

CHAPITRE 1 : INTRODUCTION A LA GEOSTATISTIQUE

1.1. INTRODUCTION

Au passé, les ingénieurs des mines faisaient des calculs pour évaluer les ressources en minerai d'un gisement à partir d'un petit nombre de sondages prélevés en des sites irrégulièrement repartis dans le domaine d'étude.

La réserve totale disponible était inconnue et traitée comme une variable aléatoire. Mais il était impossible d'assimiler les teneurs mesurées aux différents sites sondés à des réalisations de variables aléatoires indépendantes.

Si on suppose une indépendance statistique entre les mesures réalisées entre différents points de l'espace, la meilleure prédiction que l'on peut faire de la teneur en un site non informé (i.e. ou l'on n'a pas réalisé de sondage) est d'attribuer la moyenne de l'échantillon.

On sent bien que cette "solution" a quelque chose de sous-optimal, en particulier il semble souhaitable d'utiliser une méthode qui donne plus de poids aux sites proches qu'aux autres points de mesure.

Cette situation :

- (i) un formalisme probabiliste pour représenter des quantités inconnues ;
- (ii) l'impossibilité de supposer une indépendance entre les données ;
- (iii) l'existence d'une structuration de la variable étudiée par rapport aux coordonnées d'espace est caractéristiques de la statistique spatiale.

Quand les mesures sont réalisées en des sites irrégulièrement espacées choisis par l'expérimentateur (on dit alors que la position des sites est non informative), on se trouve exactement dans la situation du problème d'évaluation d'un gisement, qui a donné lieu à de nombreux d'enveloppements méthodologiques: la **Géostatistique**.

La Géostatistique : méthodes probabilistes pour l'étude des phénomènes corrélés dans l'espace. Créée par Georges Matheron pour résoudre des problèmes d'estimation minière.

La Géostatistique : une démarche orientée prédiction pour caractériser la variabilité spatiale, la modéliser et prédire.

1.2. METHODES EXPLORATOIRES

1.2.1 Vocabulaire

En général un jeu de données spatiales se présente sous la forme d'une liste de valeurs numériques $z = (z_1, \dots, z_n)$ (la variable) et des coordonnées des sites $(s_1; \dots, s_n)$ auxquelles la variable a été mesurée.

Quand on réalise une mesure physique, cette mesure est associée à une longueur, une surface ou un volume élémentaire. Par exemple on peut mesurer la pluie avec un pluviomètre qui donne la hauteur d'eau tombée sur une surface d'environ 400cm^2 , mesurer une teneur en minerai dans le sol avec une carotte de 500cm^3 . Ces longueurs, surfaces, volumes, élémentaires constituent ce que l'on appelle le support de la mesure. Une mesure réalisée à support spatial donné peut être réalisée à différents supports temporels.

Par exemple, on peut mesurer la pluie avec un pluviographe pendant une journée et en déduire la pluie moyenne sur la journée ou la pluie maximale sur la journée. Il s'agit de deux variables distinctes qui auront des propriétés géostatistiques très différentes. Lorsqu'on cherche à déduire les propriétés statistiques d'une variable à un certain support à partir de cette même variable à un autre support, on dit qu'on réalise un changement de support.

Il ne faut pas confondre la notion de support avec la notion d'échelle qui désigne quelque chose de beaucoup plus vague selon le contexte. Il peut s'agir de la taille du domaine étudié ou du support temporel de la mesure.

1.2.2 Distribution monovariante

Avant d'envisager quoi que ce soit il est bon de représenter l'histogramme de z , ce qui permet d'évaluer à quel point la distribution s'écarte d'une loi gaussienne. On verra par la suite que beaucoup de méthodes atteignent une efficacité maximale lorsque la distribution des z est celle d'un vecteur multi-gaussien, ce qui suppose entre autre que la distribution marginale soit gaussienne. L'écart au caractère gaussien peut se manifester par

- une dissymétrie
- un ensemble de valeurs possible borné
- une accumulation de valeurs en un (ou plusieurs) points (distribution atomique)

1.2.3 Répartition spatiale

La distribution de z doit être complétée par une étude de la répartition spatiale des valeurs de z . On peut le faire en traçant un "symbol plot" dans lequel on représente en chaque site un symbole (croix, étoile, cercle) dont la taille est proportionnelle à la valeur observée z_i .

