--- Polycopie du cours Géostatistique ---

Master M2 Eau et Environnement

Enseignant : Pr. B. Azouzi

Partie 1

- **Définition**
- **Concept de variable régionalisée**
- **Problèmes d'estimation**
- **Combinaisons linéaires pondérées**
- **Estimation globale et Estimation locale**
- **Procédés d'estimation analytiques (non stochastiques)**
  - **(a) Interpolation linéaire par triangulation**
  - **(b) Interpolation polynomiale**
  - **(c) Interpolation par moindres carrés**
  - **(d) Interpolation par les méthodes de pondération**
    - **d.1-Méthode de THIESSEN**
    - **d.2-Pondération en fonction de la distance**
- **Conclusion sur les méthodes non stochastiques**

Année universitaire 2020/21

Définition

La géostatistique est la science qui traite les données spatiales d'une variable appelée « **variable régionalisée** » en caractérisant la structure spatiale du phénomène naturel (régionalisé) sous une forme mathématique et en estimant, à partie d'un échantillonnage connu, des données dans des points inconnus.

La géostatistique vient, donc, résoudre le problème d'estimation d'une variable régionalisée.

La géostatistique doit son nom aux origines minières de la discipline (Krige, Matheron). Nombre des concepts fondamentaux de la discipline sont issus des travaux de Georges Matheron (tels que : variable régionalisée, fonction aléatoire, hypothèse intrinsèque, effet de pépite…etc (MATHERON et al. 1965).

Concept de variable régionalisée

Les méthodes d'analyse spatiale s'intéressent à l'étude des **phénomènes régionalisés**, c'est-à-dire des phénomènes qui se déploient dans l'espace et y présentent une certaine structure. Par "espace", nous entendons en général l'espace géographique (à 1, 2 ou 3 dimensions), mais il peut aussi s'agir du temps ou d'espaces plus complexes.

L'objet sur lequel nous allons travailler ne sera pas le phénomène régionalisé lui-même, qui est une réalité physique, mais une description mathématique de cette réalité, à savoir une (voire plusieurs) fonction numérique appelée variable régionalisée ou encore **régionalisation**, censée représenter et mesurer correctement ce phénomène.
Par exemple:
• un phénomène géologique tel que l'épaisseur de la couche subhorizontale peut être vu comme la distribution dans l'espace à 2 dimensions de la variable épaisseur ;
• un phénomène 'de minéralisation peut être caractérisé par la distribution dans l'espace à trois dimensions de variables telles que teneur, densité, granulométrie, ...
• la concentration dans le sol d'un polluant, la mesure de la pluie en un point, la production de la récolte d'une parcelle, le taux en carbone mesuré sur un échantillon de terre sont des variables régionalisées.

D'un point de vue mathématique, une variable régionalisée est simplement une fonction traditionnellement notée z, définie en tout point s de l'espace. Cette définition est purement descriptive et ne fait appel à aucune interprétation probabiliste. En général, cette fonction varie très irrégulièrement dans l'espace et échappe à toute représentation fonctionnelle simple mais présente aussi une structure spatiale (zones riches/pauvres).
Un phénomène régionalisé n'ayant jamais une étendue infinie, nous n'étudierons la variable régionalisée z(s) qu'à l'intérieur d'un domaine borné Œ, appelé champ de la variable. Ce champ Œ peut représenter une zone naturelle en dehors de laquelle z n'est pas définie; il peut aussi s'agir d'un domaine particulier où la variable régionalisée présente un intérêt, par exemple les endroits où z est non nulle, ou supérieure à un certain seuil.

