Introduction à la statistique spatiale

Edith Gabriel

Introduction à la statistique spatiale

Edith Gabriel

IUT STID
Université d'Avignon et des Pays de Vaucluse
BP 1207
84911 Avignon

Résumé : La géologie, la météorologie, l'épidémiologie, la foresterie, les sciences du sol, l'écologie, ..., sont autant de domaines de recherche où les données ont pour point commun d'être localisées dans l'espace géographique et d'être ni indépendantes, ni identiquement distribuées. Il s'agit d'observations d'un processus aléatoire $\{Z(s); s \in \mathcal{D}\}$, où $\mathcal{D} \subset \mathbb{R}^d$, s est une localisation spatiale et $Z(s)$ est une quantité aléatoire. Leur modélisation nécessite de caractériser la dépendance (spatiale) entre différentes observations et leur caractère non homogène : moyenne non constante (variations à grande échelle) et/ou hétéroscédasticité (variations à petite échelle). On s'intéressera au cadre géostatistique où la variable d'étude se déploie continûment sur le domaine $\mathcal{D}$ et $Z(s)$ est un vecteur aléatoire en $s \in \mathcal{D}$. Il s'agira dans un premier temps de justifier le choix d'outils de statistique spatiale pour faire de l'estimation, prédiction à partir de telles données. Nous verrons ensuite comment caractériser l'organisation spatiale des variables étudiées (analyse variographique) et présenterons la méthode de krigeage qui permet de prédire la valeur prise par une variable en un site non échantillonné à partir d'observations ponctuelles en des sites voisins.

Mots-clés: Géostatistique, Statistique spatiale.

Abstract: Data in geology, meteorology, epidemiology, forestry, soil science, ecology, ..., have in common that they are collected from different spatial locations and that they are neither independent nor identically distributed. They result from observations of a stochastic process $\{Z(s); s \in \mathcal{D}\}$, where $\mathcal{D} \subset \mathbb{R}^d$, s is a spatial location and $Z(s)$ is a random quantity. Modeling such data needs to account for their spatial dependence and the small/large scale variations. Here, we focus on a spatially continuous phenomena and consider that $Z(s)$ is a random vector at location $s \in \mathcal{D}$ (referred to as geostatistical data). First, we justify the choice of spatial statistics tools to estimate, predict geostatistical data. Then, we present methods to model the spatial structure of the data (variogram modeling) and to make prediction at an unsampled location from observed data (kriging).

Key words: Geostatistics, Spatial statistics.

1 Introduction

La statistique spatiale concerne l'étude de phénomènes observés dans un domaine spatial. On note $s \in \mathcal{D} \subset \mathbb{R}^2$ la localisation d'un site de mesure et $Z = \{Z(s) : s \in \mathcal{D}\}$ le phénomène étudié, où Z est une variable aléatoire indexée par l'ensemble $\mathcal{D}$. On distingue trois types de données, selon la nature du domaine $\mathcal{D}$, appelant des traitements statistiques spécifiques :

- Les processus ponctuels : $\mathcal{D}$ est un processus ponctuel dans $\mathbb{R}^2$.

La localisation s est elle-même l'objet de l'étude. L'ensemble des sites d'observations est la réalisation d'un processus ponctuel ; le nombre de réalisations ponctuelles et leur localisation sont aléatoires. Par exemple, une question centrale lors de l'étude de la répartition spatiale d'une espèce d'arbres dans une forêt est de savoir si la répartition est plutôt régulière ou aléatoire ou si elle présente des aggrégats.

- Les données latticielles : $\mathcal{D}$ est discret et fixé.

Les sites de localisation ne sont pas des points, ni répartis de manière aléatoire, mais sont agencés selon un réseau structuré à partir d'un graphe de voisinage : lattices régulières (analyse d'images) ou non (épidémiologie). Dans ce cas, Z intègre la variable d'intérêt sur l'unité géographique s.

- Les données géostatistiques : $\mathcal{D}$ est un sous-espace continu de $\mathbb{R}^2$.