L'examen de ce symbole plot permet de détecter (de manière qualitative encore) la présence de variations "systématiques" dans les données, on parlera de tendance ou de dérive.

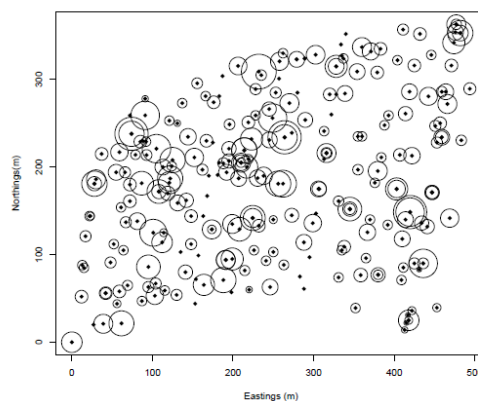


Figure1 : Symbol Plot.

Sur la figure on représente les profondeurs de sol en plusieurs points d'une parcelle de terrain. Le rayon du cercle est proportionnel à la profondeur mesurée. Ces profondeurs sont assez hétérogènes. Les valeurs sont assez hétérogènes mais aucune tendance régulière ne se manifeste.

1.2.4 Nuée variographique

A ce stade, on ne s'est pas encore préoccupé de l'éventuelle dépendance entre les données. C'est ce que l'on fait en construisant ce que l'on appelle la nuée variographique.

Il s'agit du nuage de points formés par les couples $(s_i - s_j); (z_i - z_j)^2$.

Ce nuage présente l'allure de la Figure 2, ci-dessous. La nuée variographique de profondeurs de sol (Figure de Gauche) ne fait pas apparaître de liaison entre l'écart $(z_i - z_j)^2/2$ et la distance correspondante.

En revanche, la différence quadratique $\gamma_{ij} = (z_i - z_j)^2/2$ des résistivités a tendance à croître en fonction de la distance $h_{ij} = s_i - s_j$. (Figure de Droite)

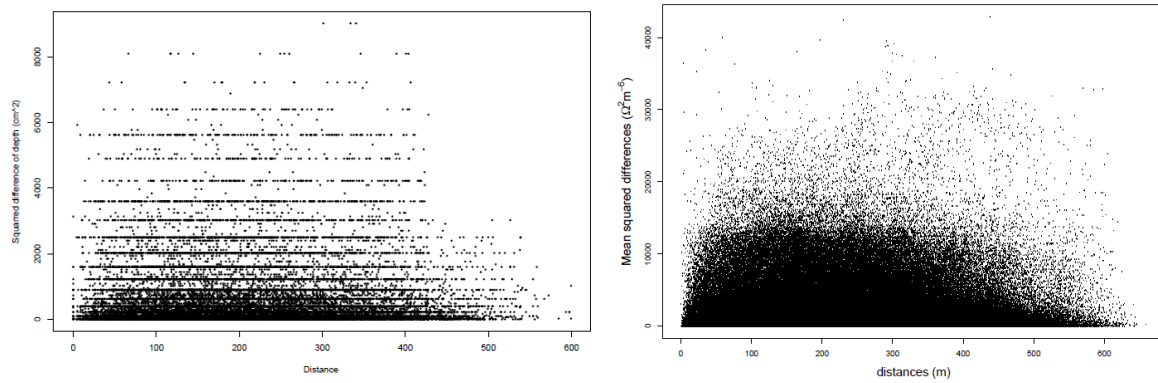


Figure 2 : Nuée variographique

Autrement dit, les paires de valeurs de z mesurées en des sites géographiquement proches ont tendance à être plus similaires que celle mesurées en des sites éloignés. C'est ce genre de comportement qui fait qu'on doit rejeter l'hypothèse d'indépendance entre les données.

