CHPITRE 2

Chapitre II : Introduction à la géostatistique

Objectifs

Ce cours propose une initiation à quelques concepts et méthodes de la géostatistique, qui regroupe dans un formalisme de fonctions aléatoires des méthodes pour l'estimation de phénomènes déployés dans l'espace géographique. Cette estimation est faite sur la base d'observations, qui peuvent inclure des prévisions effectuées avec des modèles numériques.

II.1 Définition et historique

Le mot "géostatistique" est formé à partir de la combinaison de deux racines grecques: "geo" qui signifie "terre" et "statistique" qui se rapporte à la théorie des probabilités et à l'analyse statistique. La géostatistique est donc une branche de la statistique qui s'applique spécifiquement à l'analyse des données spatiales et géographiques. Elle est utilisée pour modéliser et estimer des variables géographiques inconnues à partir d'observations partielles ou dispersées dans l'espace.

Introduit dans les années 1950 par Georges Matheron, le terme "géostatistique" désigne l'application de la statistique aux problèmes de géologie. Matheron, qui était un ingénieur des mines et un géologue français, est considéré comme le fondateur de la géostatistique moderne.

Dans les années 1960 et 1970, Matheron a développé la théorie de la géostatistique en travaillant sur des problèmes de modélisation de dépôts minéraux. Il a développé des méthodes pour estimer la qualité et la quantité de minerais dans les gisements en utilisant des données spatiales, ce qui a eu un impact significatif sur l'industrie minière.

Depuis lors, la géostatistique a été appliquée à de nombreuses autres disciplines, notamment la géologie, la météorologie, l'agronomie, l'environnement et l'ingénierie. Ainsi, la géostatistique est devenue prometteur de plus en plus, car elle reste une méthode essentielle pour l'analyse de données spatiales et géographiques. La géostatistique continue de connaitre des développements et des tendances actuelles importantes :

1. Intégration de la géostatistique avec l'apprentissage automatique : l'apprentissage automatique est de plus en plus utilisé pour l'analyse de données géospatiales, et la géostatistique peut fournir des méthodes complémentaires pour la modélisation des structures spatiales.

2. Expansion de la géostatistique en dehors des sciences de la terre : la géostatistique a été initialement développée pour la modélisation des gisements minéraux, mais elle a depuis été appliquée à d'autres domaines tels que la climatologie, l'écologie et la santé publique.
3. Utilisation croissante de données massives et de données en temps réel : la géostatistique doit être adaptée pour gérer les défis liés à la gestion et à l'analyse de données de grande taille et de données qui sont collectées en temps réel.
4. Développement de nouvelles méthodes pour la modélisation de la variabilité spatiale : la géostatistique continue d'évoluer pour répondre aux besoins de modélisation de la variabilité spatiale, en utilisant de nouvelles méthodes telles que la simulation stochastique.

La géostatistique continuera d'être une méthode importante pour l'analyse et la modélisation de données spatiales et géographiques dans de nombreux domaines scientifiques et techniques.

II.2 Notions de variables régionalisées (géostatistique)

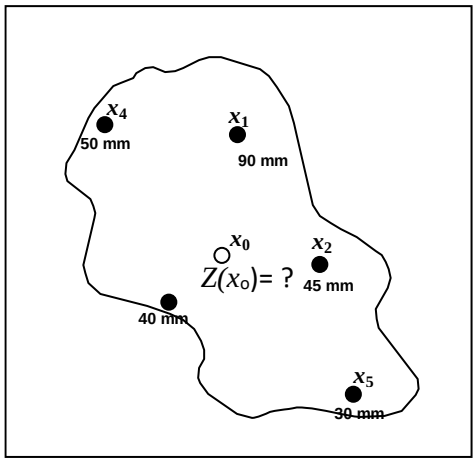
La géostatistique appelée également théorie des variables régionalisées, car elle s'intéresse à l'analyse statistique des variables qui varient de manière spatiale, c'est-à-dire à travers une région. Elle permet notamment d'analyser la structure spatiale de la variation des variables, en prenant en compte leur dépendance spatiale.

La théorie des variables régionalisées repose sur l'hypothèse que les variables spatiales présentent une structure de dépendance spatiale, c'est-à-dire que la valeur d'une variable en un point de l'espace est influencée par les valeurs de cette même variable observées dans les points voisins. Cette dépendance spatiale est modélisée à travers des fonctions de covariance ou de corrélation spatiale, qui mesurent la similitude entre les valeurs de la variable à différentes distances spatiales.