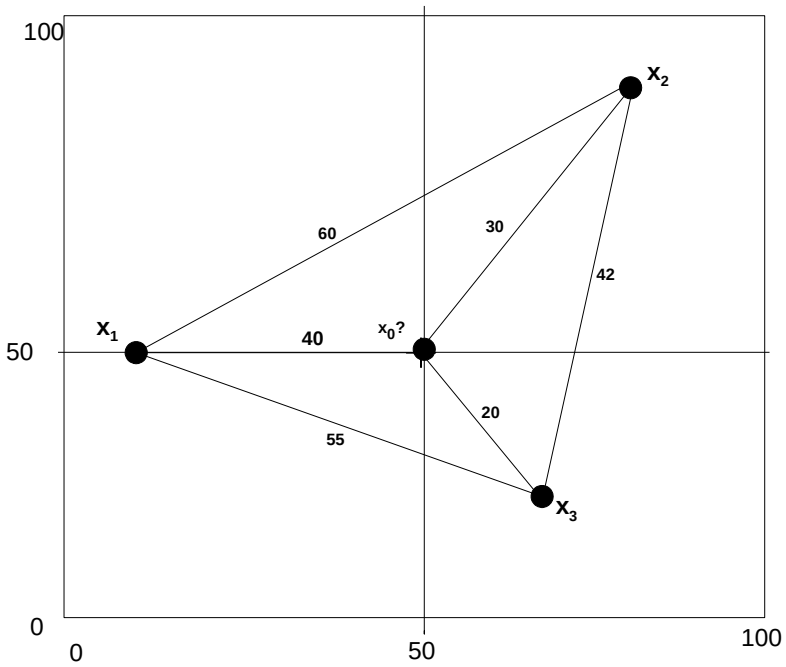
Problèmes d'estimation

L'estimation des variables régionalisées est parmi l'un des objectifs importants de l'analyse des données spatialisées. A partir des valeurs recueillies sur des observations localisées dans une zone géographique, il s'agit d'estimer soit la valeur en un site non échantillonné (interpolation), soit encore la valeur d'un bloc (surface ou volume), soit la valeur globale de la zone entière. Dans certains cas, on essayera d'estimer seulement la moyenne; dans d'autres cas, il faudra également estimer divers paramètres caractérisant la distribution des valeurs observées.

Les méthodes d'estimation locale que nous allons examiner, appliquées souvent dans les Systèmes d'Informations Géographiques (SIG), permettront de cartographier les valeurs d'une variable régionalisée, soit sous forme d'isovaleurs (lignes de niveaux à valeur constante pour la variable), soit sous forme de blocs diagrammes en 3 dimensions (vue en perspective de la surface dont la dimension verticale est représentée par la variable mesurée). Certaines techniques comme la géostatistique permettront, en outre, d'analyser la structure de la variable mesurée, la dépendance et la régularité spatiale, les changements d'échelle, les phénomènes d'anisotropie, ...

Pour fixer les idées, prenons l'exemple suivant :

On veut estimer la variable inconnue Z_0 au niveau du point X_0 en utilisant les autres variables Z_1, Z_2 et Z_3 mesurées aux points X_1, X_2 et X_3 comme indiqué sur le schéma. On donne Z_1=4, Z_2= 6 , Z_3 = 5



Combinaisons linéaires pondérées

La plupart du temps, la valeur estimée sera une combinaison linéaire pondérée des valeurs observées sur les sites échantillonnés. On peut donner comme exemple connu d'estimateur la moyenne arithmétique, $z = 1/n \, \Sigma z(si)$, qui est une combinaison linéaire donnant des poids égaux à tous les points d'observation. Mais, cette estimation sera vite abandonnée dans le cadre spatial. En effet, lorsque les observations sont réparties dans une zone géographique D,

la moyenne arithmétique est un estimateur médiocre: elle ne tient pas compte de la disposition des sites dans la zone D, des secteurs sous échantillonnés ou sur échantillonnés, des zones de valeurs riches ou pauvres pour la variable z.

Utilisées aussi bien pour des raisons de commodité de calcul que de bon sens, les combinaisons linéaires pondérées sont de la forme :

$$z = \Sigma \lambda i z(s_i)$$

où **λi** est le poids affecté à $z(s_i)$, valeur observée sur le sites s_i.

On impose, en général, $\Sigma \lambda i = 1$, de sorte que, dans le cas où les données sont toutes égales à une constante, la valeur estimée redonnera cette constante.

Les différentes méthodes d'estimation se distingueront, en général, par les divers choix ou modes de calcul des poids **λi**. Certaines reposent sur des notions de bon sens, d'autres sur des théories faisant intervenir la statistique. On verra que ces deux approches ne sont pas incompatibles et que souvent elles se rejoignent.

Estimation globale et Estimation locale

L'estimation globale concerne le champ entier D, que l'on désire caractériser par une valeur unique. L'estimation globale obtenue en calculant la moyenne des valeurs des observations est assez bien représentative de la valeur réelle si les sites sont localisés sur une grille régulière ou répartis au hasard. Malheureusement, dans la pratique, cette situation n'est pas toujours réalisée.
On peut avoir plus d'observations dans certains secteurs que dans d'autres et la moyenne peut alors ne pas être représentative de la zone entière. Les méthodes globales devront donc tenir compte des zones où les observations sont plus denses. Les observations issues de ces zones devront avoir "moins de poids" que celles provenant de zones de plus faible densité.
Cependant, il est rare qu'une estimation globale soit suffisante. Il est souvent nécessaire de la compléter par des estimations locales. Par exemple, dans une étude de pollution il ne suffit pas de connaître la pollution moyenne de l'ensemble de la zone, mais il faut aussi distinguer les secteurs fortement pollués de ceux qui le sont moins.