En effet, dans le cas d'un échantillon i.i.d les γ_{ij} seraient des variables aléatoires de même loi et de même variance.

Rappels : En [théorie des probabilités](#) et en [statistique](#), des **variables indépendantes et identiquement distribuées** sont des [variables aléatoires](#) qui suivent toutes la même [loi de probabilité](#) et sont [indépendantes](#). On dit que ce sont des **variables aléatoires iid** ou plus simplement des **variables iid**. Cas discret - deux variables aléatoires réelles X, Y sont dites indépendantes et identiquement distribuées si elles sont :

- indépendantes : c'est-à-dire, $P(X=x, Y=y)=P(X=x) \times P(Y=y)$ pour tous réels x, y .
- de même loi : c'est-à-dire, $P(X=x) = P(Y=x)$ pour tout réel x .

Cas continu - deux variables aléatoires réelles X, Y sont dites indépendantes et identiquement distribuées si elles sont :

- indépendantes : c'est-à-dire, $f_{X,Y}(x, y)=f_X(x) \times f_Y(y)$ pour tous réels x, y .
- de même loi : c'est-à-dire, $f_X(x)= f_Y(x)$ pour tout réel x .

où $f_{X,Y}$, f_X , f_Y sont les densités de probabilité du couple (X,Y) et des variables X et Y .

1.3 FONCTIONS ALEATOIRES, STATIONNARITE, COVARIANCE

1.3.1 Fonction aléatoire

La question qui se pose maintenant est de modéliser la distribution statistique de l'échantillon $z_1, z_2, \dots, z_n$. La modélisation géostatistique repose sur les notions sur les notions fondamentales de fonction aléatoire et de loi spatiale. Pour en faciliter la compréhension, un certain nombre de définitions s'impose. Dans toute la suite, D désigne le domaine étudié et s les points de ce domaine.

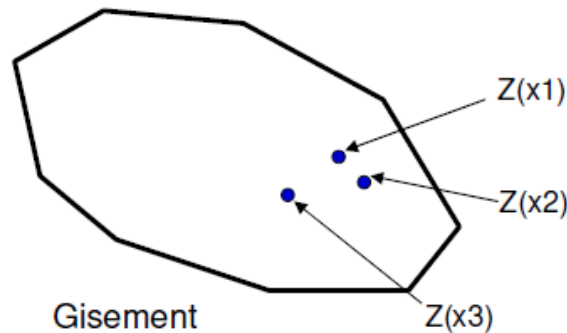
Variable aléatoire : variable dont les réalisations possibles sont connues mais dont le résultat final ne peut être déterminé avant d'effectuer la mesure.

En un point s , la variable aléatoire considérée est notée $Z(s)$.

$z(s)$, valeur observée, est une réalisation de $Z(s)$.

Fonction aléatoire : famille de toutes les variables aléatoires $\{Z(s) ; s \text{ appartient à } D\}$. Elle est définie dans tout le domaine, y compris aux points où il n'y a pas de mesure disponible.

La figure suivante représente le point de vue de la géostatistique :



Gisement => infinité de points ou très grand nombre de « quasi-points »

x : emplacement géographique

Chaque point $\rightarrow$ teneur $\rightarrow Z(x)$

Chaque teneur $\rightarrow$ v. a. (ensemble forme une fonction aléatoire (de x))

Variable régionalisée : réalisation $\{z(s) ; s \text{ appartient à } D\}$ d'une fonction aléatoire.

Champ : domaine de l'espace où la variable régionalisée prend ses valeurs.

Loi spatiale : donnée de toutes les lois de probabilité qui régissent les vecteurs aléatoires $\{Z(s_1) \dots Z(s_n)\}$ que l'on peut extraire de $\{Z(s) ; s \text{ appartient à } D\}$. En pratique la description de la loi spatiale se limite à ses deux premiers moments : moyenne et covariance.

