

5. Géostatistique

JEAN-MICHEL FLOCH

Insee

5.1	Fonctions aléatoires	116
5.1.1	Définitions	116
5.1.2	Stationnarité	117
5.2	Variabilité spatiale	118
5.2.1	Covariance et corrélogramme	118
5.2.2	Variogramme	119
5.2.3	Application empirique	120
5.3	Ajustement du variogramme	124
5.3.1	Allure générale du variogramme	124
5.3.2	Les variogrammes usuels	126
5.3.3	L'ajustement du variogramme	127
5.4	Le krigeage ordinaire	130
5.4.1	Principe	130
5.4.2	Application aux données de précipitations	132
5.5	Support et changement de support	138
5.5.1	Variance de dispersion empirique et relations d'additivité de Krige	138
5.5.2	Variogramme de la variable régularisée	139
5.5.3	Krigeage par bloc	140
5.6	Extensions	140
5.6.1	Le cokrigeage	140
5.6.2	Le krigeage universel	144
5.7	Modèles mixtes avec variogramme	144

Résumé

La géostatistique constitue une branche très importante de la statistique spatiale. Développée à partir de préoccupations très pratiques (recherche minière), elle a fait l'objet, sous l'impulsion de Georges Matheron et de ses collègues de l'école des Mines de Fontainebleau, de très importants développements méthodologiques. Les illustrations les plus simples concernent des problèmes tels que l'interpolation des températures ou des précipitations. Mais les travaux les plus importants concernent les applications géologiques et minières que l'on peut trouver par exemple chez Chilès (CHILES et al. 2009). L'application à des exemples démographiques ou sociaux est plus difficile, mais il semble important de présenter les grandes lignes de la méthode : traitement de la stationnarité avec l'introduction de la stationnarité intrinsèque ; introduction du semi-variogramme pour l'étude des relations spatiales ; interpolation des données par la méthode du krigeage. Au delà des applications minières, l'analyse variographique peut être utilisée dans des modèles mixtes, pour analyser les résidus.

- R** La lecture préalable du chapitre 1 : "Analyse spatiale descriptive" et du chapitre 4 : "Les configurations de points" est recommandée.

La modélisation des données spatiales est rendue difficile du fait que l'on n'observe qu'une seule réalisation du phénomène. Dans le cas des données ponctuelles, on n'observe également qu'une seule réalisation, pour laquelle on dispose de toutes les données. Pour les données continues (potentiellement observables en tout point de l'espace), on ne dispose que de données partielles, à partir desquelles on peut être amené à prédire des valeurs aux points non observés. C'est ce manque d'information qui va conduire à l'utilisation de modèles probabilistes.

L'aléatoire n'est pas une propriété du phénomène, mais une caractéristique du modèle utilisé pour le décrire. La géostatistique, qui étudie les phénomènes continus, a permis le développement de méthodes spécifiques pour étudier les relations spatiales entre les observations et construire des outils prédictifs.

La géostatistique doit son nom aux origines minières de la discipline (Krigé, Matheron). Nombre des concepts fondamentaux de la discipline sont issus des travaux de Georges Matheron (*variable régionalisée, fonction aléatoire, hypothèse intrinsèque, effet de pépite*¹, MATHERON et al. 1965). La figure 5.1 présente un résumé illustrant le cheminement de la réalité vers un modèle plus abstrait, modèle qui permettra lui-même d'agir d'une façon que l'on espère optimale (CHAUVET 2008).

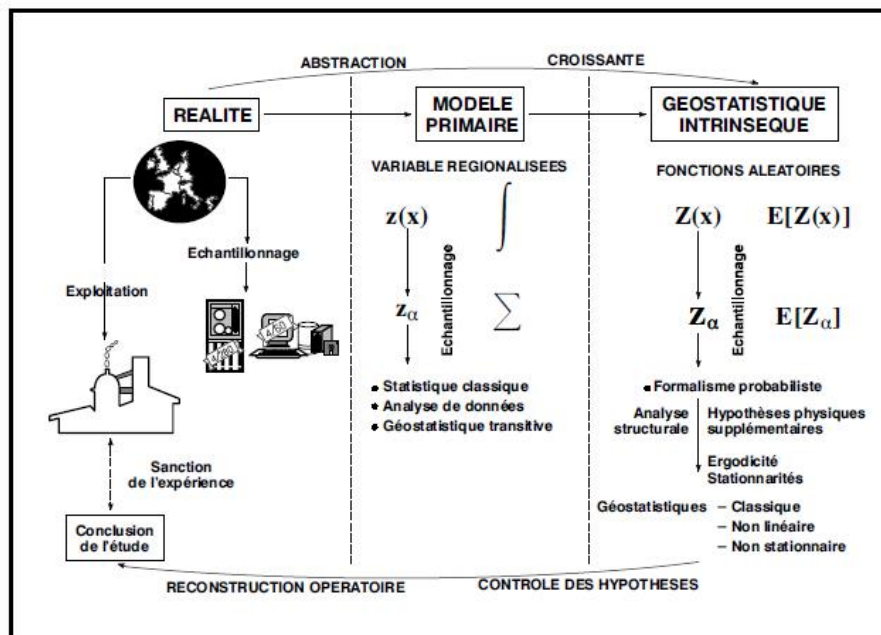


FIGURE 5.1 – Schéma d'une analyse géostatistique

Source : CHAUVET 2008

5.1 Fonctions aléatoires

5.1.1 Définitions

Comme dans la figure 5.1, on note $z(s)$ la *variable régionalisée*, et $Z(s)$ la *fonction aléatoire*, la lettre s désignant la position dans l'espace. On conservera ici cette formulation spécifique à la

1. Ces termes sont définis plus loin.

géostatistique. Un phénomène qui se déploie dans l'espace est qualifié de régionalisé. Une variable régionalisée est une fonction qui décrit de manière satisfaisante ce phénomène. C'est un premier niveau d'abstraction, où l'on reste dans la description, sans recourir à un modèle probabiliste. Si on ne fait pas d'hypothèse supplémentaire, on reste dans le cadre de la *géostatistique transitive*.