La quantité d'intérêt, Z, est à valeurs réelles et mesurée en des sites expérimentaux choisis $\{s_1, \ldots, s_n\}$. Dans l'exemple emblématique du domaine de la prospection minière, Z est la hauteur d'un filon mesurée par carottage en différents sites. Ici, la géographie des sites joue un rôle central dans la loi de variation de Z. Cette propriété définit le cadre de travail de la géostatistique.

Les domaines d'application de la statistique spatiale sont nombreux : géologie, écologie, sciences du sol, météorologie, épidémiologie, ..., et sont autant de domaines de recherche où les données ont pour point commun d'être localisées dans l'espace géographique et d'être ni indépendantes, ni identiquement distribuées. Nous nous intéresserons ici à la modélisation de données géostatistiques. Un des objectifs de l'étude de telles données est de fournir une cartographie des variables intégrant l'ensemble des informations disponibles relatives au phénomène d'étude, tout en précisant les incertitudes associées à cette cartographie.

Nous verrons dans la section 2 comment caractériser les dépendances spatiales entre observations et les variations à petite et à grande échelles, i.e. comment identifier et estimer la structure spatiale des données. Dans la section 3 nous présenterons rapidement la méthode de krigeage qui permet de prédire la valeur prise par une variable en un site non échantillonné à partir d'observations ponctuelles en des sites voisins. Pour plus de détails, le lecteur pourra par exemple se référer à Cressie (1993). La section 4 aborde la question de la mise en pratique du krigeage.

2 Modélisation de la variabilité spatiale

En géostatistique la réalisation du phénomène étudié est unique. Il est donc nécessaire de postuler des hypothèses (l'ergodicité et la stationnarité) afin de rendre possible l'inférence statistique malgré l'unicité de la réalisation.

L'hypothèse d'ergodicité permet d'inférer les paramètres (i.e. les moments) de la loi à partir d'une réalisation unique : pour $\mathcal{D}_1 \subset \cdots \subset \mathcal{D}_n \subset \ldots$, on a : $\mathbb{E}\left[Z(\mathcal{D}_n)\right] \longrightarrow m$ $(= \mathbb{E}\left[Z(s)\right])$, lorsque $n \to \infty$.

Au sens strict, la stationnarité signifie que la loi de probabilité de Z est invariante par translation. Cependant en théorie du krigeage, on utilise une hypothèse plus faible :

– soit la stationnarité intrinsèque $(\mathcal{H}_1)$:

$$\mathbb{E}\left[Z(s+h) - Z(s)\right] = 0 \text{ et } \mathrm{Var}(Z(s+h) - Z(s)) = 2\gamma(h),\ \forall s,\ \forall h,$$

i.e. l'espérance de tout accroissement est nulle et la variance de tout accroissement existe, dépend uniquement de h et est appelée *variogramme*.

– soit la stationnarité d'ordre 2 $(\mathcal{H}_2)$:

$$\mathbb{E}\left[Z(s)\right] = m \text{ et } \mathrm{Cov}(Z(s), Z(s+h)) = C(h),\ \forall s,\ \forall h,$$

i.e. l'espérance de Z est la même en tout site et la covariance de Z ne dépend que du vecteur de translation entre les points s et $s+h$. Dans ce cas, $\gamma(h) = C(0) - C(h)$.
Sous ces hypothèses, le variogramme ne dépend que du vecteur de translation h et donc de la distance entre s et $s+h$ et de l'orientation de h. Il est dit isotrope lorsqu'il ne dépend que de la norme de h et anisotrope s'il dépend de la direction de h. Dans la suite nous supposerons l'isotropie.

L'analyse de la régularité du processus des variations locales repose sur le (semi)-variogramme $\gamma(h) = \frac{1}{2}\mathrm{Var}\left(Z(s+h) - Z(s)\right)$. Pour une valeur de h donnée, on obtient une estimation empirique de $\gamma(h)$ de la manière suivante : soit $N(h)$ l'ensemble des couples (s_i, s_j) de sites de mesure tels que $s_i - s_j = h$, alors

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(s_i,s_j)\in N(h)} \left\{Z(s_i) - Z(s_j)\right\}^2,$$

où $|N(h)|$ désigne le nombre d'éléments de l'ensemble $N(h)$.