La géostatistique est l'application du formalisme des fonctions aléatoires à la reconnaissance et à l'estimation des phénomènes naturels. Une fonction aléatoire Z(x) est un ensemble de variable aléatoire Z(xi), définie en chaque point xi du domaine D (gisement par exemple) Matheron (1971) :

Z(x) = {Z(xi), $\forall$ **xi** $\in$**D}**

Un phénomène naturel peut être caractérisé par la répartition dans l'espace d'un certain nombre de grandeurs mesurables, que nous appelons " **variables régionalisées**".



Ces variables ont une structure d'auto-corrélation qui dépend du module et de la direction du vecteur séparant deux points de mesure. Mathématiquement, une variable régionalisée est une fonction du point x. Cette fonction est généralement irrégulière et montre deux aspects complémentaires (Matheron, 1973) :

- Un aspect aléatoire qui explique les irrégularités locales ;
- Un aspect structuré qui reflète les tendances du phénomène à grande échelle.

Si au point x_i de l'espace, la variable régionalisée $Z(x_i)$ est considérée comme valeur unique (valeur vraie) alors la géostatistique étudiera la corrélation spatiale de la variable régionalisée et la structure de cette variable dans l'espace. C'est **la géostatistique transitive** (Chauvet, 1999).
Le choix constitutif, de la géostatistique minière consiste à interpréter chaque valeur de la variable régionalisée Z(x), comme une réalisation particulière d'une variable aléatoire $Z(x_i)$ implanté au point x_i, donc plusieurs réalisations sont possibles. C'est **la géostatistique intrinsèque.**

II.3 L'interpolation spatiale

Les phénomènes naturels, tels que les précipitations, la température, topographie, couches géologiques, teneur en éléments chimiques d'un gisement minier, faciès géologiques d'un réservoir pétrolier, qualité de l'eau ou de l'air, la concentration en polluants, ou d'un site pollué, etc., peuvent varier de manière significative dans l'espace. Ces phénomènes se déployant dans l'espace et dans le temps se présentent à l'observateur sous forme régionalisée où ils manifestent une certaine structure spatiale. Ils sont généralement caractérisés localement par des informations géoréférencées. Ces dernières, sont à l'origine de nature continue dans l'espace géographique tridimensionnel et il est évidemment impossible de mesurer leurs valeurs en tous points de l'espace qu'elles occupent, mais uniquement au niveau des endroits types d'une manière irrégulière et selon une technique d'échantillonnage appropriée.

Dans le but de mieux comprendre et modéliser ces phénomènes, il très important de connaitre les valeurs prises par ces phénomènes observés en d'autres points de l'espace. En effet, il est souvent nécessaire de réaliser des interpolations spatiales pour estimer les valeurs de ces phénomènes à des emplacements non observés. Il s'agit donc d'une procédure consistant à estimer la valeur d'une grandeur en un site à partir d'échantillons de cette grandeur récoltés dans d'autres sites.

II.3.1 Introduction

L'interpolation spatiale consiste à estimer la valeur d'une variable à des emplacements spatiaux non observés, à partir des valeurs observées à des emplacements connus. Cette technique est largement utilisée en géostatistique pour modéliser les phénomènes naturels qui varient dans l'espace.

L'interpolation spatiale est utilisée dans de nombreux domaines, tels que la météorologie, la qualité de l'air, l'hydrologie, la géologie, l'agriculture, etc. Elle permet de mieux comprendre la variabilité spatiale des phénomènes naturels, de prédire leur évolution future, et de prendre des décisions éclairées en matière de gestion des ressources naturelles et de l'environnement.

L'intérêt de l'interpolation spatiale réside dans le fait qu'elle permet de mieux comprendre la variabilité spatiale des phénomènes naturels, de prédire leur évolution future, et de prendre des décisions éclairées en matière de gestion des ressources naturelles et de

l'environnement. Les objectifs de l'interpolation spatiale peuvent varier en fonction du contexte de l'étude, mais peuvent inclure :

1. La création de cartes spatiales : l'interpolation spatiale peut être utilisée pour créer des cartes spatiales représentant la distribution de la variable étudiée dans une région donnée. Ces cartes peuvent être utilisées pour mieux comprendre la variabilité spatiale des phénomènes naturels et pour planifier des actions de gestion.