L'étape ultérieure, qualifiée de *géostatistique intrinsèque*, introduit la notion de fonction aléatoire. Elle résulte d'un choix, celui de considérer la variable régionalisée comme la réalisation d'une fonction aléatoire. Ce choix permet d'utiliser des outils probabilistes puissants, la contrepartie étant un éloignement de la réalité. Le modèle probabiliste est un intermédiaire de calcul dont on attend qu'il rende des services pour la compréhension du phénomène régionalisé.

La fonction aléatoire est caractérisée complètement par la donnée de sa fonction de répartition

$$F(s_1, s_2, \dots, s_n; z_1, z_2, \dots, z_n) = P\{Z(s_1) \leq z_1, Z(s_2) \leq z_2, \dots, Z(s_n) \leq z_n\}. \quad (5.1)$$

Comme on ne dispose que d'une seule réalisation de notre phénomène régionalisé, il faut trouver une autre façon de faire de l'inférence. Citant MATHERON et al. 1965 : " Pour que l'inférence soit possible, il est nécessaire d'introduire des hypothèses supplémentaires sur la fonction aléatoire $Z(s)$ de façon à réduire le nombre des paramètres dont dépend sa loi. Tel est le but de l'hypothèse stationnaire que nous allons définir : une fonction stationnaire se répète en quelque sorte elle-même dans l'espace, et cette répétition rend à nouveau possible l'inférence statistique à partir d'une réalisation unique." On va de fait traiter chaque observation comme la réalisation d'une variable aléatoire.

5.1.2 Stationnarité

Trois acceptions de la stationnarité sont utilisées en géostatistique :

- la stationnarité stricte ;
- la stationnarité au second ordre ;
- la stationnarité intrinsèque.

Définition 5.1.1 — Stationnarité stricte. La stationnarité stricte renvoie directement à la loi de probabilité du processus. Il y a stationnarité stricte si en se déplaçant par translation, toutes les caractéristiques de la fonction aléatoire restent les mêmes.

Formellement la distribution jointe des $Z(s_i)$ est la même que celle des $Z(s_i + h)$, h indiquant une translation par rapport à la position initiale. Cette forme de stationnarité n'est pas opérationnelle, et très restrictive.

Définition 5.1.2 — Stationnarité du second ordre. La stationnarité au second ordre ou stationnarité faible n'impose plus de conditions sur la loi de probabilité, mais seulement sur la moyenne et la covariance. Ces indicateurs doivent être invariants par translation.

Sachant que l'on décompose $Z(s)$ en une composante déterministe et une composante aléatoire $Z(s) = m(s) + R(s)$

la stationnarité au second ordre requiert les conditions suivantes :

- $\mathbb{E}[Z(s)] = m(s) \forall s$.

L'invariance de l'espérance par translation entraîne la constance de la composante déterministe.

$$m(s+h) = m(s) = m \forall s;$$

- La variance est constante : $\mathbb{E}[(Z(s) - m)^2] = \sigma^2$;

- La covariance ne dépend que du décalage spatial :

$$\text{Cov}[Z(s+h), Z(s)] = \mathbb{E}[(Z(s+h) - m)(Z(s) - m)] = C(h).$$

En pratique, cette hypothèse de stationnarité s'avère souvent trop forte. La limite la plus importante vient de ce que la moyenne peut changer sur le territoire d'intérêt, et que la variance peut ne pas être bornée lorsque cette aire d'intérêt croît. C'est Georges Matheron qui a tiré les conséquences de ces limites de la stationnarité faible en proposant la notion, encore plus faible, de la stationnarité intrinsèque (MATHERON et al. 1965).

Définition 5.1.3 — Stationnarité intrinsèque. L'hypothèse de la stationnarité intrinsèque est la suivante : $\mathbb{E}[(Z(s+h) - Z(s))^2] = 0$.

Les accroissements peuvent être stationnaires sans que le processus lui-même le soit.

On peut alors définir une nouvelle fonction, appelée *variogramme*, fondée sur les différences entre valeurs et valeurs décalées, et qui ne dépend que du décalage :

$$\gamma(h) = \frac{1}{2} \mathbb{E}[Z(s+h) - Z(s)]^2 \quad (5.2)$$

La stationnarité à l'ordre 2 entraîne la stationnarité intrinsèque, mais l'inverse n'est pas vrai. Une fonction aléatoire peut permettre le calcul d'un variogramme sans qu'il en soit de même pour la covariance et la fonction d'autocorrélation.

5.2 Variabilité spatiale

5.2.1 Covariance et corrélogramme

Définition 5.2.1 — Covariance. La fonction de covariance va permettre de prendre en compte les relations entre l'ensemble des paires de points. Si on prend en compte deux points s_i et s_j , la covariance peut être définie par l'équation 5.3.

$$\text{Cov}[Z(s_i), Z(s_j)] = \mathbb{E}[(Z(s_i) - m)(Z(s_j) - m)] \quad (5.3)$$

Lorsque le processus est stationnaire au second ordre, la covariance ne va plus dépendre que de la distance entre les points $|s_i - s_j|$. Si on note h cette distance, on va définir $C(h)$ calculée pour toutes les valeurs de h en prenant en compte tous les couples de points situés à une distance h les uns des autres. Cette fonction de covariance $C(h)$ est définie par l'équation 5.4.

$$C(h) = \text{Cov}[Z(s+h), Z(s)] = \mathbb{E}[(Z(s+h) - m)(Z(s) - m)] \quad (5.4)$$

Elle traduit la façon dont évoluent la covariance des observations lorsque leur distance augmente. Lorsque h est égal à 0, la covariance est égale à la variance.

$$C(0) = \mathbb{E}[(Z(s) - m)^2] = \sigma^2 \quad (5.5)$$

Les propriétés de la fonction de covariance sont les suivantes :

$$C(-h) = C(h) \quad (5.6)$$

$$|C(h)| \leq C(0). \quad (5.7)$$

Pour que la fonction de covariance soit dite *admissible*, il faut que la variance d'une combinaison linéaire de variables soit positive : $\text{Var}[\sum_{i=1}^n \lambda_i z(s_i)] = \sum_{i=1}^n \sum_{j=1}^n C(s_i - s_j)$.