Les **estimations locales** s'intéressent, au contraire, aux différents secteurs de la zone. Elles doivent tenir compte de la distance entre le secteur à estimer et les sites d'observation. Les sites proches du secteur à estimer auront intuitivement plus de poids que les sites éloignés. Les estimations locales sont aussi sensibles aux zones où les observations sont plus denses. Des observations proches ont souvent des valeurs similaires et contiennent donc une information redondante.
En résumé, dans l'estimation locale, les poids affectés aux observations doivent tenir compte à la fois de la distance entre le secteur à estimer et les points échantillonnés et de la possible redondance entre les valeurs observées due à leur proximité.

A partir d'un échantillonnage discret d'une variable spatialement repartie, les procédés d'éstimation interviennent en utilisant les données disponibles , et permettent alors d'évaluer la valeur de la variable étudiée en un lieu non mesuré.
Ces procédés d'estimation peuvent être **analytiques** ou probabilistes (stochastiques).

- les méthodes **non** stochastiques qui sont **des procédés d'estimation analytique**s qui reposent sur des propriétés mathématiques déterministes ;

- les méthodes **stochastiques** qui font appel à des modèles **probabilistes.**

Procédés d'éstimation analytiques :

A. Interpolation lineaire par triangulation :

c'est une pratique manuelle d'interpolation linéaire qui consiste à former un réseau de triangles dont les sommets sont representés par les points expérimentaux les plus voisins possibles retenus. Le lissage des courbes isovaleurs se fait entre les points des sommets à travers les côtés de ces triangles.

L'interpolation linéaire par triangulation est trop fastidieuse à exécuter; et comporte un grand risque d'erreur.

B. Interpolation polynomiale :

On essaie dans ce cas d'ajuster, une fonction polynomiale (ou trigonométrique) aux données experimentaux X et Y (mesurés) supposées être assez reguliers,

$$Z(Xi) = f (Xi , Yi)$$

Le modèle doit être éxact de tel sorte que la surface d'ajustement doit passer par les données d'echantillonnage.
L'utilisation de telle méthode est trop difficile, car elle suppose que les données sont bien réparties avec une structure regulière, ce qui n'est pas toujours le cas.

C. Interpolation par moindre carrées :

Elle se pratique dans le cas d'une structure chaotique, en faisant l'hypothèse que le phénomère réel est en fait regulier, mais qu'il est affecté d'une erreure $\varepsilon(x)$ de moyenne nulle et sans auto-corrélation spatiale:

on a

$$Z(x) = m(x) + \varepsilon(x)$$

$m(x)$ peut prendre un modèle généralement polynomial .

La méthode des moindres carrés consiste à minimiser la somme des carrés des écarts entre valeur observée et valeur estimée (théorique) par le modèle choisi:

$$\Sigma\,(Z^{*}(x) - Z(x))^{2} \quad \text{minimum}$$

sans pour autant obliger la surface d'ajustement à passer par les points expérimentaux.

Cette approche lisse le résultat et entraine une perte de détails. Une variation locale si elle éxiste peut affecter l'ajustement globale

D. Interpolation par les méthodes de pondération:

On calcule ici la valeur d'une variable spatiale Zi par la moyenne pondérée des autres valeurs échantillonnées.

Parmi ces méthodes on a :

- Méthode des polygones d'influence de **THIESSEN.**
- Méthode de pondération en fonction de la distance.

D-1 Méthode de THIESSEN :

C'est une méthode géométrique assez arbitraire conçue pour être employé à grande échelle dans les régions sans influence orographique trés marqués .

Elle consiste à déterminer un certain nombre de polygônes formés à partir des médiatrices des droites reliant les points adjacents (REMINIERAS, 86). La valeur expérimentale Zi est supposée être la moyenne representative du polygône **Si** :

$$\bar{Z}i = 1/S \sum Si.Zi$$

Si = est la surface du polygône élémentaire

S = surface totale du domaine étudié.

Le pourcentage de **Si** par rapport à **S** ,sert de coeficients de pondération propre à chaque valeur Zi.

La méthode des polygône d'influence présente l'inconvénuent d'être d'une application laborieuse.