Le variogramme empirique sert de support au choix d'un modèle théorique de variogramme décrivant de manière satisfaisante la régularité des variations locales. L'ensemble des modèles théoriques de variogramme est schématisé par le graphique de la figure 1a. Ce graphique met en avant trois paramètres fondamentaux :

– le palier (ou seuil) du variogramme. Il s'agit de sa valeur limite pour de grandes valeurs de h. Ce paramètre suscite un intérêt très important dans l'analyse de la régularité des variations locales. En effet, si le seuil d'un variogramme est infini, alors le processus des variations locales n'est pas stationnaire.

(a) (b)

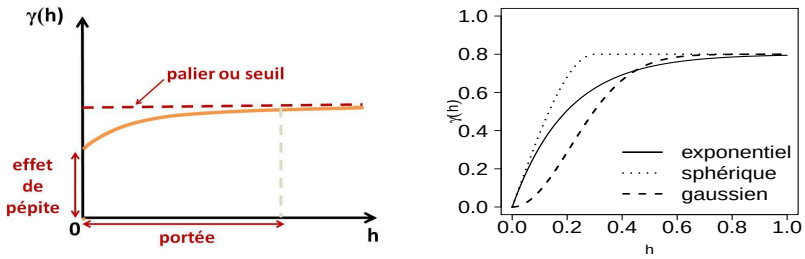

Figure 1: **(a)** Schéma type d'un variogramme. **(b)** Exemples de variogrammes.

– la portée. Elle représente la valeur h pour laquelle le variogramme atteint une limite et caractérise la distance entre sites de mesures au-delà de laquelle les dépendances entre mesures du processus sont nulles. Elle représente l'échelle de l'hétérogénéité du processus.

– l'effet de pépite. Il représente une discontinuité du variogramme à l'origine et traduit une forte irrégularité du processus des variations locales.

Les modèles classiques de variogrammes sont :

– le modèle exponentiel : $\gamma(h) = \begin{cases} 0, & \text{si } h = 0 \\ \sigma_0^2 + \sigma^2 \left(1 - \exp(-h/a)\right), & \text{si } h > 0 \end{cases}$,

– le modèle sphérique : $\gamma(h) = \begin{cases} 0, & \text{si } h = 0 \\ \sigma_0^2 + \sigma^2 \left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right), & \text{si } 0 < h \leq a \\ \sigma_0^2 + \sigma^2, & \text{si } h > a \end{cases}$,

– le modèle gaussien : $\gamma(h) = \begin{cases} 0, & \text{si } h = 0 \\ \sigma_0^2 + \sigma^2 \left(1 - \exp(-(h/a)^2)\right), & \text{si } h > 0 \end{cases}$,

– le modèle pépitique : $\gamma(h) = \begin{cases} 0, & \text{si } h = 0 \\ \sigma_0^2, & \text{si } h > 0 \end{cases}$,

où a désigne la portée, σ_0^2 la variance à l'origine et $\sigma_0^2 + \sigma^2$ le palier. La figure 1b représente les modèles exponentiel, sphérique et gaussien.

3 Le krigeage

Le krigeage est essentiellement utilisé pour compléter les données manquantes, simuler des jeux de données, estimer une valeur dans une région particulière ou représenter l'incertitude d'estimation.