Cela entraîne que C soit *semi-définie positive*.

Définition 5.2.2 — Fonction d'autocorrélation. On définit la fonction d'autocorrélation $\rho(h)$ comme une fonction de h par le rapport $\frac{C(h)}{C(0)}$. Sa valeur est comprise entre -1 et +1. On peut montrer les relations suivantes lorsque la stationnarité à l'ordre 2 est vérifiée :

$$\begin{aligned}\gamma(h) &= C(0) - C(h) \\ \gamma(h) &= \sigma^2 (1 - \rho(h)).\end{aligned}\tag{5.8}$$

Encadré 5.2.1 — Estimation de la fonction de covariance. La fonction de covariance est estimée à partir des $n(h)$ paires de points, comme définies ci dessous, pour i variant de 1 à $n(h)$.

$$\widehat{C}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(s_i) - m)(z(s_i + h) - m)\tag{5.9}$$

avec $n(h) = \text{Card} \{(s_i, s_j) / |s_i - s_j| \approx h\}$

5.2.2 Variogramme

On trouve dans la littérature les expressions de *variogramme* ou de *semi-variogramme*. Certains auteurs (MATHERON et al. 1965) estiment qu'il faut utiliser le terme de semi-variogramme pour $\gamma(h)$ tel que défini dans l'équation 5.10, le variogramme correspondant à $2\gamma(h)$. C'est le choix que nous faisons dans cet article.

Des trois indicateurs que sont la fonction de covariance, la fonction d'autocorrélation et le variogramme, ce dernier est le plus utilisé dans la mesure où il renvoie à la forme la plus faible de stationnarité et donc aux conditions les moins restrictives sur le comportement local de la moyenne.

$$\gamma(h) = \sigma^2 (1 - \rho(h))\tag{5.10}$$

Les propriétés du variogramme sont les suivantes :

$$\begin{aligned}\gamma(h) &= \gamma(-h) \\ \gamma(0) &= 0 \\ \frac{\gamma(h)}{\|h\|^2} &\rightarrow 0 \quad \text{quand} \quad \|h\| \rightarrow \infty\end{aligned}\tag{5.11}$$

Pour tout ensemble de réels $\{a_1, a_2, \dots, a_m\}$ vérifiant $\sum_{i=1}^m a_i = 0$, on a la propriété suivante :

$$\sum_{i=1}^m \sum_{j=1}^m a_i a_j \gamma(s_i - s_j) \leq 0\tag{5.12}$$

Lorsque le processus est isotrope :

$$\gamma(h) = \gamma(\|h\|)\tag{5.13}$$

Encadré 5.2.2 — Estimation du variogramme expérimental. Un variogramme expérimental peut être estimé à partir des couples de points définis comme précédemment.

$$\widehat{\gamma}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} (z(s_i + h) - z(s_i))^2 \quad (5.14)$$

avec $n(h) = \text{Card} \{ (s_i, s_j) / |s_i - s_j| \approx h \}$

Le variogramme peut être estimé selon différentes directions pour mettre en évidence l'éventuelle anisotropie du phénomène étudié.

5.2.3 Application empirique

La géostatistique avec R

Les packages dédiés à la géostatistique dans ce chapitre sont **gstat** et **geoR**, les plus couramment utilisés. Sur le site du CRAN, de nombreux autres packages sont proposés. On trouvera ci dessous une liste commentée réalisée par Roger Bivand.

Le package **gstat** fournit un large éventail de fonctions pour la géostatistique univariée et multivariée, y compris pour les ensembles de données plus volumineux, tandis que **geoR** et **geoRglm** contiennent des fonctions pour la géostatistique basée sur un modèle. Le diagnostic du variogramme peut être effectué avec **vardiag**. L'interpolation automatisée utilisant **gstat** est disponible dans **automap**. Cette famille de packages est complétée par **intamap** avec des procédures d'interpolation automatique et **psgp**, qui implémente le krigeage de processus gaussien clairsemé. Un large éventail de fonctions similaires se trouve dans le package **fields**. Le package spatial est livré avec la base R et contient plusieurs fonctions principales. Le package **spBayes** est compatible avec les modèles univariés et multivariés gaussiens avec MCMC. **rampes** est un autre ensemble de modélisation géostatistique bayésienne. Le package **geospt** contient des fonctions de base géostatistiques et radiales, y compris la prédiction et la validation croisée. En outre, il inclut des fonctions pour la conception de réseaux d'échantillonnage spatial optimaux basés sur la modélisation géostatistique. **spspan** est un autre package qui offre des fonctions pour optimiser les configurations d'échantillons, en utilisant un recuit spatial simulé. Le package **geostatsp** offre des fonctions de modélisation géostatistique utilisant des objets Raster et SpatialPoints. Les modèles non gaussiens sont adaptés à l'aide de l'INLA, et les modèles géostatistiques gaussiens utilisent l'estimation du maximum de vraisemblance. Le package **FRK** est un outil de modélisation et de prédiction spatiales / spatio-temporelles avec de grands ensembles de données. L'approche, discutée dans Cressie et Johannesson (2008), décompose le champ, et donc la fonction de covariance, en utilisant un ensemble fixe de n fonctions de base, où n est généralement beaucoup plus petit que le nombre de points de données (ou polygones).

Le package RGeostats

RGeostats est un package du langage R. Il a été développé par l'équipe de géostatistique du centre de Géosciences de Mines ParisTech. Il met en œuvre toutes les fonctions géostatistiques disponibles de la librairie (commerciale) Geoslib (écrite en C/C++). Il bénéficie donc de l'expérience accumulée dans le domaine minier, et il est spécialement dédié à de telles applications. Il permet de réaliser toutes les implémentations présentées dans ce chapitre, et de traiter les effets de support.

