Paris School of Business (PSB)

Interpolation des données spatiales

MSc Data Management

Projet : **R**

par :

Arnaud Bruel YANKO, Adrien JUPYTER

Sous la direction de :

M. Henri LAUDE

Enseignant

Année académique 2020-2022

♣ Table des matières ♣

1 Introduction

Les techniques d'interpolation spatiale peuvent être séparées en deux principales catégories : les approches déterministes et géostatistiques. Pour faire simple, les méthodes déterministes n'essayent pas de capturer la structure spatiale des données. Elles utilisent seulement des équations mathématiques prédéfinies pour prédire des valeurs à des positions où aucun échantillon n'est disponible (en pondérant les valeurs attributaires des échantillons dont la position dans la parcelle est connue). Au contraire, les méthodes géostatistiques cherchent à ajuster un modèle spatial aux données. Cela permet de générer une valeur prédite à des positions non échantillonnées dans la parcelle (comme les méthodes déterministes) et de fournir aux utilisateurs une estimation de la précision de cette prédiction.

Les approches géostatistiques regroupent le krigeage et ses dérivés. Toutes les méthodes qui seront discutées doivent être appliquées sur des variables continues (NDVI, rendement, teneur en carbone du sol) et pas factoriel (ex : une classe issue d'une méthode de classification) ou binaires (variables avec des valeurs de 0 ou 1 ? il existe des méthodes pour ce type de données, notamment des méthodes de krigeage

⋆ ⋆ Chapitre Un ⋆ ⋆

Les méthodes d'estimation déterministes et non stochastiques

1 Estimation globale

La moyenne arithmétique d'un ensemble de valeurs situées dans un espace géographique est souvent un estimateur peu représentatif de la moyenne globale de la zone. En effet, si, par exemple, l'échantillonnage privilégie certains secteurs où les valeurs sont très faibles, elles ne seront pas représentatives de l'ensemble de la zone ainsi que leur valeur moyenne. Pour obtenir une estimation représentative de la moyenne globale il est indispensable de pondérer la valeur des observations de telle manière que celles qui sont proches les unes des autres aient moins d'influence dans l'estimation. Cette méthode qui donne un poids faible aux observations rassemblées en "grappes" a cependant deux inconvénients :

— elle peut donner un poids nul à de l'information très importante car l'information utile est complètement inconnue a priori ;

— il est souvent impossible d'identifier les observations importantes et celles qui le sont moins.

On distingue un nombre important de techniques que nous détaillerons dans la suite.

1.1 La méthode des polygones d'influence de Thiessen

On définit pour chaque site d'observations ; un polygone d'influence qui est déterminé de telle sorte que chaque point du polygone est plus proche du site s_i que de tout autre site. Le polygone d' influence d'un site s_i est obtenu géométriquement en traçant les médiatrices des droites joignant si et les autres sites immédiatement voisins et en prenant le plus petit polygone contenant s_i. La zone géographique σ est alors partitionnée en polygones, appelés *polygones de Thiessen ou de Voronoï ou encore cellules de Dirichlet*. Les sites proches d'autres sites auront des polygones d'influence de petites surfaces alors que les sites éloignés des autres sites auront de plus grandes surfaces.

Autrement dit

La méthode du polygone de Thiessen permet d'estimer des valeurs pondérées en prenant en considération chaque station pluviométrique. Elle affecte à chaque pluviomètre une zone d'influence dont l'aire, exprimée en $\%$, représente le facteur de pondération de la valeur locale.

Les polygones de Thiessen construits, leurs surfaces vont servir à construire 1' estimation globale de la moyenne de la variable régionalisée z sur la zone σ. Celle-ci sera égale à la moyenne des valeurs observées sur chaque site pondérées par leur surface d'influence relative.

$$EG(T) = \sum_{i=1}^{n} \frac{S_i}{S} z(s_i) \tag{1.1}$$

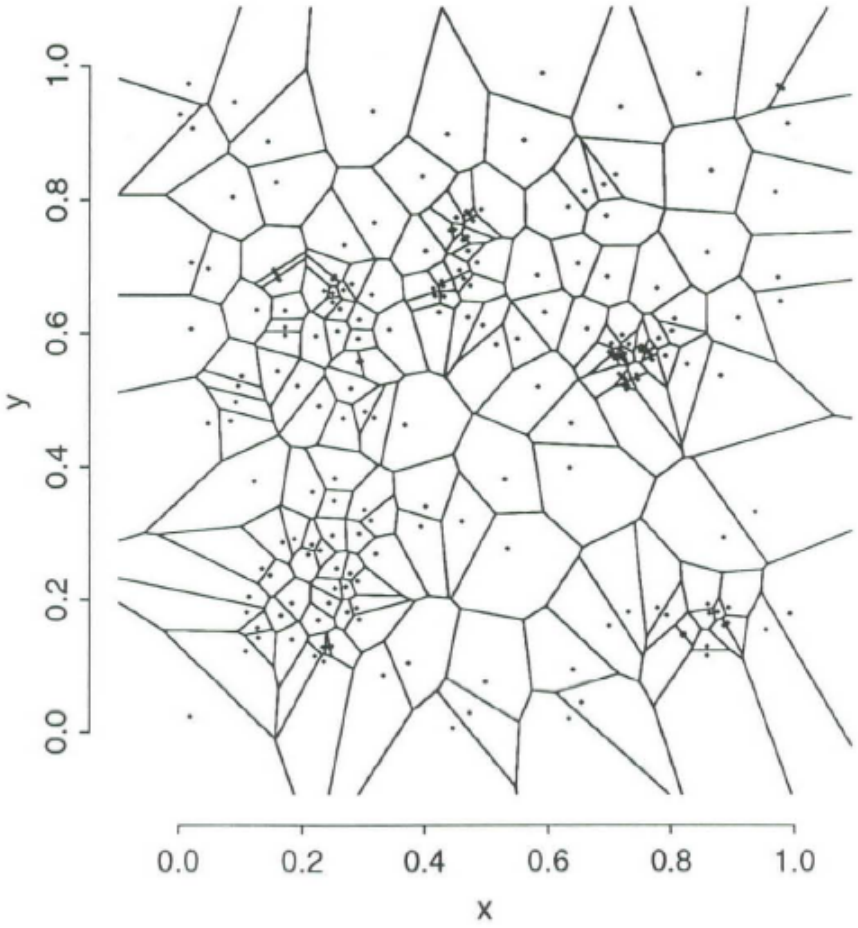
où

— EG(T) :=estimation globale (Thiessen) ;

— S_i est la surface du polygone d'influence du Site s1 ;

— S la surface de la zone entière σ tel que : $S = \sum_{i=1}^{n} S_i$

FIGURE 1.1 – Les polygones de Thiessen de 200 points simulés



la figure ci-dessus nous permet de visualiser les polygones de Thiessen de 200 points dont la position a été simulée à partir de plusieurs lois uniformes, permettant ainsi d'obtenir des amas de points. Les zones à forte concentration en points ont des polygones de Thiessen de surfaces moindres.

