprimé en numéro de jour, compté depuis le 1er janvier 2002.

La concentration est notée z, les valeurs mesurées étant indicées par i.

Le résumé statistique de cette « population » de taille n est reporté au tableau 1. La moyenne $\bar{z}$ et la variance s^2 sont calculées comme

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad \text{et} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2$$

l'écart-type s étant la racine positive de la variance. La variance et l'écart-type caractérisent la dispersion de la population autour de sa moyenne. La variable étant positive, le coefficient de variation $s / \bar{z}$ caractérise la dispersion relative de la population.

Minimum, maximum, moyenne et écart-type ont la même unité que les données. La dimension physique de la variance est le carré de celle des données. Le coefficient de variation est adimensionnel.

Support temporel	Effectif	Minimum µg/m³	Maximum µg/m³	Moyenne µg/m³	Variance (µg/m³)²	écart-type µg/m³	coefficient de variation %
1 jour	364	1.30	121.00	57.82	425.05	20.62	35.60

Tableau 1. Résumé statistique des 364 mesures journalières de la concentration en Ozone.

Première partie : rappels

Les calculs statistiques usuels reposent sur les hypothèses suivantes :

- Hypothèse 1 : les données constituent un échantillon issu d'une distribution *Z,* d'espérance (de moyenne probabiliste) *E(Z)* = *m* et de variance $E((Z-m)^2) = \sigma^2$.

  Les concentrations $z_1,...z_n$ sont donc considérées comme un tirage de *n* variables aléatoires $Z_1,...Z_n$ de même espérance *m* et même variance σ^2.

- Hypothèse 2 : ces tirages sont indépendants.

  Les variables aléatoires $Z_1,...Z_n$ sont donc supposées mutuellement indépendantes.

1. La moyenne temporelle par intervalles de *k* jours $\left[i, i+k-1\right]$, qui peut être référencée à la date $i+\frac{k-1}{2}$, est notée $R_i^k = \frac{1}{k}\sum_{j=i}^{i+k-1} Z_j$.

Calculer l'espérance et la variance notée des variables aléatoires R_i^k qui seront notées $E\left(R^k\right)$ $D_0^2\left(R^k\right)$.

2. Estimation de l'espérance *m.*

L'espérance *m* de la distribution est estimée par la moyenne $\overline{Z}$ de l'échantillon : $m^* = \frac{1}{n}\sum_{j=1}^{n} Z_j$

L'erreur d'estimation est *m-m**. En exprimer l'espérance et la variance, notée $D^2\left(m-m^*\right)$.

Réponse : $E(m-m^*) = 0$ et pour un échantillon de taille *n*, $D^2\left(m-m^*\right) = \sigma^2 / n$

Deuxième partie : calculs statistiques

La figure 2 présente les concentrations mesurées, moyennées (ou « régularisées ») par intervalle de sept jours (une semaine), 14 jours (deux semaines) et 28 jours (quatre semaines). Les résumés statistiques sont les suivants :

Support temporel	**effectif**	**Minimum µg/m^3**	**Maximum µg/m^3**	**Moyenne µg/m^3**	**Variance (µg/m^3)²**	**Écart-type µg/m^3**	**Coefficient de variation %**
7 jours	52	16.97	98.64	57.82	269.74	16.42	28.40
14 jours	26	20.01	88.74	57.82	226.55	15.05	26.03
28 jours	13	30.04	83.07	57.82	200.24	14.15	24.47

Tableau 2. Résumé statistique des concentrations régularisées par intervalles de 7, 14 et 28 jours.

3. Comparer la moyenne et la dispersion (minimum, maximum, variance) des données en fonction de leur support temporel : 1,7,14 et 28 jours.
4. En utilisant le résultat de la question 1, déduire de la variance des données journalières, la variance « théorique » des régularisées sur 7, 14 ou 28 jours.
5. Comparer aux résumés statistiques du tableau 2 et commenter.
6. À l'aide des résultats de la question 2, calculer numériquement la variance d'estimation de l'espérance *m*, à partir des données journalières et de leurs régularisées sur respectivement 7,14 ou 28 jours.

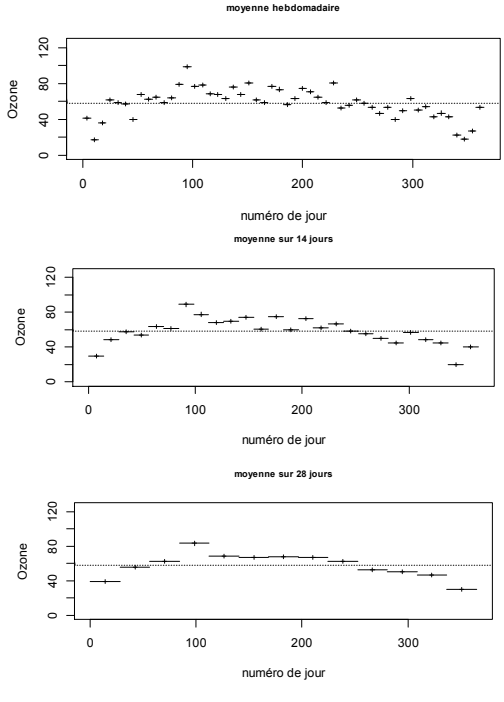


Figure 2. Concentration régularisée sur 7, 14 et 28 jours. La largeur des segments représente le support temporel des données. Le pointillé horizontal indique la moyenne annuelle de la concentration.