RGeostats permet à l'utilisateur de R, moyennant son chargement et installation, d'accéder aux fonctionnalités usuelles de la Géostatistique. C'est également une plate-forme qui permet à l'équipe de Géostatistique du Centre de Géosciences de développer des prototypes pour l'application de nouveaux modèles (ex : simulations booléennes, simulations bi-plurigaussiennes) ou nouvelles techniques (écoulement de fluides, simulations du temps de première arrivée en géophysique). Les fonctions sont décrites, mais le code n'est pas accessible, et il y a peu d'exemples.

RGeostats est disponible par simple téléchargement à l'adresse suivante : cg.ensmp.fr/rgeostats.

Analyses exploratoires

L'application proposée utilise les données *Swiss rainfall* (précipitations en Suisse). Cette base de données est très utilisée dans les études spatiales, notamment dans DIGGLE et al. 2003. Elle est fournie dans le package *geoR* de R. Les observations correspondent aux relevés pluviométriques de 467 stations météorologiques de Suisse, et sont effectuées le 8 mai 1986. Il s'agit bien d'une donnée continue, les précipitations pouvant être potentiellement relevées en tout point du territoire. Elle peut donc relever d'une modélisation géostatistique, mais elle est plus simple à appréhender que les données minières. En plus de la pluviométrie, mesurée en millimètres, on trouve des données sur l'altitude des stations météorologiques. Dans le package *geoR*, on dispose de trois bases *Sic* (*Spatial interpolation comparaison*) :

- *Sic.100* : échantillon de 100 observations qui pourront servir à effectuer les interpolations ;
- *Sic.367* : observations non incluses dans l'échantillon qui permettront de comparer les estimations et les observations ;
- *Sic.all* : ensemble.

On utilisera ici les fonctionnalités de *geoR*, mais *gstat* et *RGeostats* fournissent des outils d'analyse.

On trouvera dans la figure 5.2, les données échantillonnées (en vert) et les données témoins, les cercles étant proportionnels à la pluviométrie enregistrée.

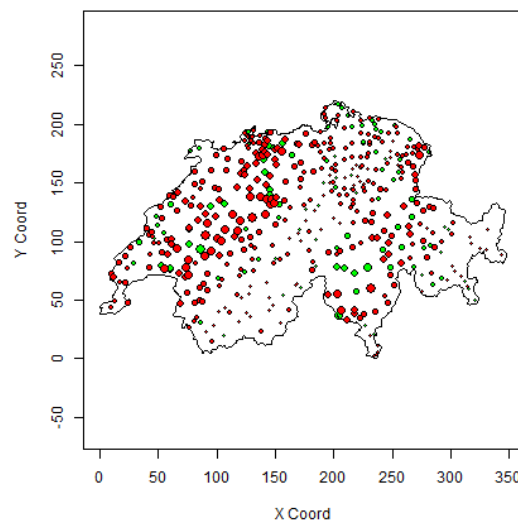


FIGURE 5.2 – Les précipitations en Suisse

Source : *Swiss rainfall* du package *geoR*

```
library(geoR)
points(sic.100, borders=sic.borders,col="green")
points(sic.367, borders=sic.borders,col="red",add=TRUE)
```

Le package *geoR* fournit quelques représentations descriptives grâce à la fonction *plotgeodata*. On trouve dans la figure 5.3, de gauche à droite et de haut en bas :

- la représentation du niveau de la pluviométrie, selon les quantiles de la variable ;
- la pluviométrie en fonction de la latitude ;
- la pluviométrie en fonction de la longitude ;
- l'histogramme des données de pluviométrie.

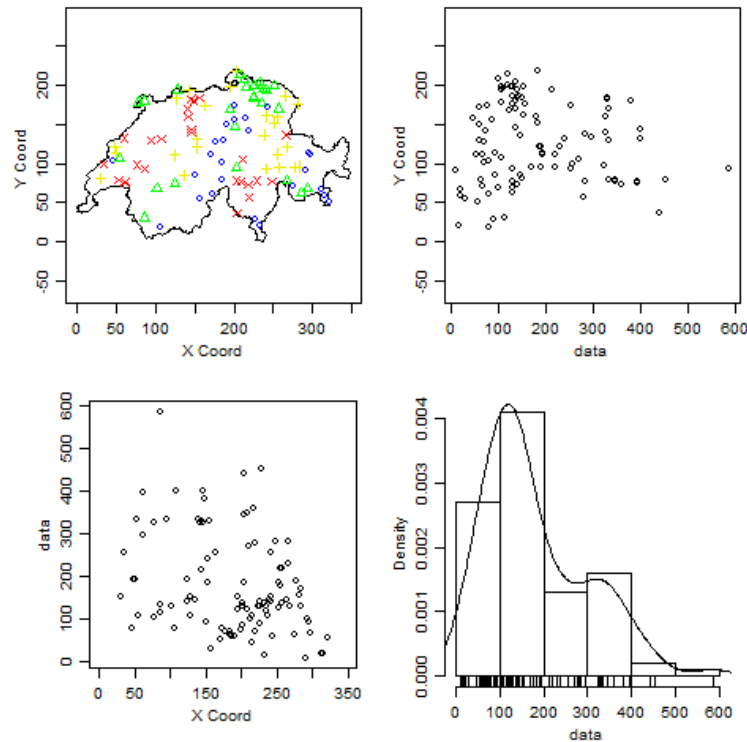


FIGURE 5.3 – Quelques statistiques descriptives

Source : *Swiss rainfall* du package *geoR*

```
library(geoR)
plot.geodata(sic.100,bor=sic.borders)
```

L'histogramme de la figure 5.3 laisse penser que la distribution de la variable n'est pas gaussienne, et qu'une transformation des données pourrait être envisagée puisque les méthodes les plus courantes n'ont des propriétés intéressantes que dans le cadre gaussien.