1.2 La méthode des cellules

La zone entière σ est divisée en cellules rectangulaires. Chaque cellule contient un nombre variable de sites. L'inverse de ce nombre sert de coefficient de pondération dans le calcul de l'estimation globale.
La procédure se fait donc en 2 étapes :

— dans chaque cellule, on calcule la moyenne des valeurs observées sur les sites de la cellule. On obtient alors la moyenne de la cellule ;

— on calcule ensuite la moyenne des moyennes de toutes les cellules, où chaque cellule a le même poids.

Ce qui nous permet de trouver et ceci de manière évidente l'estimation globale de la zone :

$$EG(C) = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{n_\alpha} \sum_{i_\alpha}^{n_\alpha} z(s_{i_\alpha}) \tag{1.2}$$

où

— EG(C) :=estimation globale (Cellule) ;

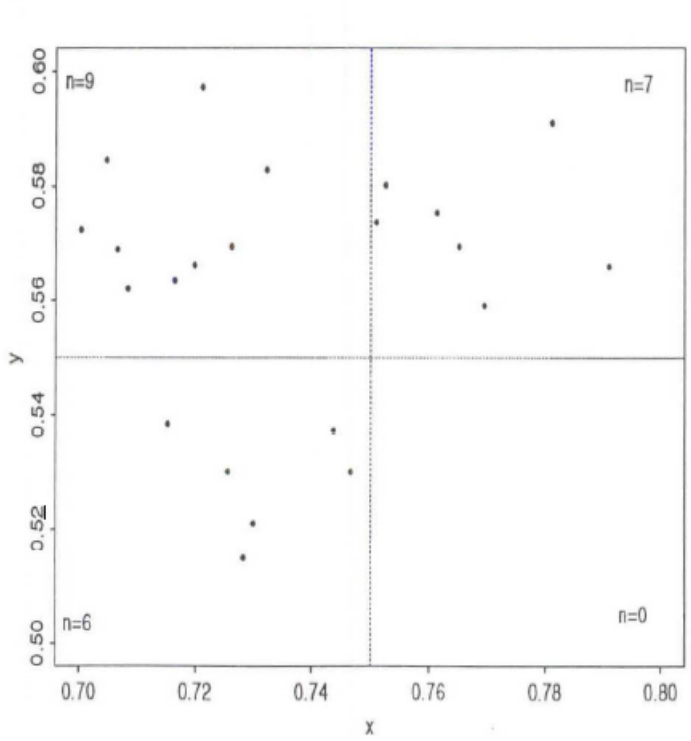
— s_{i_α} sont les sites de la cellule α ;

— N est le nombre de cellules dans la zone σ contenant au moins un site d'observation.

L'estimation obtenue dépendra de la taille de la cellule. Si les cellules sont trop petites, chacune contiendra au plus un point et tous les sites auront le même poids. Si les cellules sont trop grandes, un grand nombre de points appartiendront à la même cellule et chacun d'eux aura encore le même poids.

Pour cela on pourra faire plusieurs essais en faisant varier la taille de la cellule. Sur une sous zone de l'exemple précédent (200 points simulés), on peut voir sur la figure suivante la répartition des sites dans les 4 cellules carrées de côté 0.05. Le nombre de sites dans chaque cellule varie ici de 0 à 9.

FIGURE 1.2 – EG(C) dans une sous zone des 200 points simulés



1.3 La géostatistique transitive

La géostatistique est l'étude des variables régionalisées, à la frontière entre les mathématiques et les sciences de la Terre. Son principal domaine d'utilisation a historiquement été l'estimation des gisements miniers, mais son domaine d'application actuel est beaucoup plus large et tout phénomène spatialisé peut être étudié en utilisant la géostatistique.

La géostatistique transitive s'intéresse essentiellement aux problèmes d'estimation globale. Il s'agit de caractériser l'ensemble du champ σ par une valeur unique représentative du phénomène régionalisé. La première étape consiste à définir un concept approprié pour caractériser la variable régionalisée sur σ. Il faut que ce soit un concept objectif, en ce sens qu'il serait disponible si la réalité était connue partout. Ainsi, la moyenne globale est un concept objectif car défini sans ambiguïté par la donnée d'un domaine et par la valeur de la variable en tous les points de ce domaine.
La géostatistique transitive se propose d'estimer non pas la moyenne globale de la variable régionalisée $z(s)$ sur le champ σ, mais son abondance totale :

$$Q = \int_{\sigma} z(s)\,\mathrm{d}s \tag{1.3}$$

La moyenne se déduit de l'abondance en divisant par la surface ou le volume de σ.

Exemple 1.1.
- z est la proportion d'une culture dans un secteur rectangulaire et Q est la surface totale cultivée ;
- z est la concentration de nitrate dans le sol et Q est la quantité dans la zone entière ;
- z est la teneur en métal dans un gisement minier et Q est la quantité de métal associée.

Par convention,
$z(s) = 0$ qd $\notin \sigma$

Ce qui nous permet d'écrire logiquement,

$$Q = \int_{\sigma} z(s)\,\mathrm{d}s = \int_{espace} z(s)\,\mathrm{d}s \tag{1.4}$$

Notons que le champ désigne, le support (borné) de la fonction z, c'est-à-dire 1' ensemble des points où z est non nulle telle que :

$\sigma = \{s \in \mathbb{R}^d tq z(s) \neq 0\}$

Le problème de 1' estimation globale se formule ainsi : estimer 1' abondance $Q = \int_{\sigma} z(s)$ d'une fonction z à support σ borné (non nécessairement connu), à partir de n sites de mesure

où la valeur de z est donnée.

Le covariogramme transitif

1. Définition et propriétés mathématiques

   Considérons une variable régionalisée z de champ $\sigma \subset \mathbb{R}^d$. On va associer à $z(s)$, fonction qui varie très irrégulièrement dans σ, une fonction plus simple, appelée covariogramme transitif, et notée traditionnellement $g(h)$.

$$g(h) = z(s)z(s+h)\,\mathrm{d}s \tag{1.5}$$

   Le covariogramme transitif a une signification objective : pourvu que l'on connaisse exhaustivement la variable régionalisée z, il est parfaitement calculable dans la totalité du champ.

   mathématiquement on parle de covariogramme transitif si les propriétés suivantes sont vérifiées :

   — $g(h)$ est à support borné, car z est nulle en dehors de σ ;

   — $\forall\, h, g(h) = g(-h)$ (symétrie) ;

   — $\forall\, h, |g(h)| \leq g(0)$ (inégalité de Schwarz) ;

   — $g(h)\,\mathrm{d}s = (z(s)\,\mathrm{d}s)^2 = Q^2$

   — $g(h)$ est de type positif c'est-à-dire qu'il vérifie :

   $\forall k \in \mathbb{N}^*, \forall \lambda_1, \cdot, \lambda_k \in \mathbb{R}, \forall s_1, \cdot, s_k$ ensemble de k sites,

$$\sum_{i=1}^{k}\sum_{j=1}^{k} \lambda_i \lambda_j g(s_i - s_j) \geq 0 \tag{1.6}$$

   Cette dernière propriété, très contraignante, est difficile à vérifier en pratique. C'est la raison pour laquelle on choisira, parmi un certain nombre de fonctions de type positif.