Troisième partie : regard critique

7. Les coefficients de corrélation entre valeurs successives sont les suivants :
   - 0.75 pour les concentrations journalières ;
   - 0.67 pour les régularisées hebdomadaires ;
   - 0.62 pour les régularisées durant deux ou quatre semaines.

L'hypothèse d'indépendance des tirages vous paraît-elle réaliste ?

8. Donner l'expression de $D^2\left(R^k\right)$ variance de la régularisée R_i^k (cf. question 1), dans le cas général où le coefficient de corrélation des variables Z_j et $Z_{j'}(j \neq j')$ est $r_{jj'}$.

Les $r_{jj'}$ sont maintenant supposés positifs, l'un d'eux au moins étant non nul. Écrire une inégalité entre $D^2\left(R^k\right)$ et $D_0^2\left(R^k\right)$, variance de la régularisée sous les hypothèses du calcul statistique (question 1).

Interpréter les résultats de la question 5.

9. En Europe, les épisodes de pollution par l'ozone ont lieu principalement en été. Les concentrations mesurées sont-elles en accord avec cette saisonnalité ?

L'hypothèse de stationnarité de l'espérance durant l'année vous paraît-elle réaliste ?

10. En l'absence de lacune et d'erreur dans les mesures, la moyenne annuelle des concentrations en 2002 est connue et égale à la moyenne des valeurs journalières, soit 57.82 µg/m^3. Dans le modèle probabiliste, la variance de l'erreur d'estimation de la moyenne annuelle par la moyenne des concentrations journalières devrait donc être nulle.

L'espérance *m* estimée par le calcul statistique usuel (question 2) coïncide-t-elle avec la moyenne annuelle en 2002 ?

Selon vous, que représentent les variances d'erreur calculées à la question 6 ?

Solution

Première partie : rappels

1. Les variables aléatoires $Z_i,...,Z_{i+k-1}$ ayant même espérance *m*,

$$E\left(R_i^k\right) = \frac{1}{k}\sum_{j=i}^{i+k-1} E\left(Z_j\right)$$
$$= \frac{1}{k}\sum_{j=1}^{k} m$$
$$= m$$

noté $E(R^k)$.

Les variables $Z_i,...,Z_{i+k-1}$ étant mutuellement indépendantes sont sans corrélation ; par suite, la variance de la somme est la somme des variances, et

$$D_0^2\left(R_i^k\right) = \frac{1}{k^2}\sum_{j=i}^{i+k-1} \sigma^2$$
$$= \sigma^2 / k$$

noté $D_0^2\left(R^k\right)$.

La variance de la moyenne sur *k* points est égale à la variance de la variable aléatoire *Z* divisée par *k*.

2. Estimation de l'espérance de la distribution sous-jacente.

D'après la question précédente, l'espérance de l'erreur d'estimation de la moyenne est nulle : $E(m-m^*) = 0$: la moyenne de l'échantillon est un estimateur sans biais de l'espérance de la distribution dont l'échantillon est issu.

L'espérance étant une constante déterministe,

$$D^2\left(m - m^*\right) = D^2\left(m^*\right)$$

d'où

$$D^2\left(m - m^*\right) = \sigma^2 / n$$

Caractérisée par la variance de l'erreur d'estimation, la précision sur l'espérance (la moyenne probabiliste) *m* de la distribution varie en raison inverse de la taille de l'échantillon.

L'écart-type de l'erreur d'estimation de l'espérance est $\sigma / \sqrt{n}$. Exprimé en %, l'écart-type d'estimation relatif est $100 \times \frac{\sqrt{D^2\left(m-m^*\right)}}{m^*}$ soit $\frac{100}{\sqrt{n}}\frac{\sigma}{m^*}$.

Deuxième partie : calculs statistiques

3. Les 364 jours constituant un nombre entier de semaines, de quinzaines (14 jours) et de « mois » de quatre semaines, la régularisation revient à moyenner les données journalières par classes de même effectif. Les régularisées ont donc la même moyenne que les concentrations journalières.

La dispersion des données décroît quand l'intervalle de régularisation augmente :

- le minimum croît et le maximum décroît ;
- la variance et donc l'écart-type décroissent.

Les données respectivement journalières, hebdomadaires, bimensuelles et mensuelles constituent donc des populations statistiques différentes.

4. D'après la question 1, la variance de la régularisée de k variables aléatoires indépendantes de même variance σ^2 est égale à σ^2 / k .