Nuée variographique et variogramme expérimental

En géostatistique intrinsèque, la **nuée variographique** est un nuage de points des données exprimant leur variabilité selon leurs interdistances. La nuée variographique fournit la représentation graphique des valeurs intervenant dans le calcul du variogramme. Pour un jeu de données de la variable Z aux points $(s_1, \dots, s_i, \dots, s_n)$, il représente les points d'abscisses $\|s_i - s_j\|$ et d'ordonnées $\frac{1}{2} [z(s_i) - z(s_j)]^2$. On peut la représenter sous forme de nuage de points et sous forme de boxplot (voir figure 5.4).

```
library(geoR)
vario.b<- variog(sic.100,option =c ("bin", "cloud", "smooth"),
bin.cloud=TRUE)
vario.c <- variog(sic.100, op="cloud")
bplot.xy(vario.c$u,vario.c$v, breaks=vario.b$u,col="grey80", lwd=2,cex=0.1,
outline=FALSE)
```

Ces représentations étant peu lisibles, la représentation la plus utile est le variogramme expérimental (défini en encadré 5.2.2), présenté en figure 5.6 à partir d'un schéma de construction présenté en figure 5.5.

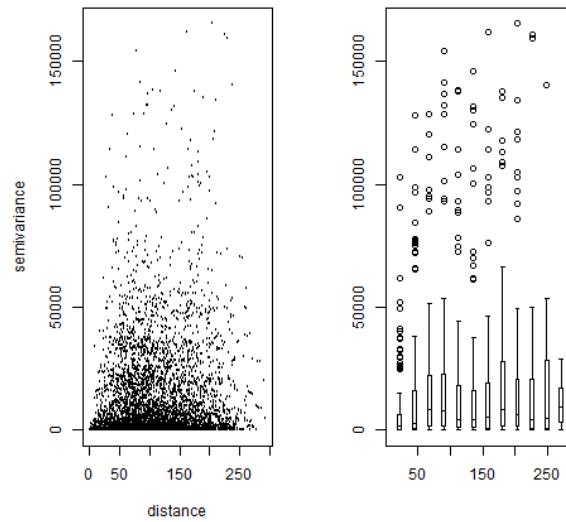


FIGURE 5.4 – Nuée variographique

Source : *Swiss rainfall du package geoR*

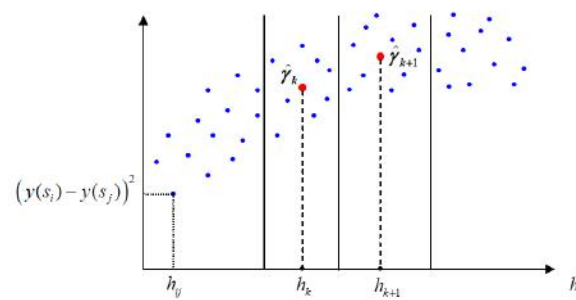


FIGURE 5.5 – Variogramme expérimental : schéma de construction

Source : *Swiss rainfall du package geoR*

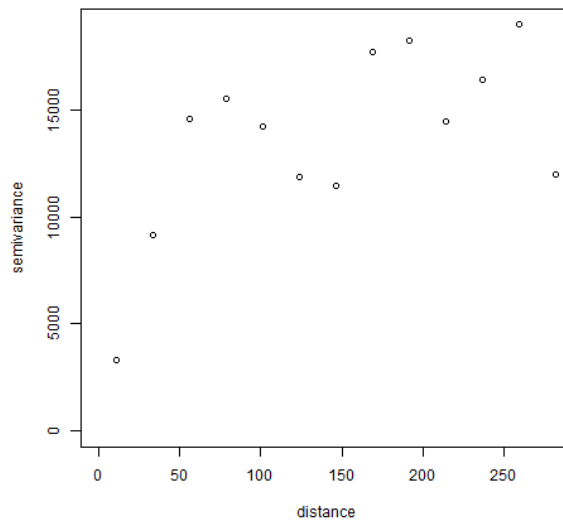


FIGURE 5.6 – Variogramme expérimental
 Source : *Swiss rainfall* du package *geoR*

```
library(geoR)
vario.ex<- variog(sic.100, bin.cloud=TRUE)
plot(vario.ex)
```

Dans la figure 5.6, tous les points observés sont pris en compte pour le calcul du variogramme. Mais les phénomènes étudiés ne sont pas forcément isotropes, et il peut être utile de calculer des variogrammes selon plusieurs directions de l'espace (voir figure 5.7).

```
library(geoR)
vario4<-variog4(sic.100)
plot(vario4,same=FALSE)
```

5.3 Ajustement du variogramme

On verra dans la partie 5.4 consacrée au krigeage que la valeur des estimateurs dépend des observations et de la structure d'autocorrélation spatiale, appréhendée par le variogramme. L'analyse variographique n'est donc pas qu'un point de passage. Elle constitue le point central de la démarche géostatistique. Les variogrammes empiriques présentés dans la partie 5.2 ne sont pas directement utilisables car ils ne respectent pas les propriétés énoncées dans la partie 5.2.2. Pour être utilisés dans les modèles géostatistiques, il faut au préalable les ajuster à des modèles théoriques ayant des formes analytiques bien définies, ce qui implique une vision de ce que doit être un semi-variogramme.

5.3.1 Allure générale du variogramme

On commencera par présenter le modèle le plus classique, à partir duquel seront construits les variogrammes théoriques qui permettront de définir entre autres les équations de krigeage (voir la partie 5.4).

Ce variogramme a une forme qui est d'abord croissante, jusqu'à un certain palier. La valeur de h correspondant à ce palier est appelée la *portée*. On le comprend en se référant à la relation entre