2. Portée

   On appelle portée, dans une direction donnée, la distance au-delà de laquelle le covariogramme est identiquement nul. C'est une quantité finie car le covariogramme transitif est à support borné. La portée dans une direction mesure la plus grande dimension de OE dans cette direction. C'est une quantité purement géométrique, car elle dépend du champ et non de la variable régionalisée elle-même.

3. Comportement à l'origine

4. Isotropie

   Le covariogramme transitif est isotrope s'il est identique dans toutes les directions de l'espace $\mathbb{R}^d$, auquel cas il ne dépend que du module $|h|$ de h. Dans le cas contraire, il y a **anisotropie**. Le cas le plus simple est l'anisotropie géométrique, où une simple transformation linéaire des coordonnées (rotation suivie d'une homothétie) suffit à rétablir l'isotropie : en dilatant ou en contractant les coordonnées dans la direction d'anisotropie, on retrouve les conditions isotropes.

L'estimation de l'abondance

Les sites échantillonnés peuvent avoir plusieurs types de configurations dans l'espace $\mathbb{R}^4$: ils peuvent être régulièrement répartis, disposés selon un échantillonnage aléatoire stratifié ou encore être implantés de manière totalement aléatoire. Dans chaque cas, nous allons examiner comment estimer l'abondance

$$Q = \int_{\sigma} z(s)\,\mathrm{d}s \tag{1.7}$$

— les sites d'observation sont sur une grille régulière

On suppose que les observations sont réparties sur une grille régulière, que l'on prendra rectangulaire pour plus de commodité. La cellule rectangulaire élémentaire, notée $[a, b]$, est définie par les vecteurs orthogonaux a et b. Elle a pour dimension les longueurs $|a|$ et $|b|$.

Les sites échantillonnés sont repérés par deux indices pet q (entiers relatifs), de la manière suivante :

$$s_{p+q} = s_0 + pa + qb, \forall p, q \in \mathbb{Z} \tag{1.8}$$

Définition de l'estimateur

A partir des mesures $\{z(s_0 + pa + qb), p, q \in \mathbb{Z}\}$, l'abondance Q peut être estimée par la combinaison linéaire suivante :

$$\hat{Q}(s_0) = |a||b| \sum_{p,q \in \mathbb{Z}} z(s_0 + pa + qb) \tag{1.9}$$

Le champ σ étant borné et $z(s)$ étant nulle en dehors de σ, la somme ci-dessus ne porte que sur un nombre fini de termes non nuls. En pratique, il suffit donc de s'assurer que le réseau de prélèvements "déborde" le champ de la variable régionalisée z pour pouvoir calculer $\hat{Q}(s_0)$.

Randomisation du réseau d'échantillonnage

En l'absence de toute information sur les données (zones pauvres, zones riches, ...), il n'y a aucune raison objective de commencer l'échantillonnage en un endroit précis : on a positionné d'une manière quelconque l'ensemble du réseau d'échantillonnage $\{s_{p,q}p, q \in \mathbb{Z}\}$ dans l'espace.

Cela revient à dire que l'origine s_0 de ce réseau a été implantée "au hasard" dans une cellule élémentaire de la grille ; la position de celle-ci peut donc être considérée comme une variable aléatoire S_0, uniformément distribuée dans la surface d'une cellule élémentaire $[a, b]$ de la grille. On peut alors considérer :

$$\hat{Q}(s_0) = |a||b| \sum_{p,q\in\mathbb{Z}} z(s_0 + pa + qb) \tag{1.10}$$

également comme une variable aléatoire, et en calculer l'espérance et la variance :

$$E\{\hat{Q}(s_0)\} = \int_{[a,b]} \hat{Q}(s)\,\mathrm{d}s \frac{1}{|a||b|} \tag{1.11}$$

$$= \int_{[a,b]} \sum_{p,q\mathbb{Z}} z(s + pa + qb)\,\mathrm{d}s \tag{1.12}$$

$$= \sum_{p,q\mathbb{Z}} \int_{[a,b]} z(s + pa + qb)\,\mathrm{d}s \tag{1.13}$$

$$= z(s)\,\mathrm{d}s = Q \tag{1.14}$$

Le passage de la ligne de (1.12) à (1.13) est possible par continuité.

L'estimateur global est donc sans biais :

$$E\{\hat{Q}(s_0)\} = Q \tag{1.15}$$

Cela signifie que, pour une variable régionalisée z(s) donnée, les erreurs que l'on ferait en construisant l' estimateur $\hat{Q}(s_0$ pour un grand nombre de grilles, dont l' origine s_0 est tirée au hasard, finiraient par se compenser.

Répartition non aléatoire

La démarche qui vient d'être présentée n'est pertinente que si l'implantation des échantillons est modélisable de façon appropriée par un processus ponctuel connu. Dans le cas d'un échantillonnage non aléatoire, il devient difficile de probabiliser !'estimateur, faute de connaître le mode de construction de l'échantillonnage.

Le bon sens voudrait que l'on cherche une homogénéité de la répartition de l'information, pour ne pas mélanger des informations de qualité différente. Dans le cas d'un échantillonnage irrégulier, on pourra par exemple chercher à diviser le domaine d'étude en sous-zones où la reconnaissance des données est homogène. Une autre alternative consiste à construire un estimateur du type surfaces d'influence. Il est en effet intuitif, pour estimer une abondance à partir d'échantillons répartis de façon irrégulière dans r espace, de chercher à pondérer chaque valeur par son importance relative en surface par rapport au champ total. Dans le cas précédent, cette pondération provenait de la densité du processus ponctuel. Ici, le processus ponctuel est inconnu, et faute de mieux on fait appel à la surface d'influence de chaque échantillon.

aissant les surfaces d' influence ϵ_i des sites s_i, l'abondance totale sera estimée par :

$$\hat{Q} = \sum_{i=1}^{n} \epsilon_i z(s_i) \tag{1.16}$$

L'estimation globale en géostatistique intrinsèque

l'estimation globale consiste-t-elle à adopter une formule, en général très simple, pour caractériser la variable régionalisée sur σ (pour fixer les idées, ce sera le plus souvent la moyenne arithmétique des valeurs mesurées, censée caractériser la moyenne exhaustive sur le champ σ), et à calculer la variance de l'erreur associée à cet estimateur, compte tenu du modèle proposé et du mode d'échantillonnage. Cette variance est une mesure de la précision de l'estimation. L'intérêt de la géostatistique par rapport à la statistique classique est la prise en compte de la structure spatiale de la variable régionalisée et de la position des observations dans le calcul de la variance d'estimation.