La variance des données journalières étant égale à 425.05, le calcul statistique prévoit donc les variances suivantes :

- régularisées hebdomadaires (k=7) : 60.72 $(\mu g/m^3)^2$
- régularisées par quinzaine (k=14) : 30.36 $(\mu g/m^3)^2$
- régularisées « mensuelles » (k=28): 15.18 $(\mu g/m^3)^2$

5. La variance prévue par le calcul statistique est très largement inférieure à la variance expérimentale des concentrations régularisées (tableau 2), respectivement égale à 269.74 pour les concentrations hebdomadaires, 226.55 pour les données par quinzaine et 200.24 pour les données régularisées sur quatre semaines. s_k^2 désignant la variance expérimentale des données régularisées sur k jours successifs, les rapports s^2 / s_k^2 sont les suivants :

   - 7 jours : 1.58 , au lieu de 7 par le calcul statistique ;
   - 14 jours : 1.88 , au lieu de 14 par le calcul statistique ;
   - 28 jours : 2.12 au lieu de 28 par le calcul statistique.

   L'une au moins des hypothèses du calcul statistique n'est donc pas vérifiée.

6. D'après la question 2, les estimations de l'espérance à partir des quatre ensembles de données sont identiques : 57.82 $\mu g/m^3$. Les variances d'estimation prévues par le calcul statistique sont les suivantes :

Support temporel en jours	**Variance d'estimation** $D^2(m\text{-}m^*)$	**Écart-type relatif (en %)** $100\,\frac{\sqrt{D^2(m-m^*)}}{m^*}$
1	1.17	1.87
7	5.19	3.94
14	8.71	5.10
28	15.40	6.79

La variance d'estimation issue du calcul statistique croît avec le support temporel des données car leur effectif décroît.

Troisième partie : regard critique

7. L'hypothèse d'indépendance des tirages n'est pas réaliste, les concentrations présentant une corrélation temporelle marquée. Cette corrélation est ici une fonction décroissante du support.

8. Variance de la régularisée, en présence de corrélation temporelle :

$$\begin{aligned} D^2\left(\frac{1}{k}\sum_{j=i}^{i+k-1} Z_j\right) &= E\left(\frac{1}{k}\sum_{j=i}^{i+k-1}\left(Z_j - m\right)\right)^2 \\ &= \frac{1}{k^2}\sum_{j=i}^{i+k-1} E\left[\left(Z_j - m\right)^2\right] + \frac{1}{k^2}\sum_{j\neq j'} E\left[\left(Z_j - m\right)\left(Z_{j'} - m\right)\right] \\ &= \frac{\sigma^2}{k} + \frac{\sigma^2}{k^2}\sum_{j\neq j'} r_{jj'} \end{aligned}$$

Lorsque les coefficients de corrélation $r_{jj'}$ sont tous positifs pour $j \neq j'$, l'un au moins étant non nul, le deuxième terme est strictement positif. La variance de la régularisée est alors strictement supérieure au résultat prévu par le calcul statistique usuel :

$$D^2\left(R_k^i\right) = D_0^2\left(R_k^i\right) + \frac{\sigma^2}{k^2}\sum_{j\neq j'} r_{jj'}$$

La présence de corrélation temporelle positive entre les concentrations à des dates voisines explique que les variances des régularisées sur 7, 14 ou 28 jours soient largement supérieures aux résultats prévus par le calcul statistique.

9. Les figures 1 et 2 montrent une variation saisonnière des concentrations, les données par quinzaine ou « mensuelles » étant toutes supérieures à la moyenne annuelle d'avril à septembre, et toutes inférieures à cette moyenne en automne et en hiver. Les concentrations mesurées à la station sont donc en accord avec la saisonnalité connue de l'ozone. L'hypothèse de stationnarité de l'espérance durant l'année ne décrit donc pas avec précision la réalité.

Noter qu'en 2002, la croissance des concentrations apparaît plus rapide en début d'année que leur décroissance en fin d'année.

10. Les mesures étant effectuées en continu, en l'absence de lacunes et d'erreurs dans les mesures, la moyenne des données (journalières, hebdomadaires...) représente exactement la moyenne annuelle des concentrations en 2002, égale à 57.82 µg/m^3. L'erreur d'estimation de cette moyenne annuelle par la moyenne des données (journalières, hebdomadaires, par quinzaine ou "mensuelles") est donc nulle.

La moyenne temporelle en 2002, $z_{2002} = \frac{1}{364}\sum_{i=1}^{364} z_i$ correspond dans le modèle probabiliste à la régularisée $Z_{2002} = \frac{1}{364}\sum_{i=1}^{364} Z_i$. Comme $Z_{2002} = \overline{Z}$, l'erreur d'estimation de Z_{2002} par la moyenne des variables aléatoires représentant les concentrations journalières ou leurs régularisées sur 7, 14 ou 28 jours est nulle, et par suite $D^2\left(Z_{2002} - \overline{Z}\right) = 0$.