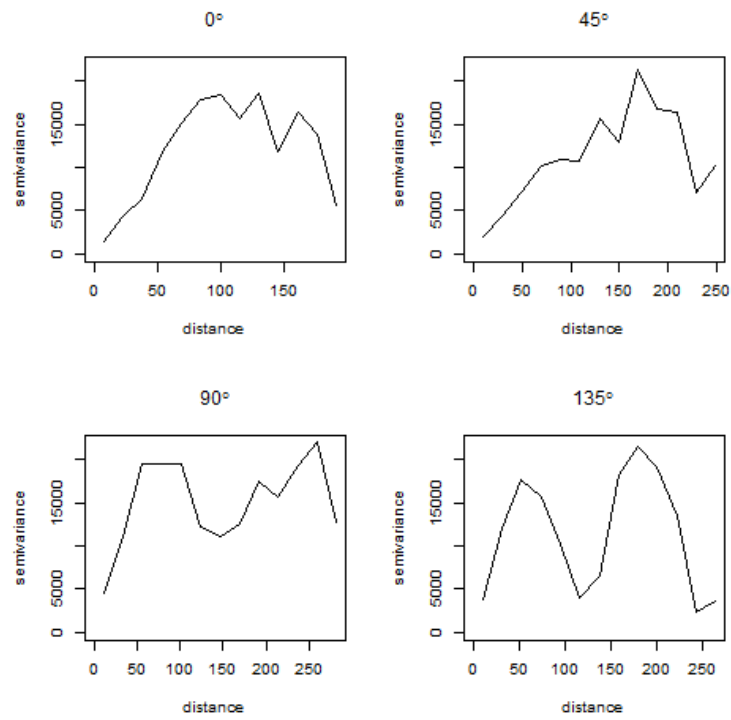


FIGURE 5.7 – Variogrammes directionnels
Source : *Swiss rainfall* du package *geoR*

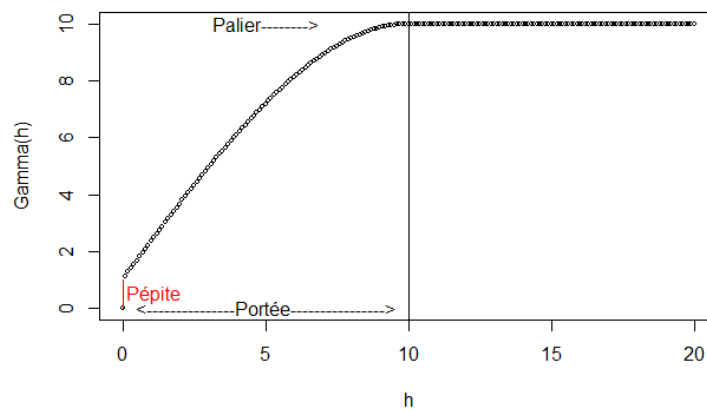


FIGURE 5.8 – Variogramme théorique

la covariance, lorsque celle-ci est définie et le semi-variogramme, à savoir $\gamma(h) = C(0) - C(h)$. La covariance est très fréquemment une fonction décroissante de la distance, ce qui implique la croissance du semi-variogramme, mais ce n'est pas toujours le cas (modèle sinus-cardinal par exemple).

À une certaine distance, qui correspond à la portée, la covariance va s'annuler. Cette portée est la portée de la dépendance spatiale. Il n'y a plus de relation entre les valeurs observées à une distance observée au-delà de cette portée. Pour le semi-variogramme cela signifie qu'au-delà de cette portée, sa valeur est constante, et que l'on a : $C = C(0) = \sigma^2$.

Pour $h = 0$, la valeur du variogramme est par définition nulle. Mais on constate en pratique que pour des valeurs très proches de 0, le variogramme prend des valeurs supérieures à 0 et qu'il y a de fait une discontinuité à l'origine. On appellera *pépite* la limite du variogramme en zéro. Comme l'explique Matheron (MATHERON et al. 1965) : "La notion d'échelle joue ici un rôle primordial. À l'échelle de la dizaine de mètres, un phénomène de transition dont la portée est centimétrique ne se manifeste sur $\gamma(h)$ que comme une discontinuité à l'origine, c'est à dire un effet de pépite." Elle représente la variation entre deux mesures effectuées à des emplacements infiniment proches, et peut donc provenir de deux effets :

- une variabilité de l'instrument de mesure : la pépite mesure donc en partie l'erreur statistique de l'instrument de mesure
- un réel effet pépite : une variation brutale du paramètre mesuré ; le cas historique est le passage sans transition d'une pépite d'or à un sol ne contenant quasiment pas d'or.

On peut rencontrer d'autres formes de variogramme. La figure 5.9 montre deux cas classiques. Le premier est celui du **variogramme linéaire**, le second celui du **pur effet de pépite**. Lorsque le variogramme n'est pas borné, la moyenne et la variance ne sont pas définies : cela indique le plus souvent une tendance à grande échelle, qu'il sera nécessaire de modéliser. L'effet de pépite pur traduit l'absence de dépendance spatiale.

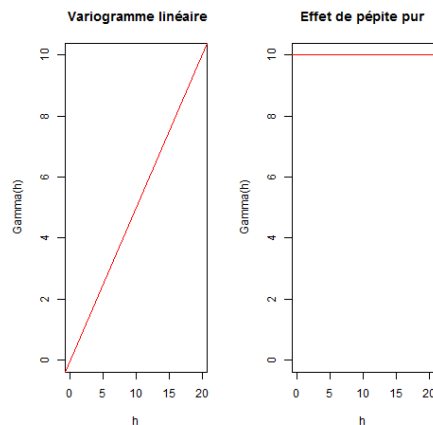


FIGURE 5.9 – Deux semi-variogrammes atypiques

5.3.2 Les variogrammes usuels

La littérature géostatistique propose de nombreuses fonctions qui satisfont les propriétés du semi-variogramme tel qu'il est présenté dans la figure 5.8. Ces fonctions paramétrées doivent permettre de décrire les différentes composantes (portée, palier, pépite). Elles doivent aussi gérer le comportement de la fonction à l'origine (tendance linéaire, tangence horizontale ou verticale).

On ne présentera ici que quatre exemples de modèles de variogramme (figure 5.10), les autres

étant décrits dans les ouvrages de référence (ARMSTRONG 1998, CHILES et al. 2009, WALLER et al. 2004).