2. La prédiction de valeurs manquantes : l'interpolation spatiale peut être utilisée pour prédire les valeurs manquantes de la variable à partir des données observées. Cette technique est utile lorsque les données manquantes sont rares ou lorsqu'il est difficile d'obtenir des données supplémentaires.

3. La création de surfaces continues : l'interpolation spatiale peut être utilisée pour créer des surfaces continues représentant la distribution de la variable étudiée dans l'espace. Ces surfaces peuvent être utilisées pour estimer les volumes et les masses de la variable, ainsi que pour déterminer les zones où la variable dépasse certains seuils critiques.

4. La comparaison de différentes méthodes d'interpolation : l'interpolation spatiale peut être utilisée pour comparer différentes méthodes et déterminer celle qui convient le mieux aux données et aux objectifs de l'étude.

En somme, les méthodes d'interpolation spatiale sont des outils importants pour comprendre les phénomènes naturels qui varient dans l'espace et pour prendre des décisions éclairées en matière de gestion des ressources naturelles et de l'environnement

Il existe plusieurs méthodes d'interpolation spatiale utilisées pour la prévision d'une valeur inconnue à partir d'observations et de données géoréférencées. On parle alors d'interpolation pour l'estimation de cette valeur. Selon les modèles mathématiques sur lesquels elles reposent, les méthodes de prévision peuvent être classées en deux groupes :

- **Méthodes déterministes :** se sont des méthodes qui se basent sur des propriétés purement mathématiques, généralement géométriques, sans tenir compte du phénomène physique considéré.
- **Méthodes stochastiques :** ces méthodes supposent une modélisation probabiliste du phénomène, dont les observations résultent de la réalisation de variables aléatoires; ces méthodes font alors appel à des modèles découlant de l'analyse statistique des données considérées. On parle alors de techniques géostatistiques bas é sur la théorie des variables régionalisées.

II.3.2 Méthodes d'interpolation déterministes

Afin de mettre en évidence l'intérêt de l'interpolation spatiale, nous allons prendre comme exemple certains problèmes rencontrés généralement lors de l'estimation ou hydrologique, météorologique, géologique, topographiques…etc :

- Le passage des mesures ponctuelles des facteurs climatique (précipitation, température…) à une estimation spatiale de celles-ci, souvent nécessaire en hydrologie et délicat.

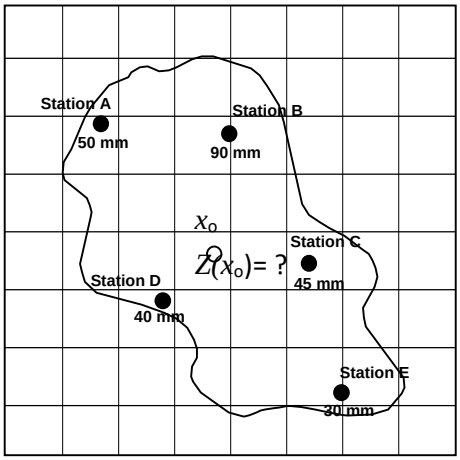
- Estimation d'une cote piézométrique dans un aquifère, souvent nécessaire pour déterminer le sens d'écoulement et les paramètres hydrodynamiques d'un aquifère.
- Détermination d'une cote topographique d'un point à l'aide d'une carte topographique.

Les méthodes les plus couramment utilisées sont les méthodes de calcul de moyennes ou les méthodes d'interpolation des données collectées localement. Ces méthodes permettent notamment le calcul de la variable considérée à l'échelle globale (moyenne à l'échelle de l'espace considéré par exemple le bassin versant), ou à l'échelle locale (à l'échelle d'un point ou d'un site bien déterminé).

Nous nous intéressons ici aux méthodes déterministes pour l'estimation locale et ponctuelle d'une valeur de la variable régionalisée. Cette estimation sera réalisée à partir de combinaisons linéaires des observations en tenant compte de leur disposition les unes par rapport aux autres mais aussi de la distance entre le secteur à estimer et les points de données.

Soit par exemple une variable Z() connue en cinq (05) point (valeurs de précipitations connues en 05 stations pluviométrique), et les hydrologues souhaitent connaître la précipitation au point x_0.

Il existe plusieurs méthodes qui sont utilisé pour répondre à cette question.