Le modèle de base du krigeage a une expression analogue à celle du modèle de régression classique :

$$Z(s) = \mu(s) + \varepsilon(s)$$

où $\mu(s)$ désigne une structure déterministe pour l'espérance de Z et $\varepsilon(s)$ est une fonction aléatoire, stationnaire, d'espérance nulle et de structure de dépendance spatiale connue. La modélisation en pratique des dépendances spatiales a été abordée dans la section 2. Le krigeage dépend de la forme de la tendance $\mu(\cdot)$: krigage simple si $\mu(s) = m$, où m est une constante connue ; krigage ordinaire si $\mu(s) = m$, où m est une constante inconnue ; krigeage universel si $\mu(s) = \sum_{k=1}^{p} \beta_k f_k(s)$ est une combinaison linéaire de fonctions (connues) de la position s, les paramètres β_k étant inconnus.
Dans tous les cas, la prédiction en un site s_0, $\hat{Z}(s_0)$, satisfait les contraintes du krigeage :

- contrainte de linéarité : $\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i(s_0) Z(s_i)$,
- contrainte d'autorisation : $\mathbb{E}\left[\hat{Z}(s_0) - Z(s_0)\right]$ et $\mathrm{Var}\left(\hat{Z}(s_0) - Z(s_0)\right)$ existent.
- contrainte de non-biais : $\mathbb{E}\left[\hat{Z}(s_0) - Z(s_0)\right] = 0$.
- contrainte d'optimalité : $\mathrm{Var}\left(\hat{Z}(s_0) - Z(s_0)\right)$ est minimale.

La valeur minimale de la variance de prédiction est appelée variance de krigeage.
Les poids $\lambda_i(s_0)$ sont solutions du système de krigeage résultant de ces contraintes. Ainsi, on obtient :

- prédicteur du krigeage simple (sous l'hypothèse $\mathcal{H}_2$) :

$$\hat{Z}(s_0) = m + \mathbf{c}_0' \mathbf{C}^{-1} (\mathbf{Z} - m\mathbf{1}_n),$$

la variance de krigeage simple est $\sigma^2_{KS}(s_0) = \sigma^2(s_0) - \mathbf{c}_0' \mathbf{C}^{-1} \mathbf{c}_0$.

- prédicteur du krigeage ordinaire (sous l'hypothèse $\mathcal{H}_2$) :

$$\hat{Z}(s_0) = \left(\mathbf{c}_0 + \frac{1 - \mathbf{1}_n' \mathbf{C}^{-1} \mathbf{c}_0}{\mathbf{1}_n' \mathbf{C}^{-1} \mathbf{1}_n} \mathbf{1}_n \right) \mathbf{C}^{-1} \mathbf{Z},$$

la variance de krigeage ordinaire est $\sigma^2_{KO}(s_0) = \mathbf{c}_0' \mathbf{C}^{-1} \mathbf{c}_0 - \dfrac{(1 - \mathbf{1}_n' \mathbf{C}^{-1} \mathbf{c}_0)^2}{\mathbf{1}_n' \mathbf{C}^{-1} \mathbf{1}_n}$.

- prédicteur du krigeage universel (sous l'hypothèse $\mathcal{H}_1$) :

$$\hat{Z}(s_0) = \left(\gamma_0 + \mathbf{X} (\mathbf{X}' \mathbf{\Gamma}^{-1} \mathbf{X})^{-1} (\mathbf{x}_0 - \mathbf{X}' \mathbf{\Gamma}^{-1} \gamma_0) \right)' \mathbf{\Gamma}^{-1} \mathbf{Z},$$

la variance de krigeage universel est $\sigma^2_{KU}(s_0) = (\mathbf{x}_0 - \mathbf{X}' \mathbf{\Gamma}^{-1} \gamma_0)' (\mathbf{X}' \mathbf{\Gamma}^{-1} \mathbf{X})^{-1} (\mathbf{x}_0 - \mathbf{X}' \mathbf{\Gamma}^{-1} \gamma_0)$,
où $\mathbf{Z}$ est le n-vecteur des variables aléatoires $Z(s_1), \ldots, Z(s_n)$, $\mathbf{C}$ est la $n \times n$-matrice de covariance d'éléments $C_{ij} = \mathrm{Cov}(Z(s_i), Z(s_j))$, $\mathbf{\Gamma}$ est la $n \times n$-matrice d'éléments $\Gamma_{ij} = \gamma(Z(s_i) - Z(s_j))$, $\mathbf{c}_0$ est le n-vecteur dont le $i$ème élément est $\mathrm{Cov}(Z(s_i), Z(s_0))$, $\sigma^2(s_0) = \mathrm{Cov}(s_0, s_0)$, γ_0 est le n-vecteur dont le $i$ème élément est $\gamma(Z(s_i) - Z(s_0))$, $\mathbf{x}_0$ est le n-vecteur des fonctions $f_1(s_0), \ldots, f_p(s_0)$, $\mathbf{X}$ est la $n \times p$-matrice d'éléments $f_j(s_i)$ et $\mathbf{1}_n$ est le n-vecteur constitué de 1.
Le prédicteur de krigeage est un BLUP (Best Linear Unbiased Predictor). Il s'agit d'un interpolateur exact, i.e. $\hat{Z}(s_i) = Z(s_i)$, lissant la réalité, i.e. $\mathrm{Var}(\hat{Z}(s_i)) \leq \mathrm{Var}(Z(s_i))$.