Le calcul statistique usuel ne correspond pas à l'estimation de la régularisée Z_{2002} , mais à celle de l'espérance m de « la » distribution sous-jacente. Dans ce modèle $m = E\left(Z_{2002}\right)$ et $D^2\left(m - Z_{2002}\right)$, variance d'estimation de *m* par Z_{2002}, est non nulle.

Les variances d'estimation « statistique » issues du calcul à partir des concentrations journalières ou de leurs régularisées (question 6) ne reflètent donc pas l'incertitude sur la moyenne annuelle Z_{2002}. Du fait de la corrélation temporelle des concentrations, ces résultats ne représentent pas non plus l'incertitude sur l'espérance m supposée constante durant l'année.

Dans le cas d'une Fonction aléatoire *Z(t)* stationnaire d'ordre deux et ergodique, l'espérance m est la limite de la moyenne temporelle de *Z* sur des intervalles tendant vers l'infini. Cette limite *m* diffère des moyennes annuelles $Z_{2002} = \frac{1}{T}\int_{2002} Z(t)dt$, $Z_{2003} = \frac{1}{T}\int_{2003} Z(t)dt$ etc., qui présentent des variations autour de *m*.

En résumé, la variance d'estimation de la moyenne résultant du calcul statistique

- ne correspond pas à l'estimation d'une moyenne temporelle mais à celle d'une espérance mathématique, qui n'est pas la grandeur recherchée pour le « rapportage environnemental » ;
- ne tient pas compte de la corrélation temporelle présente dans la plupart des variables étudiées (température ou concentrations en différentes substances dans l'atmosphère, les nappes ou les cours d'eau par exemple).

Par ailleurs, le calcul statistique ne tient pas compte des dates de mesures. En présence de saisonnalité, un échantillonnage intermittent préférentiel (plus dense ou au contraire lacunaire durant les périodes de valeurs fortes ou faibles) induit un biais dans l'estimation de la moyenne annuelle par la moyenne arithmétique des données.

Enfin, dans le cas d'un échantillonnage régulier suffisamment dense, la variance d'estimation issue du calcul statistique est généralement trop élevée par rapport à l'incertitude sur la moyenne annuelle, car la variance des données reflète non seulement les fluctuations aux petits intervalles de temps mais aussi la saisonnalité des concentrations.

La variance d'estimation issue du calcul statistique peut ainsi n'avoir aucun sens (C. Bernard-Michel, 2006). En géostatistique, le calcul d'une variance d'estimation nécessite d'expliciter la grandeur à estimer : moyenne temporelle ou espérance mathématique, par exemple. Ce calcul fait intervenir les dates et le support des données, ainsi que la corrélation temporelle modélisée.

Remarque : pour l'ozone, les normes relatives à la pollution atmosphérique portent sur les concentrations élevées (quantiles, cumul de dépassements de valeur-seuil). En qualité de l'air, les valeurs réglementaires sur la moyenne annuelle concernent par exemple le dioxyde d'azote.

Exercice 2

Étude exploratoire d'une pollution de sol

Le sol d'une friche industrielle comporte un niveau horizontal de remblais superficiels d'épaisseur constante recouvrant la formation sous-jacente. Le site a été reconnu suivant une grille horizontale à maille carrée, comportant 10 multiplié par 10 nœuds. En chaque nœud, un sondage vertical a été foré. La carotte a été découpée en deux tronçons de même longueur, l'une prélevée dans les remblais superficiels et l'autre dans la formation sous-jacente. Par niveau, 100 données sont ainsi disponibles.

$x=(x_1,x_2)$ désigne un point du plan. Le centre des carottes dans la couche superficielle est à la cote $x_3= -0.5m$, et celui des carottes du niveau inférieur à la cote $x_3= -1.5m$.

Nous allons examiner (schématiquement) la variabilité des teneurs sur le site, à l'aide des outils de l'analyse exploratoire et de l'analyse variographique.

La teneur du niveau superficiel est interprétée comme une réalisation d'une fonction aléatoire $S(x)$, et celle du niveau inférieur, comme une réalisation d'une fonction aléatoire $Y(x)$, ces fonctions étant supposées stationnaires d'ordre deux.

Pour simplifier, la teneur (exprimée en %) est discrétisée en trois valeurs par niveau.

Première partie : analyse statistique exploratoire

1. Remblais superficiels

Dans la couche superficielle, les teneurs peuvent prendre les valeurs s_1, s_2 et s_3, telles que $s_1 = s_2-\delta$ et $s_3 = s_2+\delta$, avec $\delta > 0$. Parmi les 100 teneurs mesurées, 40% sont égales à s_1, 40% à s_2, et 20% à s_3. Dans les questions suivantes, on exprimera s_1 et s_3 en fonction de s_2 et de δ.

La fréquence des teneurs égales à s_i dans les remblais superficiels est notée p_i.

a) Tracer l'histogramme des teneurs des 100 échantillons, en exprimant l'ordonnée en % de l'effectif.

b) Calculer la moyenne m de cette distribution.

c) Calculer la variance σ^2 de cette distribution (moyenne du carré de l'écart à la moyenne m).