-) Notations

Soit V un domaine borné inclus dans le champ σ de la variable régionalisée z. On note $Z(V)$ la variable aléatoire obtenue en faisant la moyenne spatiale de $Z(s)$ sur V :

$Z(V) = \frac{1}{|V|} \int_v Z(s)\,\mathrm{d}s,$

$Z(V)$ est aussi appelée régularisée de la variable ponctuelle $Z(s)$ sur le support V. Dans le cas stationnaire d'ordre deux, $Z(V)$ admet la même espérance que $Z(s)$, et sa variance peut s'écrire :

$C(V,V) = VarZ(V) \frac{1}{|V|^2} \int_v \int_v C(s-t)\,\mathrm{d}s\mathrm{d}t, V \in \sigma$

-) Variance d'extension

On appelle variance d'extension d'un domaine V à un domaine V', notée $\sigma_E^2(V,V')$, la va-

riance de la différence $Z(V) - Z(V')$:
On peut interpréter la différence $Z(V) - Z(V')$ en considérant $Z(V)$ comme une moyenne à estimer et $Z(V')$ comme son estimateur ; $Z(V) - Z(V')$ est alors l'erreur d'estimation, d'espérance nulle (sous réserve de stationnarité d'ordre deux ou intrinsèque) et de variance égale à la variance d'extension $\sigma_E^2(V, V')$.

Le formalisme des variances d'extension est le plus général qui soit pour calculer une variance d'estimation : $Z(V)$ est la quantité à estimer, et $Z(V')$ son estimateur, sans contrainte sur V ni sur V'. Le cas qui nous intéresse dans la suite est celui où le domaine V est le champ OE de la variable régionalisée et le domaine V' est constitué par les n sites échantillonnés ; $Z(V)$ est alors la moyenne exhaustive sur le champ σ et $Z(V')$ la moyenne arithmétique des échantillons :

$Z(V^{'}) = \hat{Z} = \frac{1}{n} \sum_{n_i}^{n} Z(s_i)$

La formule générale de la variance d'extension s'écrit alors :

$$var(Z(\sigma) - \hat{Z}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} C(s_i - s_j) - 2 \sum_{i=1}^{n} \frac{1}{n|\sigma|} \int_{\sigma} C(s_i - s)\, \mathrm{d}s + \frac{1}{|\sigma|^2} \int_{\sigma} \int_{\sigma} C(s - t)\, \mathrm{d}s\mathrm{d}t \tag{2.1}$$

Cette expression ré présente ce que l'on appelle la **variance d'estimation de** σ par les échantillons $Z(s_i)$. Conceptuellement, il n'existe pas de différences entre variances d'extension et variances d'estimation : la variance d'extension de v à V est simplement la variance d'estimation de $Z(V)$ par $Z(v)$.
En général, on préfère cependant utiliser la terminologie de variance d'extension pour les cas élémentaires, par exemple le cas où l'on étend un échantillon à sa zone d'influence. L'expression variance d'estimation est employée pour des situations plus générales dans lesquelles les échantillons sont utilisés simultanément pour estimer une quantité déterminée, comme la moyenne globale du champ.
L'expression ci-dessus est toutefois difficile à calculer numériquement pour peu que le nombre d'échantillons soit important ou que la géométrie de

σ

soit compliquée. Nous allons voir comment se simplifie cette formule pour certains types d'échan-

tillonnage particuliers.

0.1 L'estimation globale

L'apport de la géostatistique est la prise en compte de la géométrie de l'information et de la structure de la variable étudiée dans le calcul d'une variance d'estimation globale "réaliste". Il est en effet intuitif de supposer que la précision d'une estimation dépend de ces deux paramètres : une variable chaotique donnera lieu à une plus grande imprécision qu'une variable régulière. Les différences d'approche que nous allons développer tiennent aux hypothèses faites sur le mode d'échantillonnage.

1. Echantillonnage ponctuel aléatoire pur

   Les emplacements des sites échantillonnés sont tirés au hasard, indépendamment les uns des autres, et uniformément dans le champ σ. On cherche à estimer la moyenne exhaustive sur le champ σ :

   $$Z(\sigma) = \int \sigma Z(s)\,\mathrm{d}s \tag{2.2}$$

   par la moyenne des échantillons :
   $\hat{Z} = \frac{1}{n}\sum_{i=1}^{n} Z(s_i)$
   La variance s'exprimera de la manière suivante :

   $$var\hat{Z} - Z(\sigma) = \frac{D^2(o|\sigma)}{n} \tag{2.3}$$

   où $D^2(o|\sigma)$ désigne la variance de dispersion d'un point dans le domaine σ.

   On pourra alors comparer l'équation précédente à celle donnée par la statistique classique $var\hat{Z} - m = \frac{\sigma^2}{n}$.
   en géostatistique, le numérateur de l'expression n'est pas la variance a priori du modèle, mais la variance de dispersion d'un point dans le champ ; cette quantité dépend du domaine à estimer et ne coïncide avec la variance a priori du modèle que dans le cas particulier d'un domaine infiniment grand ou d'un effet de pépite pur : $C(h) = 0$ pour $|h| > O$. Par ailleurs, la variance de dispersion peut être définie dans le cadre intrinsèque strict, contrairement à la variance a priori. Ainsi, la géostatistique permet d'utiliser une gamme de modèles plus riche que la statistique classique, limitée au cadre stationnaire du second ordre.

2. Echantillonnage ponctuel aléatoire stratifié

   Le domaine σ escoupé en n cellules Vi toutes identiques à une même cellule de référence V. On tire au hasard, à l'intérieur de chaque cellule, un échantillon indépendamment des autres prélèvements. La variance d'estimation globale est,

$$var\hat{Z} - Z(\sigma) = \frac{D^2(o|V}{n} \tag{2.4}$$

   où $D^2(o|V)$ est la variance de dispersion d'un point dans la cellule de référence V. On peut noter que l'échantillonnage aléatoire stratifié conduit toujours à une variance d'estimation inférieure à celle de l'échantillonnage aléatoire pur, puisque, d'après la relation d'additivité ouformule de Krige $D^2(o|V) - D^2(o|\sigma) = D^2(V|\sigma) \leq 0$

3. Echantillonnage ponctuel systématique ou régulier

   ici on découpe le domaine σ en n cellules V_i toutes identiques à une même cellule de réference V et on échantillonne le point central s_i de chacune de ces cellules V_i.
   La variance d'obtient par :

$$var(Z(\sigma)-\hat{Z}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C(s_i-s_j) - 2\sum_{i=1}^{n}\frac{1}{n|\sigma|}\int_{\sigma} C(s_i-s)\,\mathrm{d}s + \frac{1}{|\sigma|^2}\int_{\sigma}\int_{\sigma} C(s-t)\,\mathrm{d}s\mathrm{d}t \tag{2.5}$$