4 Mise en pratique

En pratique, la mise en oeuvre du krigeage se décline en trois étapes : analyse exploratoire, modélisation et krigeage.

Comme dans toute analyse statistique, il est recommandé de commencer par une analyse exploratoire des données, en se rappelant qu'en géostatistique les données sont multivariées (chaque donnée est la réalisation d'une variable aléatoire qui a sa propre distribution). L'exploration spatiale (graphiques 3D, courbes de niveaux, ...) des données permet en outre de juger de la stationnarité de la variable, d'examiner le comportement directionnel de la variable (anisotropie) et d'identifier des valeurs aberrantes.

La formulation du modèle nécessite le choix de la forme de $\mu(\cdot)$ et l'estimation du variogramme. Les paramètres du modèle de variogramme sont usuellement estimés par la méthode des moindres carrés (ordinaires ou pondérés) ; le vecteur des paramètres, θ, minimise $\sum_{k=1}^{N} w_k \left\{\hat{\gamma}(h_k) - \gamma(h;\theta)\right\}^2$, où les h_k sont les distances pour lesquelles une estimation du variogramme $\hat{\gamma}(h_k)$ est faite (en général $\max_k h_k$ ne dépasse pas la demi-distance maximale entre deux points d'observation), $\gamma(;\theta)$ est le modèle variographique à valider et w_k désigne le poids associé à la donnée $\hat{\gamma}(h_k)$. Le modèle de variogramme peut ensuite être validé par la méthode de bootstrap paramétrique (approche de type Monte Carlo) ou par la méthode validation croisée qui consiste à éliminer à tour de rôle chaque observation $Z(s_i)$ et à la prévoir par krigeage $\hat{Z}(s_i)$ sans réestimer le variogramme. L'erreur quadratique normalisée moyenne, $\frac{1}{n}\sum_{i=1}^{n} \frac{\left\{\hat{Z}(s_i)-Z(s_i)\right\}^2}{\sigma_K^2(s_i)}$, est alors proche de 1 lorsque le modèle est bien estimé ; $\sigma_K^2(s_i)$ désigne la variance de krigeage. Cette méthode permet aussi de comparer la qualité des prédictions provenant de différents modèles et surtout d'en choisir un.

Une fois le modèle sélectionné, l'interpolation peut être effectuée en n'importe quels points. Le krigeage est souvent réalisé sur une grille régulière afin d'obtenir une cartographie de la variable d'étude. Il est également possible de cartographier l'incertitude associée aux valeurs interpolées grâce à la variance de krigeage, puisqu'elle représente la dispersion possible de la valeur réelle et inconnue autour de la valeur obtenue par krigeage. Les faibles valeurs indiquent une interpolation de bonne qualité, inversement pour les valeurs fortes, qui peuvent ainsi traduire des zones sous-échantillonnées.

Il existe de nombreux packages du logiciel `R` pour le traitement de données spatiales disponibles sur le serveur du CRAN (Comprehensive R Archive Network). Les méthodes de krigegae sont par exemple implémentées dans les packages `geoR` et `gstat`.

Bibliographie

[1] Cressie, N. (1993) *Statistics for Spatial Data*, Revised Edition, John Wiley & Sons, New York.