2. Formation sous-jacente

Les teneurs sont globalement plus élevées dans la formation sous-jacente. Les trois valeurs mesurées sont y_1,y_2 et y_3 avec $y_1 = s_1+D$, $y_2= s_2+D$, $y_3= s_3+D$, D étant une

constante déterministe positive. Les fréquences des y_i, notée q_i, sont ici égales à 20%, 60% et 20%.

a) Tracer l'histogramme des teneurs des 100 échantillons, en exprimant l'ordonnée en % de l'effectif.

b) Calculer la moyenne m' des 100 échantillons du niveau plus profond.

c) Calculer la variance σ'^2 de cette distribution.

3. Les données des deux niveaux sont mélangées.

Notons z_i l'ensemble des valeurs possibles de la teneur dans les deux niveaux. On posera $p_i = 0$ pour $z_i > s_3$ et $q_i = 0$ pour $z_i < y_1$.

a) Exprimer la moyenne M des 200 échantillons en fonction de s_2, δ et D.

b) Exprimer la variance S^2 des 200 échantillons en fonction de z_i, p_i, q_i et M.

c) En exprimant l'ordonnée en % de l'effectif total, tracer l'histogramme des 200 données dans les cas suivants : $D = 0$, $D = \delta$, $D = 2\delta$, $D = 3\delta$, $D = 4\delta$, $D = 5\delta$.

4. Pour quelles valeurs de D l'histogramme global peut-il être supposé unimodal ou bimodal ?

5. Conclusion : lorsque les données des deux niveaux sont mélangées, l'histogramme global permet-il toujours de détecter l'hétérogénéité des teneurs (c'est-à-dire la présence de deux populations statistiquement différentes) ?

Deuxième partie : variogramme vertical

6. Soit ρ le coefficient de corrélation au point x entre la teneur $S(x)$ du niveau superficiel et celle $Y(x)$ de la formation sous-jacente.

Exprimer l'espérance de l'écart quadratique $\left(Y(x)-S(x)\right)^2$ à l'aide de m, m', σ, σ' et ρ.

En déduire le premier pas (pour $h_3 = 1$ m) du variogramme vertical de la teneur.

7. Examiner les cas particuliers $\rho = 0$ et $\rho \to 1$.

Troisième partie : variogramme horizontal des remblais superficiels

Lors du démantèlement du site, les remblais superficiels ont été déposés en « andains » suivant des lignes parallèles à Ox_1. Par suite, la teneur S est anisotrope dans le plan horizontal. Elle présente une corrélation spatiale suivant la direction Ox_1, et est pépitique suivant la direction Ox_2. Les teneurs des lignes d'ordonnées x_2 et x_2+h_2 n'étant pas corrélées :

$$\forall (h_1,h_2) \; tel\; que \; |h_2| > 0, \; Cov\left(S(x_1,x_2), S(x_1+h_1, x_2+h_2)\right) = 0$$

8. Le champ est supposé suffisamment grand pour que σ^2 soit une bonne approximation de la variance dans un champ infini. Tracer le variogramme de S dans la direction Ox_2, noté $\gamma(0,h_2)$, en indiquant le palier.

9. Variogramme suivant la direction Ox_1.

Pour calculer le variogramme suivant la direction Ox_1, nous supposerons la droite discrétisée au pas d. Les teneurs aux points de même ordonnée x_2 et d'abscisses respectives x_1 et x_1+d sont telles que:

- avec la probabilité p, $S(x_1+d, x_2)=S(x_1, x_2)$,
- avec la probabilité complémentaire $q = 1-p$, $S(x_1+d, x_2)$ est supposé indépendant des $S(u, x_2)$ pour $u \leq x_1$.

a) Calculer la covariance $C(d,0)$ entre $S(x_1+d, x_2)$ et $S(x_1, x_2)$

b) Calculer la covariance $C(2d,0)$ entre $S(x_1+2d, x_2)$ et $S(x_1, x_2)$

c) Par récurrence, en déduire la covariance entre $S(x_1+nd, x_2)$ et $S(x_1, x_2)$.

d) En déduire la forme de la covariance discrète $C(h_1,0)$ et en préciser la portée et le palier. Donner l'expression du variogramme $\gamma(h_1,0)$ associé.

10. Variogramme horizontal « moyen »

Un géostatisticien débutant a calculé le variogramme omnidirectionnel, sans contrôler la présence d'une éventuelle anisotropie. Pour simplifier, nous admettrons qu'il a effectué la moyenne des variogrammes directionnels dans les directions Ox_1 et Ox_2 :

$$\gamma_{12}(h) = \frac{1}{2}\left(\gamma(h,0) + \gamma(0,h)\right)$$

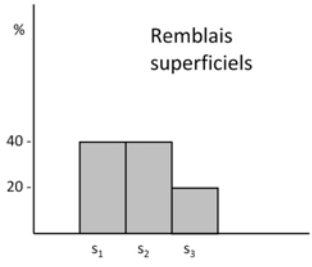
Tracer schématiquement sur la même figure le variogramme directionnel suivant Ox_1, celui suivant Ox_2, ainsi que le variogramme « omnidirectionnel » $\gamma_{12}(h)$.