   Cette expression est souvent difficile à calculer lorsque le nombre n d'échantillons est élevé ou lorsque la géométrie du champ σ est complexe. Pour la simplifier, l'idée est de décomposer
   l'erreur d'estimation globale :
   $(\hat{Z} - Z(\sigma) = \frac{1}{n}\sum_{i=1}^{n} Z(s_i) - Z(\sigma))$
   en une somme d'erreurs élémentaires $(Z(s_i) - Z(V_i))$, dont il est facile de calculer la variance, et de supposer ces erreurs élémentaires non corrélées.
   On montre que la variance d'estimation s' écrit :
   $var(Z(\sigma) - \hat{Z}) = \frac{\sigma_E^2(o,V)}{n}$
   où $\sigma_E^2(o, V)$ est la variance d'extension d'un échantillons s_i dans son domaine V_i :

0.2 Estimation globale en support non ponctuel

On suppose à présent que les unités statistiques échantillonnées ne sont plus ponctuelles, mais sont des surfaces ou des volumes identiques à un même support de référence v. On fait l'hypothèse que le champ σ peut être partitionné en cellules qui se déduisent toutes de v par translation, et que l'échantillonnage s'opère parmi ces cellules. En particulier, le nombre d'unités statistiques à l'intérieur du champ σ est fini, contrairement au cas ponctuel, d'où des modifications dans les définitions précédentes et des corrections dans certaines formules. Ainsi, le nombre de localisations possibles des unités statistiques étant fini, on travaille avec des sommes discrètes au lieu d'intégrales : les moyennes sur le champ σ ainsi que les variances de dispersion et d'extension sont définies par des sommes et non par des intégrales.

Si on désigne par v le support de l'unité statistique échantillonnée, n le nombre d'échantillons et N le nombre total d'unités statistiques dans σ, alors la variance d'estimation s'exprime par :

$$var(\hat{Z} - \hat{\sigma}) = \frac{N-n}{N-1} \frac{D^2(v|\sigma)}{n} \tag{2.6}$$

Dans le cas des échantillonnages systématique et stratifié aléatoire avec une unité statistique échantillonnée par cellule, les sites d'observation peuvent toujours être considérés comme tirés indépendamment les uns des autres, et les relations trouvées précédemment restent valables, à condition de remplacer, dans les formules des variances d'extension et de dispersion, le point (o) par l'unité statistique v.

0.3 Comparaison avec la statistique descriptive

il existe une forte parenté entre les formules de variances d'estimation obtenues en géostatistique et celles de la statistique descriptive. Cette dernière n'interprète pas les valeurs observées comme des réalisations de variables aléatoires, mais utilise les moyennes et variances statistiques calculées sur les échantillons, ce qui permet de prendre en compte la non-indépendance des observations. Elle s'affranchit même de l'hypothèse de stationnarité, puisqu'il n'y a plus d'interprétation probabiliste des valeurs.

Le tableau suivant nous donne un bon récapitulatif :

statistique descriptive	**ponctuel**	**bloc**
échantillonnage aléatoire pur	$\frac{s^2(o\|\mathfrak{G})}{n}$	$\frac{N-n}{N-1}\frac{s^2(v\|\mathfrak{G})}{n}$
échantillonnage aléatoire stratifié avec p unités par cellule (p>1) et n/p cellules	$\frac{p}{n^2}\sum_{i=1}^{n/p} s^2(o\|V_i)$	$\frac{p(N-n)}{Np-n}\frac{p}{n^2}\sum_{i=1}^{n/p} s^2(v\|V_i)$
échantillonnage systématique	non calculable	non calculable

Récapitulatif sur l'estimation globale par les méthodes statistiques et géostatistiques

support ponctuel		**statistique classique** (stationnarité d'ordre 2)	**statistique descriptive** (pas d'hypothèse de stationnarité)	**géostatistique intrinsèque** (stationnarité d'ordre 2 ou hypothèse intrinsèque)
échantillonnage aléatoire pur		$\frac{\sigma^2}{n}$	$\frac{s^2(o\|\mathfrak{G})}{n}$	$\frac{\mathfrak{G}^2(o\|D)}{n}$
échantillonnage stratifié aléatoire	1 point par cellule		incalculable	$\frac{D^2(o\|V)}{n}$
	p points par cellule et n/p cellules		$\frac{p}{n^2}\sum_{i=1}^{n/p}\frac{s^2(o\|V_i)}{n}$	
échantillonnage systématique			incalculable	$\approx\frac{\sigma_E^2(o\|V)}{n}$

L'unité échantillonnée est une surface (ou un volume) identique à une surface (ou volume) de référence v, et le nombre de localisations possibles dans le champ est fini.

support v non ponctuel		**statistique classique** (stationnarité d'ordre 2)	**statistique descriptive** (pas d'hypothèse de stationnarité)	**géostatistique intrinsèque** (stationnarité d'ordre 2 ou hypothèse intrinsèque)
échantillonnage aléatoire pur		$\frac{N-n}{N-1}\frac{\sigma^2}{n}$	$\frac{N-n}{N-1}\frac{s^2(v\|\mathcal{G})}{n}$	$\frac{N-n}{N-1}\frac{D^2(v\|\mathcal{G})}{n}$
échantillonnage stratifié aléatoire	1 point par cellule		incalculable	$\frac{D^2(v\|V)}{n}$
	p points par cellule et n/p cellules		$\frac{p(N-n)}{Np-n}\frac{p}{n^2}\sum_{i=1}^{n/p}\frac{s^2(v\|V_i)}{n}$	$\frac{p(N-n)}{Np-n}\frac{D^2(v\|V)}{n}$
échantillonnage systématique			incalculable	$\approx\frac{\sigma_E^2(o\|V)}{n}$

L'estimation locale : Le Krigeage et le Cokrigeage

Le krigeage s'appuie sur l'interprétation de la variable régionalisée comme la réalisation d'une fonction aléatoire, dont on suppose modélisée la structure spatiale (covariance ou variogramme). Il s'agit en l'occurrence de rechercher, parmi les estimateurs linéaires, celui qui présente les "meilleures" propriétés (à savoir absence de biais et variance minimale). Mathématiquement, le krigeage n'est, ni plus ni moins, qu'une technique de régression multiple qui minjmise l'erreur quadratique moyenne, à partir de données corrélées. L'avantage du krigeage sur les techniques d'interpolation déterministes (interpolation linéaire ou polynomiale après triangulation de l'espace, splines, ...) est d'une part qu'il évite de commettre des erreurs systématiques dans l'estimation, et d'autre part qu'il fournit une variance d'estimation (c'est-à-dire une variance de l'erreur d'estimation, aléatoire dans le modèle probabiliste). Signalons dès à présent qu'il ne faut pas confondre la variance d'estimation avec un intervalle de confiance sur l'estimation25 , auquel on a l'habitude de se référer. La variance d'estimation constitue quand même un apport non négligeable, car elle permet d'apprécier quantitativement la précision de l'estimation.