1. Méthode des polygones de Thiessen

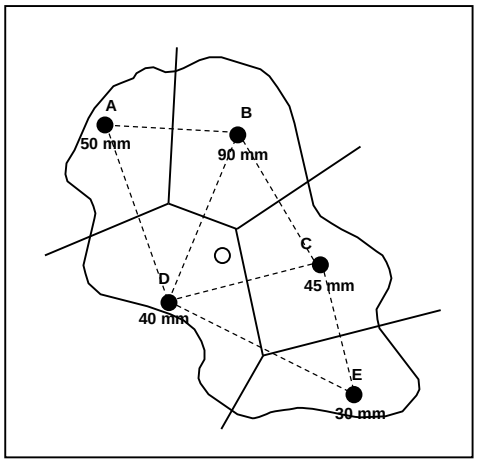
La méthode de Thiessen consiste à définit pour chaque site d'observation du champ, un polygone d'influence tel que chaque point du polygone est plus proche du point d'observation que de tout autre site :

$$\forall\, \vec{x}_0 \in P_i, \forall\, \vec{x}_j \in D \setminus P_i, \|\vec{x}_i - \vec{x}_0\| \leq \left\|\vec{x}_j - \vec{x}_0\right\|$$

D est alors partitionné en un ensemble de polygones convexes, nommés polygone de Thiessen (appelés aussi polygones de Voronoi ou cellules de Dirichlet).

Le découpage des polygones de Thiessen dépend uniquement de la configuration géométrique et non pas des valeurs observées. Les polygones ne sont pas nécessairement fermés dans certaines directions de l'espace : il faut ainsi limiter la partition aux frontières de D, ou fixer une distance d'influence limite.

L'interpolation par la méthode de Thiessen consiste à affecter à l'ensemble des points d'un polygone donné la valeur de la variable régionalisée correspondante (on parle aussi de plus proche voisin).



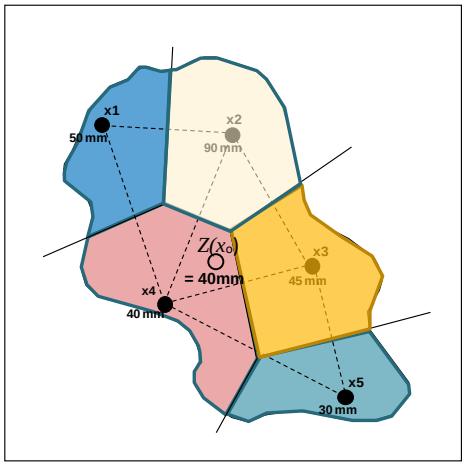
Construction des polygones de Thiessen

Thiessen admet qu'en un point quelconque différent des xi, la valeur de Z est : $\hat{Z}_{(\vec{x}_o)} = Z_{(\vec{x}_k)}$

(xk étant parmi tous les xi, celui qui est le plus proche de xo).

Parfois, on a besoin d'une estimation globale sur le domaine de travail entier ou une partie de ce domaine, que l'on désire caractériser par une valeur unique. La valeur globale est obtenue sous forme de moyenne pondérée par la surface :

$$\overline{P} = \frac{\sum_{i=1}^{n} P_i . s_i}{\sum_{i=1}^{n} s_i}$$



2. Interpolation à partir d'une triangulation (Méthode des facettes)

La triangulation consiste à diviser le champ en triangles disjoints dont les sommets sont les sites d'observation. On calcule alors la valeur en un point donné à partir des valeurs des sommets du triangle auquel il appartient.

Il existe plusieurs méthodes de triangulation, la plus utilisée étant la triangulation de Delaunay : les sommets de chaque triangle sont les sites du champ D tels que les polygones de Thiessen associés ont un côté en commun. Notons les propriétés d'une telle triangulation :

- La triangulation est indépendante de l'ordre de traitement.
- L'ensemble du domaine n'est pas recouvert : on opère uniquement dans l'enveloppe convexe des sites.
- Les cercles circonscrits à chacun des triangles ne contiennent pas d'autre site d'observation.
- Tous les points sont reliés à leur plus proche voisin.

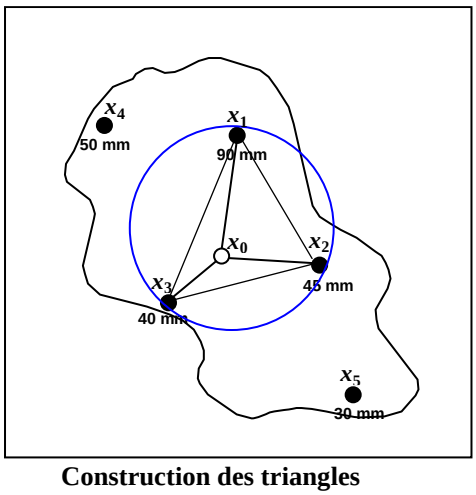
Il existe plusieurs méthodes pour l'interpolation de données à partir d'une triangulation. Ici, deux d'entre elles seront exposées.