Solution

Première partie : analyse statistique exploratoire

1. Couche superficielle

   a) L'histogramme est unimodal.



   b) La moyenne est la somme des valeurs pondérée par leur fréquence : $m = \sum_{i=1}^{3} p_i s_i$ d'où

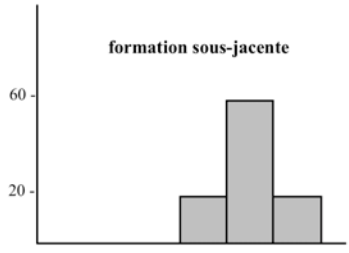
$$\begin{aligned} m &= \tfrac{2}{5}s_1 + \tfrac{2}{5}s_2 + \tfrac{1}{5}s_3 \\ &= \tfrac{2}{5}(s_2 - \delta) + \tfrac{2}{5}s_2 + \tfrac{1}{5}(s_2 + \delta) \\ &= s_2 - \tfrac{1}{5}\delta \\ &= s_2 - 0.2\delta \end{aligned}$$

   c) La variance de la distribution s'écrit

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{3} p_i (s_i - m)^2 \\ &= \tfrac{2}{5}\left(s_2 - \delta - \left(s_2 - \tfrac{1}{5}\delta\right)\right)^2 + \tfrac{2}{5}\left(s_2 - \left(s_2 - \tfrac{1}{5}\delta\right)\right)^2 + \tfrac{1}{5}\left(s_2 + \delta - \left(s_2 - \tfrac{1}{5}\delta\right)\right)^2 \\ &= \frac{\delta^2}{5}\frac{1}{5^2}\left[2\times 16 + 2\times 1 + 36\right] \\ &= \frac{2\times 7}{5^2}\delta^2 \\ &= 0.56\,\delta^2 \end{aligned}$$

2. Formation sous-jacente.

   a) L'histogramme est symétrique



   b) La loi étant symétrique, la moyenne est égale à la médiane y_2, d'où $m' = s_2+D$

Ce résultat se retrouve par le calcul suivant :

$$\begin{aligned} m' &= \tfrac{1}{5}\left(s_2 - \delta + D\right) + \tfrac{3}{5}\left(s_2 + D\right) + \tfrac{1}{5}\left(s_2 + \delta + D\right) \\ &= s_2 + D \end{aligned}$$

L'histogramme des teneurs de cette couche est caractérisé par les probabilités associées à la teneur minimale $y_{min} = y_1$, à la teneur moyenne $y_{moy} = y_2$, et à la teneur maximale $y_{max} = y_3$.

   c) Variance de la distribution :

$$\sigma'^2 = q_1\left(y_1 - m'\right)^2 + q_2\left(y_2 - m'\right)^2 + q_3\left(y_3 - m'\right)^2$$

y_2 étant égal à la moyenne,

$$\begin{aligned} \sigma'^2 &= \tfrac{1}{5}\left(s_2 - \delta + D - s_2 - D\right)^2 + \tfrac{1}{5}\left(s_2 + \delta + D - s_2 - D\right)^2 \\ &= \tfrac{2}{5}\delta^2 \\ &= 0.4\,\delta^2 \end{aligned}$$

Les teneurs de la formation sous-jacente sont moins dispersées que celles des remblais superficiels.

La variance ne dépend pas des constantes déterministes s_2 et D.

3. Mélange des données des deux niveaux

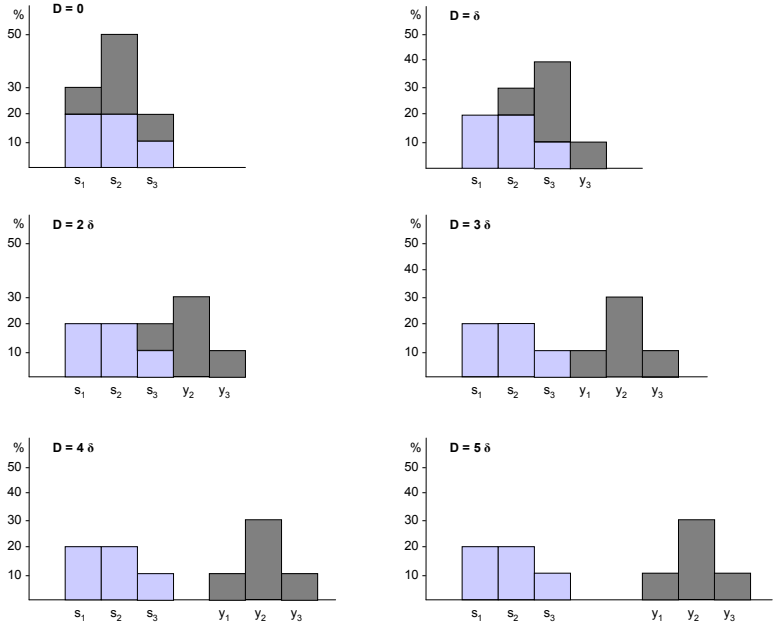
   a) Les deux sous-populations ayant le même effectif, la moyenne M de la population globale est la moyenne de m et m' :