Moyenne (ou espérance mathématique) : $E[Z(s)] = m(s)$

L'espérance est la moyenne autour de laquelle les valeurs possibles de $Z(s)$ se distribuent.

Variance:

$$\text{Var}[Z(s)] = E\{[Z(s) - m(s)]^2\} = E\{[Z(s)]^2\} - [m(s)]^2$$

La variance mesure la dispersion de $Z(s)$ autour de sa moyenne $m(s)$. Elle est égale à la covariance de $Z(s)$ avec elle-même.

Covariance :

Soient s_1 et s_2 , deux points de D

$$\text{Cov}[Z(s_1), Z(s_2)] = E\{[Z(s_1) - m(s_1)][Z(s_2) - m(s_2)]\} = E[Z(s_1)Z(s_2)] - [m(s_1)][m(s_2)]$$

La covariance quantifie le lien linéaire entre les valeurs prises en s_1 et s_2 .

Dans le cas d'un échantillon i.i.d, la caractérisation est facile, "il suffit" de spécifier une fonction de répartition (communes à toutes les variables aléatoires de l'échantillon).

1.3.2 Fonction de covariance

On peut décrire partiellement la distribution statistique d'une fonction aléatoire via des **moments** d'ordre **un** et **deux** tels que l'**espérance** et la **covariance** (qui sont ici des fonctions des coordonnées d'espace) définis par :

$$\begin{aligned}\mu(s) &= E[Z(s)] \\ C(s, s') &= \text{Cov}[Z(s), Z(s')]\end{aligned}$$

La covariance satisfait deux propriétés importantes :

1. **Symétrie** : $C(s, s') = C(s', s)$
2. **Défini-positivité** : Considérons par exemple la variance d'une somme du type $\text{Var}[\sum_1^n \lambda_i Z_{s_i}]$. En tant **que** variance, cette somme doit être positive, ceci quels que soient les λ_i , ce qui exprime que la fonction C est une fonction de type positif.

1.3.3 Stationnarité

En **géostatistique**, une **fonction aléatoire stationnaire d'ordre 2** est une **fonction aléatoire** Z sur un espace S telle que :

- sa moyenne est constante : $E_x[Z(x)] = m$
- sa **covariance** est invariante par translation, fonction du seul vecteur distance h entre les deux points : $\text{Cov}[Z(x), Z(x+h)] = C(h)$

Le processus est dit **centré** si sa moyenne est nulle en tout point

Lorsque les deux hypothèses précédentes sont vérifiées, la description de la distribution statistique de Z s'en trouve considérablement simplifiée. Tous les moments d'ordre deux se déduisent de la fonction de covariance $C(h) = C(s - s')$ et de la moyenne m . Où $s=x$ et $s'=x+h$

En toute généralité, une fonction de covariance est définie pour tous les couples $s; s'$.

Lorsque Z est **stationnaire**, la fonction de covariance se **réduit** à une fonction **d'une seule variable** d'espace, ce que l'on note abusivement $C(s; s') = C(s - s') = C(h)$.

Il arrive que cette fonction C ne dépendent pas non plus de la direction du vecteur h , mais seulement de sa norme : $C(h) = C|h|$. Ceci exprime que la variable **n'est pas structurée selon des directions préférentielles**, on dit que Z et C sont **isotropes**. Dans le cas contraire, on dit que Z et C sont **anisotropes**.

Il se trouve que l'hypothèse de stationnarité constitue souvent en première approximation un bon compromis entre fidélité aux données et simplicité mathématique. Mais bien souvent, cette hypothèse est faite d'emblée sur les données car il est difficile de la tester efficacement. On cherche ensuite le moins mauvais modèle dans ce cadre simplificateur.

REFERENCE BIBLIOGRAPHIQUE :

1. Introduction à la géostatistique, notes de cours de Mr Gilles Guillot, Institut National Agronomique de Paris-Grignon janvier 2004.
2. Guide d'utilisation des méthodes de la géostatistique linéaire, Laboratoire Central de Surveillance de la Qualité de l'Air, INERIS Décembre 2003.
1. Site internet : Wikipédia.