Définition 5.3.1 — Le modèle sphérique.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s \left[\frac{3}{2} \left(\frac{h}{a} \right) - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right] & 0 < h \leq a \\ c_0 + c_s & h > a \end{cases} \quad (5.15)$$

Définition 5.3.2 — Le modèle exponentiel.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s [1 - \exp(-\frac{h}{a})] & h > 0 \end{cases} \quad (5.16)$$

Définition 5.3.3 — Le modèle gaussien.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_s [1 - \exp(-(\frac{h}{a})^2)] & h > 0 \end{cases} \quad (5.17)$$

Définition 5.3.4 — Le modèle puissance.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + bh^p & h > 0 \end{cases} \quad (5.18)$$

Définition 5.3.5 — Modèle de Matern.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_s \left[1 - \frac{\frac{h}{a}}{2^{\alpha-1}\Gamma(\alpha)} K_{\alpha} \left(\frac{h}{a} \right) \right] & h > 0 \end{cases} \quad (5.19)$$

où Γ désigne la fonction gamma et K_{α} la fonction de Bessel modifiée de seconde espèce de paramètre α .

Définition 5.3.6 — Modèle sinus cardinal.

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_s \left[1 - \frac{a}{h} \sin \left(\frac{h}{a} \right) \right] & h > 0 \end{cases} \quad (5.20)$$

Comme l'expliquent Matheron et les géostatisticiens de l'école des Mines de Fontainebleau, la modélisation est une affaire de choix. Le choix du modèle théorique est un moment décisif dans la démarche du géostatisticien, mais on ne peut pas associer *a priori* un variogramme théorique à tel ou tel type de processus. Il faut tenir compte à la fois de la connaissance empirique du phénomène, de l'allure du variogramme expérimental obtenu. S'agissant de l'effet de pépite, on peut par exemple penser qu'il n'est pas adapté à des données comme les précipitations.

5.3.3 L'ajustement du variogramme

Il faut ensuite trouver une forme fonctionnelle adaptée au variogramme expérimental. Une première étape importante est d'obtenir une représentation lissée du variogramme, qui peut être obtenue avec la fonction *variog* de *geoR*. Comme toujours, ce lissage est tributaire du choix de la fenêtre. La représentation lissée est considérée parfois comme suffisante pour estimer "à l'oeil" le

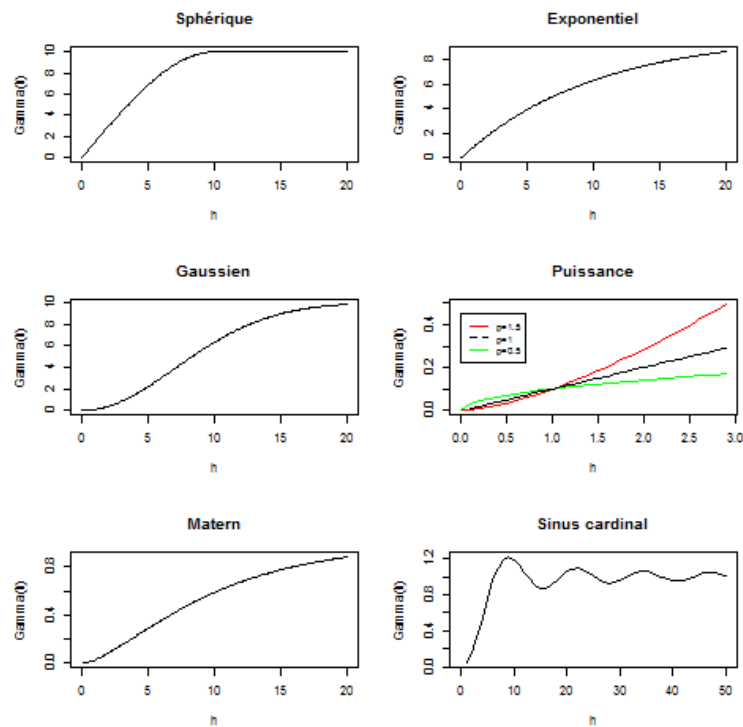


FIGURE 5.10 – Quatre exemples de variogrammes théoriques

variogramme. Cette approche est critiquée comme trop empirique par de nombreux géostatisticiens mais cette étape peut donner quelques indications, notamment sur le comportement à l'origine.

On trouvera dans la figure 5.11 trois exemples d'ajustement du variogramme expérimental des données sur les précipitations en Suisse, par un variogramme sphérique, un variogramme exponentiel sans pépite et un variogramme exponentiel avec pépite. Ces ajustements sont effectués à l'aide de la fonction `lines.variomodel` de *geoR* (RIBEIRO JR et al. 2006). C'est une première approche "à l'oeil". On voit que si le variogramme exponentiel avec pépite s'ajuste apparemment mieux aux données que le variogramme sans pépite, l'introduction de cet effet à très courte distance n'a pas de justification physique dans le cas des précipitations.

```
library(geoR)
vario.ex<- variog(sic.100,option="bin")
vario.sphe<-(variofit(vario.ex,cov.model= "spher",
ini.cov.pars=c(15000,200)))
par(mfrow=c(2,2), mar=c(3,3,1,1), mgp =c (2,1,0))
plot(vario.ex,main="Sphérique")
lines.variomodel(cov.model="sphe",cov.pars=c(15000,100),
nug=0,max.dist=350)
plot(vario.ex,main="Exponentiel")
lines.variomodel(cov.model="exp",cov.pars=c(15000,100),
nug=0,max.dist=350)
plot(vario.ex,main="Exponentiel avec pépite")
lines.variomodel(cov.model="exp",cov.pars=c(10000,100),
nug=5000,max.dist=350)
```