$$M = \frac{m + m'}{2} = s_2 + \frac{D}{2} - \frac{1}{10}\delta$$

b) Chacune des classes d'un niveau intervient avec la probabilité $p_i/2$ ou $q_i/2$ (éventuellement nulle) dans la population globale. La variance s'écrit donc :

$$S^2 = \sum_i \frac{p_i + q_i}{2} \left(z_i - M\right)^2$$

c) Les histogrammes sont obtenus en combinant les deux précédents, chaque niveau comptant pour 50% du total.



La moyenne et l'écart-type de la population globale sont reportés au tableau ci-après. La dispersion de la population globale croît avec *D*.

4. Conclusion : lorsque les données des deux niveaux sont mélangées, l'histogramme global ne permet pas nécessairement de détecter l'hétérogénéité des teneurs entre les deux niveaux. Si les deux modes sont rapprochés, l'histogramme global apparaît en effet unimodal (cas $D = \delta$ ou $D = 2\delta$).

Lorsque l'histogramme est bimodal, il convient évidemment de caractériser l'hétérogénéité.

Lorsque la cote est connue, la moyenne des teneurs par niveaux, le coloriage de l'histogramme global en fonction du niveau (cf. figure précédente), ou encore les nuages de corrélation entre teneur et cote ainsi qu'entre la teneur du niveau supérieur et celle du niveau inférieur aux mêmes points permettent de préciser la répartition des teneurs en fonction de la profondeur.

$\frac{D}{\delta}$	M	$z_i ; (z_i - M) ; \frac{(z_i - M)^2}{\delta^2}$	$\frac{(p_i + q_i)}{2}$ (en %)	$\frac{s^2}{\delta^2}$	$\frac{s}{\delta}$
0	s_2 - δ/10	$s_2 - \delta$; - 9δ/10 ; 81/100 s_2 ; δ/10 ; 1/100 $s_2 + \delta$; 11δ/10 ; 121/100	20+10=30 20+30=50 10+10=20	49/100	0.70
1	$s_2 + 2\delta/5$ $= s_2 + 4\delta/10$	$s_2 - \delta$; - 7δ/5 ; 49/25 s_2 ; - 2δ/5 ; 4/25 $s_2 + \delta$; 3δ/5 ; 9/25 $s_2 + 2\delta$; 8δ/5 ; 64/25	20 20+10=30 10+30=40 10	21/25	0.92
2	$s_2 + 9\delta/10$	$s_2 - \delta$; -19δ/10 ; 361/100 s_2 ; - 9δ/10 ; 81/100 $s_2 + \delta$; δ/10 ; 1/100 $s_2 + 2\delta$; 11δ/10 ; 121/100 $s_2 + 3\delta$; 21δ/10 ; 441/100	20 20 10+10=20 30 10	169/100	1.30
3	$s_2 + 7\delta/5$ $= s_2 + 14\delta/10$	$s_2 - \delta$; -12δ/5 ; 144/25 s_2 ; -7δ/5 ; 49/25 $s_2 + \delta$; -2δ/5 ; 4/25 $s_2 + 2\delta$; 3δ/5 ; 9/25 $s_2 + 3\delta$; 8δ/5 ; 64/25 $s_2 + 4\delta$; 13δ/5 ;169/25	20 20 10 10 30 10	76/25	1.74
4	$s_2 + 19\delta/10$	$s_2 - \delta$; -29δ/10 ; 841/100 s_2 ; -19δ/10 ; 361/100 $s_2 + \delta$; -9δ/10 ; 81/100 $s_2 + 3\delta$; 11δ/10 ; 121/100 $s_2 + 4\delta$; 21δ/10 ; 441/100 $s_2 + 5\delta$; 31δ/10 ; 961/100	20 20 10 10 30 10	489/100	2.21
5	$s_2 + 12\delta/5$ $= s_2 + 24\delta/10$	$s_2 - \delta$; -17δ/5 ; 289/25 s_2 ; -12δ/5 ; 144/25 $s_2 + \delta$; -7δ/5 ; 49/25 $s_2 + 4\delta$; 8δ/5 ; 64/25 $s_2 + 5\delta$; 13δ/5 ; 169/25 $s_2 + 6\delta$; 18δ/5 ; 324/25	20 20 10 10 30 10	181/25	2.69

5. Conclusion : lorsque les données des deux niveaux sont mélangées, l'histogramme global ne permet pas nécessairement de détecter l'hétérogénéité des teneurs entre les deux niveaux. Si les deux modes sont rapprochés, l'histogramme global apparaît en effet unimodal (cas $D = \delta$ ou $D = 2\delta$).