Le krigeage est l ?approche géostatistique la plus utilisée pour réaliser des interpolations spatiales. Les techniques de krigeage sont basées sur la définition d ?un modèle spatial entre les observations (défini par un variogramme) pour prédire les valeur attributaires d ?une variable à des positions où aucun échantillon n'est disponible. Une des spécificités du krigeage est qu ?il ne considère pas seulement la distance entre les observations (comme pour les méthodes déterministes) mais qu'il essaye aussi de capturer la structure spatiale des données en comparant deux à deux les observations séparées par des distances spatiales spécifiques.

L'objectif est de comprendre les relations existantes entre les observations séparées par des distances différentes dans l'espace. Toute cette information est prise en compte par le variogramme. Les méthodes de krigeage utilisent cette information pour attribuer un poids à chaque échantillon avant de réaliser les prédictions. Il faut noter que les techniques de krigeage permettent de préserver les valeurs attributaire des échantillons sur la carte interpolée.

Les méthodes de krigeage considèrent que le processus qui a donné naissance aux données peut être séparés en deux composantes principales : une tendance déterministe (les variations à large échelle) et une erreur auto-corrélée (les résidus) :

$$Z(s) = m + e(s) \tag{3.1}$$

Où $Z(s)$ est la valeur attributaire à la position s dans la parcelle, m est la tendance déterministe qui ne dépend pas de la position des données dans l'espace et $e(s)$ qui est le terme d ?erreur auto-corrélée (qui dépend de la position s). A noter que la tendance est déterministe ! Le variogramme est seulement calculé sur les résidus qui sont supposés être auto-corrélés ! En gros, quand on cherche à ajuster un modèle de variogramme aux données, on essaye en réalité d ?ajuster ce modèle aux résidus de ces données, après que la tendance ait été enlevée. Attention au fait que, en fonction de la forme du variogramme, la carte interpolée peut être relativement lissée. En fait, si l'effet pépite est forte, les méthodes de **krigeage** auront tendance à lisser fortement les données pour réduire le bruit dans les parcelles. C'est parfois surprenant parce que l'étendue (écart entre minimum et maximum) des valeurs après krigeage est parfois plus faible que celle des échantillons de départ. Si l ?effet pépite est effectivement très fort par rapport au pallier partiel, la structure spatiale est relativement faible et il serait peut être préférable de ne pas interpoler. Les dérivés du krigeage dépendent essentiellement de la façon dont la tendance est caractérisée. Ce point sera discuté en détail plus tard.

Grâce à l'utilisation du variogramme, une carte d'erreur peut être dérivée sur l'ensemble de la parcelle parce que les relations entre deux observations séparées par une distance h sont connues. La carte d ?erreur est un outil très utile parce qu'elle donne accès à la qualité de la prédiction sur la totalité de la parcelle. Prenez en compte que la qualité de la carte d'erreur dépend bien évidemment de la qualité d ?ajustement du modèle de variogramme aux données. Si ce modèle de variogramme n ?est pas bien ajusté aux données, il y a de grands risques que la carte d'erreurs ne soit pas fiable. Attention également au fait que l'interpolation par krigeage est aussi sensible aux données aberrantes. Dans ce cas, ces outliers peuvent masquer l'auto-

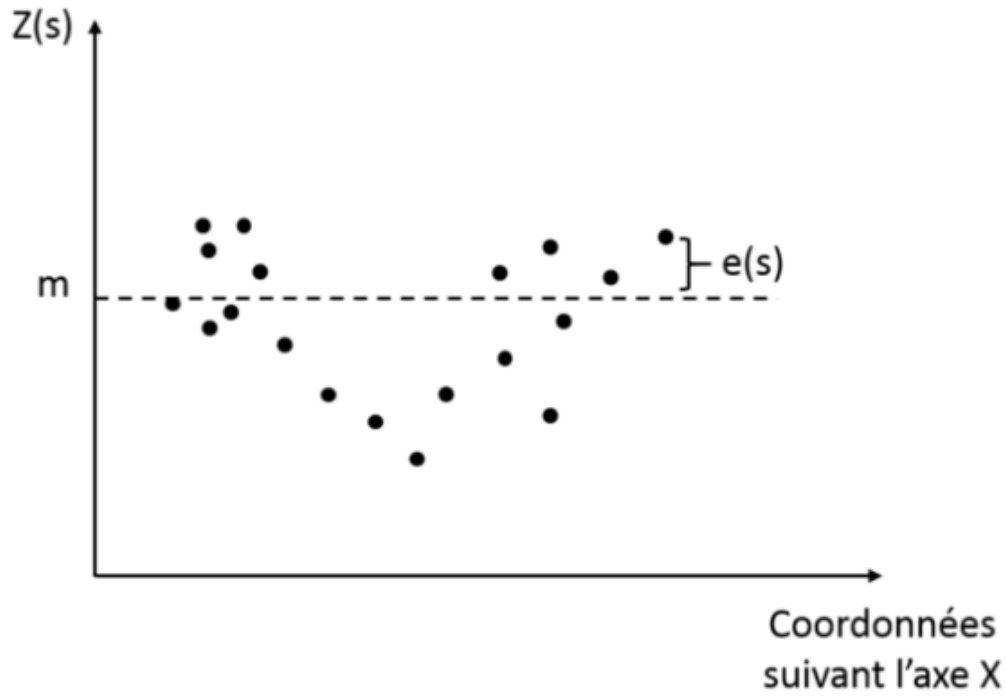
corrélation des données et empêcher la structure spatiale d ?être mise en avant. Une solution intuitive serait d ?enlever ces données aberrantes avant de réaliser le krigeage. Malgré tout, il est possible que certaines observations semblent aberrantes parce qu'elles sont beaucoup plus faibles ou fortes que le reste des observations. Si ces observations sont regroupées dans l'espace, cela pourrait être dû à un phénomène réellement existant au sein de la parcelle. Dans ce cas là, pour prendre en compte ce phénomène, mais pour ne pas empêcher la structure spatiale d ?être découverte, une solution serait de calculer le semi-variogramme sans ces observations extrêmes et ensuite de réaliser le krigeage avec l'ensemble des informations. De cette manière, la structure spatiale serait bien évaluée et les données extrêmes seraient toujours prises en compte dans le jeu de données.

Remarque 0.1. système et la variance de krigeage prennent en compte :

- les distances entre le point à estimer s_0 et les sites d'observation s_i par l'intermédiaire des termes $C(s_i - s_0)$ ou $\gamma(s_i - s_0)$;
- la configuration géométrique des sites d'observation s_i, par l'intermédiaire des termes $C(s_i - s_0)$ ou $\gamma(s_i - s_0)$;
- la structure spatiale de la variable étudiée, apportée par la covariance C ou le variogramme γ. A noter que la variance de krigeage ne dépend que de la structure C (ou γ) et de la configuration de krigeage (c'est-à-dire de la configuration géométrique des sites d'observation) et non des valeurs observées. On peut donc, connaissant C (ou γ), prévoir la qualité du krigeage avec une configuration donnée des sites d'observation.