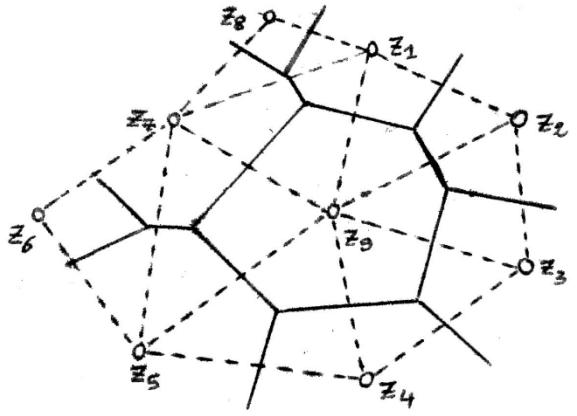


Schéma explicatif de la méthode de THIESSEN (IN G. REMINIERAS; 1986)

D-2 Pondération en fonction de la distance :

Dans ce cas la valeur de Zo non échantillonnée peut être estimé en fonction de la distance d qui sépare Zo et les points experimentaux Zi qui se trouvent dans son proche entourage par le biais de la formule ou modèle suivant:

$$Zo = \frac{\Sigma\, g(di).Zi}{\Sigma\, g(di)}$$

Avec :

g(di) : fonction de pondération:

di : distance séparant le point Zo à estime du point experentale Zi.

On peut citer parmi les fonction de pondération les plus connues:

- $g(d) = 1 / d$ interpolation par inverse des distance

- $g(d) = 1 / d^2$ interpolation par inverse des carrés des distance.

Si on revient à notre exemple ci-dessus , on peut utiliser l'interpolation par inverse des carrés des distance dont la solution est résumée dans le tableau ci-dessous.

On veut estimer la variable inconnue Z_0 au niveau du point X_0 en utilisant les autres variables Z_1, Z_2 et Z_3 mesurées aux points X_1, X_2 et X_3 comme indiqué sur le schéma. On donne Z_1=4, Z_2= 6 , Z_3 = 5

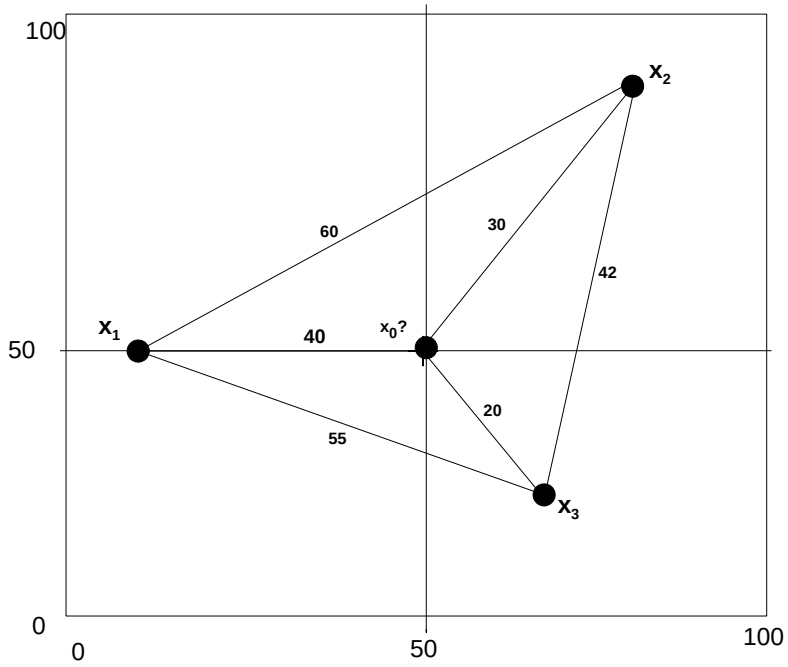


Tableau résumant la solution par 'inverse des carrés des distances :

Z	d	d au carré	1/dcarré	Zi * 1/dcarré
4	40	1600	0.000625	**0.0025**
6	30	900	0.001111	**0.00666**
5	20	400	0.0025	**0.0125**
		Somme =	0.004236	**0.02166**
			Zo =0.0216/0.00423	**8.6666**
				Donc Zo = 8.66

Conclusion sur les méthodes analytiques :

Les procédés analytiques d'éstimation de part la lourdeur de leur mise en oeuvre, présentent deux inconvénients majeurs ; à savoir:

-Ils ne donnent pas la variance d'estimation pour déterminer l'intervalle de confiance pour chaque valeur estimée.

-Ils ne tiennent pas compte de la structure spatiale du phénomène, elles s'appliquent en principe, aussi bien à un phénomene stationnaire que non stationnaire,érratique que continue (DELHOMME, 76). Or , la plupart des paramètres notamment, les grandeurs hydrogéologiques (H,T,..etc) présentent une certaine structure par un processus de dépendance ou auto-correlation régionnalisée. Face à "l'handicap" des méthodes d'éstimation analytiques, le recours aux méthodes probabilistes par procédés géostatistiques devient inévitable.