Lorsque l'histogramme est bimodal, il convient évidemment de caractériser l'hétérogénéité.

Lorsque la cote est connue, la moyenne des teneurs par niveaux, le coloriage de l'histogramme global en fonction du niveau (cf. figure précédente), ou encore les

nuages de corrélation entre teneur et cote ainsi qu'entre la teneur du niveau supérieur et celle du niveau inférieur aux mêmes points permettent de préciser la répartition des teneurs en fonction de la profondeur.

Deuxième partie : variogramme vertical

6. L'écart quadratique au point $x = (x_1, x_2)$ entre la teneur dans les remblais et celle dans la formation sous-jacente s'écrit

$$\begin{aligned}\big(Y(x)-S(x)\big)^2 &= \Big(\big(Y(x)-m'\big)+\big(m'-m\big)-\big(S(x)-m\big)\Big)^2 \\ &= \big(Y(x)-m'\big)^2+\big(S(x)-m\big)^2+\big(m'-m\big)^2+2\big(Y(x)-m'\big)\big(m'-m\big) \\ &\quad -2\big(Y(x)-m'\big)\big(S(x)-m\big)-2\big(m'-m\big)\big(S(x)-m\big)\end{aligned}$$

Les fonctions aléatoires S et Y étant supposées d'espérances stationnaires, $E(S(x)-m) = 0$. D'où

$$E\Big[\big(Y(x)-S(x)\big)^2\Big]=\big(m'-m\big)^2+E\Big[\big(Y(x)-m'\big)^2\Big]+E\Big[\big(S(x)-m\big)^2\Big]-2E\Big[\big(Y(x)-m'\big)\big(S(x)-m\big)\Big]$$

La covariance de deux variables aléatoires est égale au produit de leur coefficient de corrélation par le produit des deux écarts-types. Par suite :

$$\begin{aligned}E\Big[\big(Y(x)-S(x)\big)^2\Big] &= \big(m'-m\big)^2+Var\,Y(x)+Var\,S\big(x\big)-2Cov\big(S\big(x\big),Y(x)\big) \\ &= \big(m'-m\big)^2+\sigma^2+\sigma'^2-2\rho\sigma\sigma'\end{aligned}$$

Le premier pas (pour $h_3 = 1$m) du variogramme vertical de la teneur s'exprime, au facteur ½ près comme l'espérance de cet écart-quadratique. Par suite :

$$\begin{aligned}\gamma_{vert}\big(h_3\big) &= \tfrac{1}{2}E\Big[\big(Y(x)-S(x)\big)^2\Big] \\ &= \tfrac{1}{2}\big(m'-m\big)^2+\tfrac{\sigma^2+\sigma'^2}{2}-\rho\sigma\sigma'\end{aligned}$$

7. Cas particuliers

   - Si $\rho = 0$, $\gamma_{vert}\big(h_3\big)=\tfrac{1}{2}\big(m'-m\big)^2+\tfrac{\sigma^2+\sigma'^2}{2}$: le variogramme dépend de l'écart quadratique entre les moyennes des deux niveaux, ainsi que de la moyenne des variances.

Lorsque $m = m'$ et $\sigma = \sigma'$, on retrouve classiquement la composante pépitique σ^2.

   - Dans le cas d'une corrélation parfaite entre les deux niveaux : $\rho(h_3) = 1$. Alors, $\gamma_{vert}\big(h_3\big)\to\tfrac{1}{2}\big(m'-m\big)^2+\tfrac{1}{2}\big(\sigma-\sigma'\big)^2$.

Dans ce cas, le variogramme expérimental est nul lorsque $m = m'$ et $\sigma = \sigma'$.

Table des matières

Ce recueil d'exercices vise à faire le lien entre la théorie de la géostatistique et la pratique. Savoir poser les équations simples de la géostatistique linéaire permet en effet de comprendre la variation du coefficient de corrélation de deux concentrations avec le support (valeurs horaires, journalières ou hebdomadaires, par exemple), de raisonner le choix entre des échantillons simples ou composites pour établir un schéma de reconnaissance, ou encore de caractériser l'évolution temporelle d'une concentration en estimant la différence de deux moyennes annuelles et en calculant la variance d'erreur associée. De nombreux exercices sont directement inspirés de cas réels.

Ce recueil s'adresse aux praticiens désireux d'approfondir leurs connaissances en géostatistique et d'en diversifier les applications, ainsi qu'aux étudiants (niveau école d'ingénieur ou master) et doctorants, soucieux d'en maîtriser les concepts.

Principalement issus des applications environnementales de la géostatistique, les exemples se réfèrent à la pollution de l'air, des cours d'eau ou des sols ; beaucoup sont transposables à d'autres contextes, en particulier à l'estimation minière.

Chantal de Fouquet est directrice de recherche en géostatistique à MINES ParisTech, où elle développe et enseigne les applications environnementales de la géostatistique, notamment pour la cartographie des pollutions dans les différents milieux (air, cours d'eau et nappes, sols).

39 euros