0.1 Krigeage simple

Comme le nom le laisse à penser, le krigeage simple est la méthode dérivée du krigeage la plus facile à mettre en oeuvre. Ici, la tendance déterministe, m, est connue et considérée constante sur la totalité de la parcelle d'étude.
La figure suivante donne montre un exemple de krigeage simple et tendance et résidus correspondants



Cette méthode est globale parce qu'elle ne prend pas en compte les variations locales de cette tendance. Néanmoins, s'il n'y a pas de changements brusques dans les attributs de la variable d ?intérêt (ou s'il n'y a pas de raisons qu'il n'y en ait), cette hypothèse peut être viable et pertinente. Comme la tendance est connue, le terme d'erreur auto-corrélée est lui aussi connu ce qui rend la prédiction plus simple à faire. Il doit être compris que, ici, les valeurs prédites ne peuvent pas s'étendre au-delà des valeurs des échantillons initiaux.

0.2 Krigeage ordinaire

Cette méthode est peut-être l'approche de krigeage la plus largement reportée. Contrairement au krigeage simple, cette technique considère que la tendance est constante mais seulement au niveau d'un voisinage local . Cette hypothèse est intéressante parce qu'elle assure de prendre en compte les variations locales au sein d ?une parcelle.

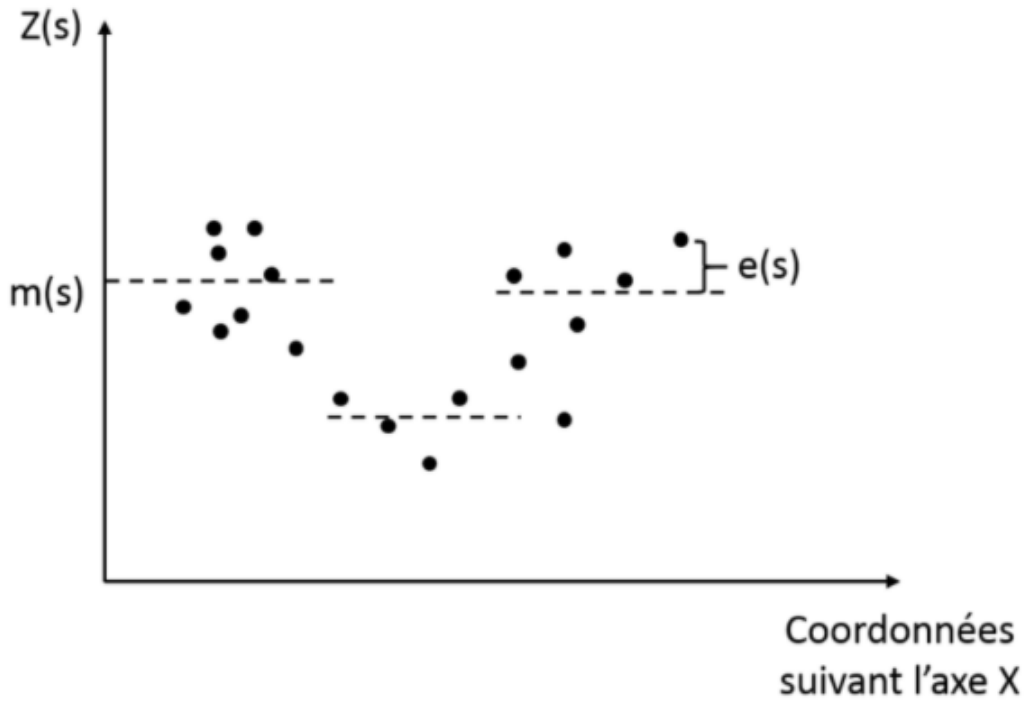
On pourrait imaginer par exemple qu'une rupture de pente au sein d'une parcelle serait intéressante à prendre en compte dans l'étude d ?une variable d'intérêt.

Le krigeage ordinaire peut être exprimé de la façon suivante :

$$Z(s) = m(s) + e(s) \tag{3.2}$$

Ici, la tendance dépend de la position spatiale des observations $(m(s))$. Cette tendance constante est considérée inconnue et doit être déterminée à partir du voisinage de données correspondant.

La figure suivante donne montre un exemple de krigeage simple et tendance et résidus correspondants

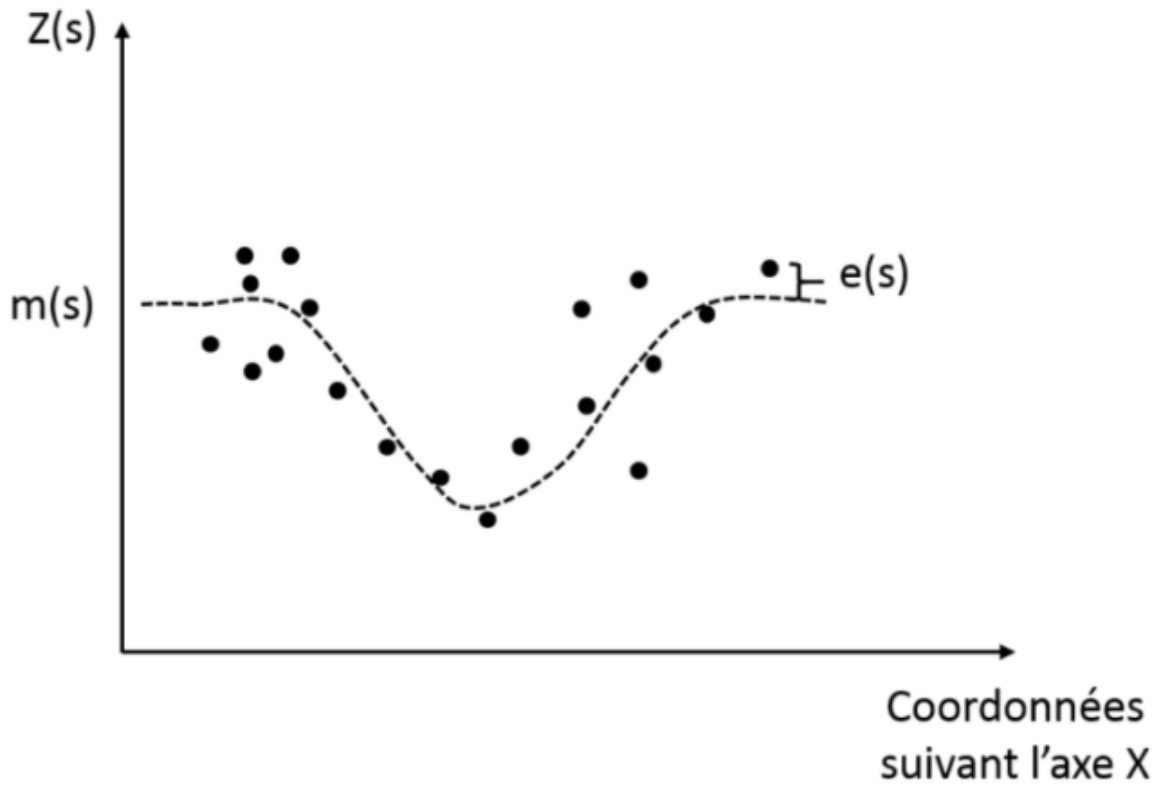


Pour cette méthode, il est nécessaire de définir l'étendue du voisinage : le nombre d'observations voisines qui seront prises en compte pour chaque prédiction. Comme la tendance n'est pas considérée connue et qu'elle doive être déterminée par les données elles-mêmes, il peut arriver que, par construction, l'étendue des valeurs prédites soient plus larges sur celle des échantillons de départ.