On considère le triangle $(\vec{x}_1\vec{x}_2\vec{x}_3)$ contenant le point d'intérêt xo. La valeur recherchée de la variable régionalisée s'écrit sous la forme :

$$\hat{Z}_{(\vec{x}_o)} = \hat{Z}_{(x,y)} = \alpha . x + \beta . y + \gamma$$

La solution est déterminée à partir d'une combinaison linéaire des valeurs observées au sommet du triangle en résolvant le système :

$$\begin{cases} \alpha . x_1 + \beta . y_1 + \gamma = z_1 \\ \alpha . x_2 + \beta . y_2 + \gamma = z_2 \\ \alpha . x_3 + \beta . y_3 + \gamma = z_3 \end{cases}$$



Construction des triangles

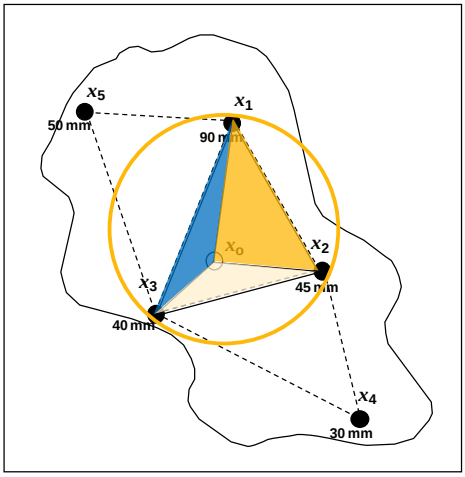
Où $z_1, z_2\ et\ z_3$ sont les valeurs observées aux sommets du triangle $(\vec{x}_1\vec{x}_2\vec{x}_3)$.

L'équation matricielle s'écrit sous la forme :

$$\begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix} . \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

Equation qu'il suffit d'inverser pour parvenir à la solution.

On peut également montrer que cette solution s'écrit sous la forme :



$$\hat{z}_{(x_0)} = \frac{|\vec{x}_1\vec{x}_0\vec{x}_2| . z_3 + |\vec{x}_1\vec{x}_0\vec{x}_3| . z_2 + |\vec{x}_2\vec{x}_0\vec{x}_3| . z_1}{|\vec{x}_1\vec{x}_2\vec{x}_3|}$$

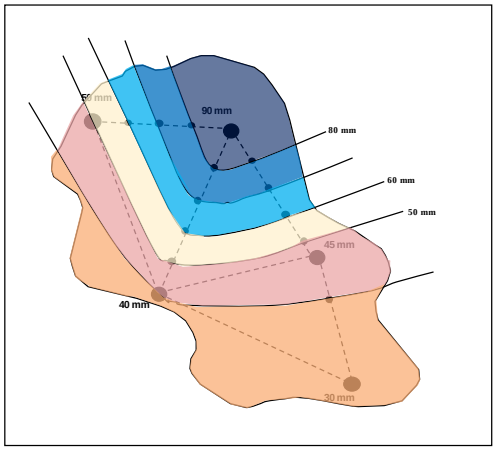
Où $|\vec{x}_1\vec{x}_2\vec{x}_3|$ représente l'aire du triangle formé par les points observés $(x_1x_2x_3)$: la solution devient donc la somme pondérée des aires des triangles formés par les 3 sommets et le point d'intérêt.

En effet, plus le point auquel on souhaite estimer la valeur de la variable régionalisée est proche d'un site d'observation, plus la proportion de surface occupée par le triangle opposé à ce sommet est importante ; le poids associé est donc plus important et la valeur estimée proche de celle du sommet.

Chaque site d'observation reçoit donc un poids égal à la proportion de surface occupée par le triangle qui lui est opposé : plus le point recherché est proche d'un site d'observation plus la valeur de la variable régionalisée est proche de la valeur observée en ce site.