0.3 Krigeage universel ou à tendance externe

Le krigeage par régression a plusieurs noms : le krigeage universel ou le krigeage avec une tendance externe. Cette méthode est similaire au krigeage ordinaire dans le sens où la tendance n'est pas constante sur la totalité de la parcelle mais dépend de la position des observations dans l'espace. Néanmoins, ici, la tendance est modélisée par une fonction plus complexe, elle n'est

pas simplement considérée constante au sein d'un voisinage local. L'objectif est le même que précédemment : enlever la tendance des données pour que les résidus auto-corrélés puissent être étudiés. A la fin de l'interpolation, la tendance est rajoutée aux résidus interpolés.



Le krigeage par régression est intéressant lorsque l'on dispose d'une variable auxiliaire à haute résolution spatiale et qu'on voudrait avoir plus d ?information sur une seconde variable pour laquelle seuls quelques échantillons sont disponibles. En général, cette deuxième variable est pénible, chère ou chronophage à obtenir. Dans ce cas, l'objectif est de modéliser la relation entre cette seconde variable et la variable auxiliaire de manière à améliorer les prédictions de la seconde variable.

Par exemple, l'information de biomasse est relativement facile à acquérir (à partir d'indices de végétation comme le $NDVI$ obtenus avec des images satellites, avions ou drone). On pourrait imaginer utiliser cette variable auxiliaire pour aider à interpoler des observations de rendement collectées à une résolution spatiale beaucoup plus grossière pendant une campagne de terrain).

A noter que si la relation entre ces deux variables est faible, le krigeage par régression reviendra à réaliser un krigeage simple ou un krigeage ordinaire.

0.4 Co-Krigeage

Lorsque l'on dispose d'une variable auxiliaire $V0$ avec une haute résolution spatiale et que l'on cherche à capturer la variabilité ou la corrélation spatiale d'une seconde variable $V1$, le co-krigeage est particulièrement intéressant.
En fait, l'objectif est d'évaluer la structure spatiale de $V1$ par rapport à $V0$ avec les échantillons à disposition et d ?interpoler la structure spatiale caractérisée sur l'ensemble de la parcelle. Comme il l'a été dit précédemment, $V1$ est généralement chronophage et coûteux à obtenir et il est beaucoup plus simple d ?utiliser une variable auxiliaire pour améliorer les prédictions de $V1$. Le co-krigeage demande à l'utilisateur de calculer un cross-variogramme et d'y ajuster un modèle de variogramme. Dans ce cas, l'idée n'est pas d'étudier l'évolution de la variance d'une variable pour des distances spécifiques entre observations mais plutôt d ?étudier l'évolution de la covariance de $V1$ par rapport à $V0$ pour ces distances spatiales particulières.
En d'autres termes, on regarde comme la structure spatiale de $V1$ évolue par rapport à $V0$. On peut rajouter que le cross-variogramme est également un outil utile pour vérifier si deux variables sont corrélées spatialement ou si elles présentent une structure spatiale similaire. Dans ce cas-là, les deux variables peuvent avoir été acquises avec une haute résolution spatiale parce que l'objectif n'est plus de les interpoler mais de les comparer l'une par rapport à l'autre. Le co-krigeage est plus difficile à mettre en place que les autres techniques de krigeage mais il peut s ?avérer plus efficace si les structures spatiales sont correctement déterminées.

0.5 Krigeage par point et krigeage par bloc

Toutes les méthodes de krigeage mentionnées précédemment ont pour vocation de prédire la valeur d'une variable d'intérêt à des positions où aucun échantillon n'est disponible. Ces positions peuvent être considérées comme des points dans l'espace (ou plus précisément comme des pixels de la grille d'interpolation avec un pixel = une valeur). Par conséquent, ces approches de krigeage peuvent être comprises comme des méthodes de krigeage par point. Quand l'incertitude de la prédiction est relativement large, on pourrait vouloir chercher à lisser les résultats interpolées en réalisant un krigeage sur une surface plus large qu'un tout petit pixel. Ce type de krigeage est connu sous le nom de krigeage par bloc. Cela a l'avantage de diminuer la variance de l'erreur de prédiction parce que l'information est plus grossière (elle est prédite sur un support spatial plus grand). Bien évidemment, avec le block kriging, il y a un risque de perdre de l'information utile mais quand l'incertitude est forte, cette approche peut être très pertinente.

Avant de terminer ce post, il est important de préciser que toutes les techniques d'interpolation qui ont été présentées produisent des résultats relativement similaires dans les cas idéaux (beaucoup d'échantillons disponibles et bien disposés dans la parcelle, peu de bruit dans les données et pas de variations brusques au sein de la parcelle). Néanmoins, quand ces conditions ne sont pas respectées, les méthodes géostatistiques sont intéressantes parce qu'elles permettent de capturer la structure spatiale sur l'ensemble de la parcelle, ce qui conduit généralement à des prédictions plus précises. Ces méthodes sont plus difficiles à mettre en place parce qu'il est nécessaire de construire un modèle spatial et de l'ajuster aux données mais cela peut permettre de mieux caractériser les parcelles. Encore une fois, je recommanderais de toujours garder une étape manuelle dans la mise en place d'approches géostatistiques parce que les résultats pourraient paraître étrange si les modèles spatiaux ne sont pas correctement ajustés aux données. Quand la structure spatiale est faible (lorsque le bruit est important), presque toutes les méthodes d'interpolation auront tendance à lisser les données dans une plus ou moins grande mesure. Cela doit être gardé à l'esprit avant d'interpoler des observations spatiales parce que les cartes interpolées pourraient paraître plus homogènes qu'elles ne le sont vraiment. On peut rajouter que les techniques d ?interpolation peuvent être également utilisées pour sous-échantillonner (le plus souvent) ou sur-échantillonner une image ou une carte. L'objectif est de changer la résolution d ?une carte ou d'une image disponible.

♣ Bibliographie ♣

[1] Thibault LAURENT, *Statistique spatiale avec R*

[2] *Internet*

[3] Michel ARNAUD,Xavier EMERY, *Estimation et interpolation de données spatiales*

[4] Sébastien Rochette *modélisation et cartographie*

[5] *Wikipedia*