3. Interpolation à partir d'une triangulation (Triangulation linéaire)

La triangulation linéaire est aussi utilisée pour l'interpolation certaines variables, afin d'obtenir une estimation globale sur le domaine de travail entier ou une partie de ce domaine.
Des courbes d'égale valeurs sont tracées (Courbes hydro-isohypses, courbes isohyètes, courbes isoteneurs…), le principe est de divisé les cotes de chaque triangle en segment proportionnels on adoptant une équidistance bien déterminer. Ensuite, les points d'égales valeurs sont raccordés à l'aide des courbes.
La valeur globale est obtenue sous forme de moyenne pondérée par la surface :



Méthode de Triangulation linéaire

$$P = \frac{\sum_{i=1}^{n} P_i \cdot s_i}{\sum_{i=1}^{n} s_i}$$

Ou P_i représente la valeur moyenne pour deux courbes successives de précipitations et s_i représente les surfaces des bandes limitées entre deux courbes successives.

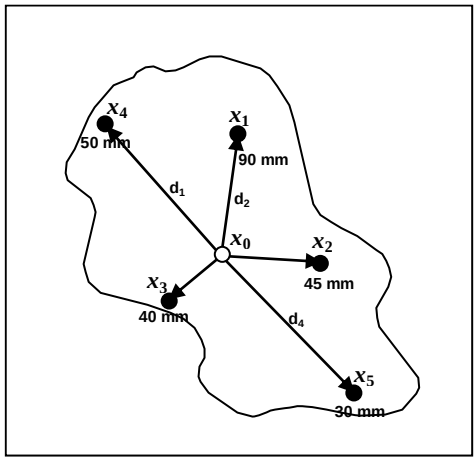
4. Méthode barycentrique (Inverse de la distance pondérée IDW)

Cette méthode consiste à estimer les valeurs d'une variable continue à des emplacements où aucune donnée n'est disponible en utilisant une combinaison linéaire pondérée des valeurs connues des points de données voisins.

Les poids sont calculés en fonction de la distance ($\boldsymbol{d_i}$) entre le point à interpoler et les points de données connus, de sorte que les points les plus proches ont une pondération plus élevée que les points plus éloignés.

Elle consiste à attribuer un poids inversement proportionnel à la distance entres les sites et le point à estimer : $\hat{\boldsymbol{z}}_{(xo)} = \frac{\sum_{i=1}^{n} \frac{(z_i)}{d_i^p}}{\sum_{i=1}^{n} \frac{1}{d_i^p}}$

Une valeur de pondération fixe ($\boldsymbol{p = 2}$), est une valeur qui donnent généralement des résultats très représentatifs de la variable estimée.



Determination des distances aux pointx xi

Cette méthode d'interpolation est exacte et fournit une surface continue. N'étant pas limitées au voisinage direct du point d'intérêt, cette méthode présente l'avantage de prendre en compte plus de données du champ d'observation. Un poids plus important est affecté aux sites les plus proches, un poids moindre aux sites plus éloignés.

Les valeurs interpolées sont limitées par les valeurs minimales et maximales du champ d'observation, la pondération étant positive. Dans le cas d'un jeu de données conséquent, il est conseillé de se limiter aux points les plus proches, le temps de calcul pouvant augmenter significativement.

Elle présente cependant quelques limites. Elle est indifférente à la configuration géométrique des observations, seule la distance compte. Elle tend à surpondérer les données groupées alors qu'elles sont redondantes.

II.4 Conclusion

En somme, les différentes techniques déterministes permettant l'estimation locale de variables régionalisées en des sites non échantillonnés. Il est en fait impossible de dire quelle méthode fournit globalement les meilleurs résultats. Aucune méthode déterministe ne parait universellement meilleure. Mais, une limite importante des méthodes présentées ici:

- L'ensemble de ces méthodes s'appliquent aveuglément sans tenir compte d'une éventuelle structure spatiale de la variable régionalisée étudiée.
- La surface obtenue peut certes être esthétique, mais pas nécessairement précise.
- L'utilisateur de ces méthodes n'a aucune idée sur la l'incertitude de les résultats d'interpolation.

De ce fait, l'estimation d'une propriété dans l'espace géographique suppose deux étapes:

- Une phase d'analyse de la structure spatiale de la propriété étudiée, pour savoir comment sont corrélés entre eux les points observés (problème d'hypothèse de linéarité).
- Une phase d'estimation proprement dite tenant compte de la structure spatiale précédemment identifiée (utilisation de la fonction de structuration ainsi définie pour l'estimation de la variable).