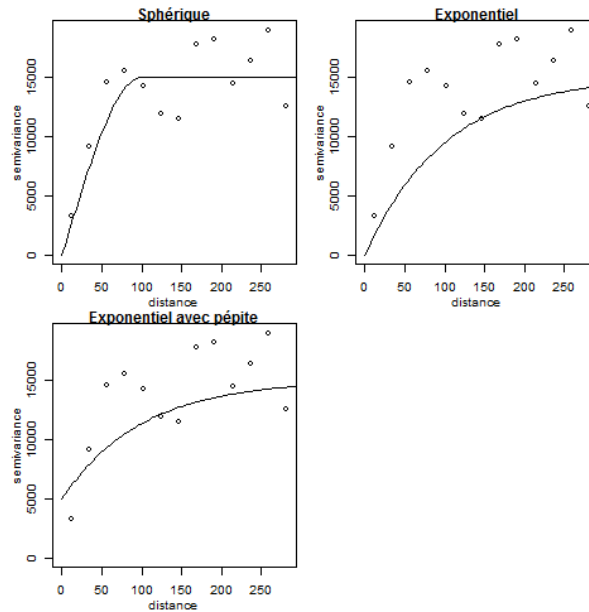


FIGURE 5.11 – Trois exemples d’ajustement du variogramme expérimental
Source : *Swiss rainfall* du package *geoR*

```
plot(vario.ex,main="Exponentiel avec pépité")
lines.variomodel(cov.model="matern",cov.pars=c(10000,100),
nug=0,max.dist=350,kappa=0.5)
```

Le choix est donc un compromis comme le rappellent WALLER et al. 2004 : "Même si un modèle particulier est jugé meilleur pour un ensemble de données particulier par une méthode statistique d’ajustement, ce n’est peut-être pas forcément le meilleur choix. Par exemple, le modèle gaussien est souvent sélectionné par un critère d’ajustement automatique, mais cela fournit un lissage qui apparaît souvent irréaliste. En fin de compte, le choix final du modèle devrait refléter à la fois le résultat de la procédure d’ajustement au modèle statistique et une interprétation cohérente avec la compréhension scientifique du processus à l’étude".

De nombreuses méthodes sont proposées pour ajuster le variogramme. On trouve des méthodes fondées sur les moindres carrés, ordinaires ou pondérés, des méthodes fondées sur la vraisemblance, ainsi que des méthodes bayésiennes. Dans *geoR*, les fonctions utilisées sont `variofit` (moindres carrés) et `likfit` (maximum de vraisemblance). Les méthodes sont assez techniques et nécessiteraient des développements importants. On peut renvoyer à RIBEIRO JR et al. 2006 pour une illustration de ces méthodes à partir de données simulées. Le code R est fourni dans l’article.

Ajustement par les moindres carrés ordinaires (MCO)

On cherche le vecteur des paramètres de la fonction qui minimise une fonction d’objectif simple, la somme des carrés des distances entre la valeur du semi-variogramme expérimental et celle du variogramme théorique.

$$\hat{\theta}_{MCO} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^k (\hat{\gamma}(h_i) - \gamma(h_i; \theta))^2 \quad (5.21)$$

Ajustement par les moindres carrés pondérés (MCP)

Les moindres carrés ordinaires ne tiennent pas compte du nombre de couples de points qui interviennent dans le calcul de chacun des points du variogramme expérimental, contrairement aux moindres carrés pondérés.

$$\hat{\theta}_{MCP} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^k \frac{\#N(h_i)}{\gamma(h_i; \theta)^2} (\hat{\gamma}(h_i) - \gamma(h_i; \theta))^2 \quad (5.22)$$

Ajustement par les moindres carrés généralisés (MCG)

D'autres auteurs ont proposé les moindres carrés généralisés pour prendre en compte l'hétéroscédasticité.

$$\hat{\theta}_{MCG} = \underset{\theta \in \Theta}{\operatorname{argmin}} (\hat{\gamma}_n - \gamma(\theta))^T \operatorname{Cov}(\gamma_n)^{-1} (\hat{\gamma}_n - \gamma(\theta)) \quad (5.23)$$

où γ est le vecteur $(\gamma_1, \gamma_2, \dots, \gamma_K)$.

Ajustement par le maximum de vraisemblance (MV)

Les paramètres du modèle sont estimés en calculant la vraisemblance. Dans le cas non gaussien, les estimations manquent de robustesse. Les calculs sont lourds, et il faut réserver cette méthode à de petits échantillons. De plus, cette méthode requiert la stationnarité au second ordre et ne peut s'appliquer à des variogrammes non bornés. Dans ce dernier cas, il faut utiliser les moindres carrés pondérés.

5.4 Le krigeage ordinaire

5.4.1 Principe

Le terme de krigeage est dû à Georges Matheron, et fait référence aux travaux pionniers de Danie Krige, ingénieur sud-africain. Le krigeage est une méthode d'interpolation très puissante. Les exemples fournis ici sont très élémentaires. Les applications à la recherche minière ou géologique fournissant de nombreux exemples où l'on s'intéresse à l'estimation de volumes et pas seulement à des interpolations simples.

On ne présentera pas ici le krigeage *simple*, qui suppose connue la valeur de la moyenne, mais le krigeage qualifié d'*ordinaire*, qui constitue le point d'orgue de la géostatistique. Dans le krigeage ordinaire, la valeur moyenne n'est pas connue. On peut en trouver des utilisations simples dans l'interpolation des températures (JOLY et al. 2009) ou dans des études sur la qualité de l'air (LLOYD et al. 2004).

Supposons que $Z(\cdot)$ est intrinsèquement stationnaire, que son variogramme $\gamma(h)$ est connu mais que sa moyenne m est inconnue. On dispose d'un ensemble de données $Z = [Z(s_1), \dots, Z(s_i), \dots, Z(s_N)]^t$. On veut prédire la valeur de $Z(\cdot)$ en un point inobservé et calculer $Z(s_0)$. L'estimateur du krigeage ordinaire va être défini comme une combinaison linéaire des observations.

$$Z_{OK}(s_0) = \sum_{i=1}^N \lambda_i Z(s_i) \quad (5.24)$$

Les valeurs des poids λ_i ne sont pas calculées à partir d'une fonction de la distance mais en utilisant le semi-variogramme et deux critères statistiques : l'absence de biais et la minimisation de

l'erreur quadratique moyenne de prédiction. L'absence de biais implique que l'on ait l'équation 5.25 :

$$\mathbb{E} [\hat{Z}_{OK}(s_0)] = \mathbb{E} [Z(s_0)] = m \quad (5.25)$$

$$\mathbb{E} [\hat{Z}_{OK}(s_0)] = \sum_{i=1}^N \lambda_i \mathbb{E} [Z(s_i)] = \sum_{i=1}^N \lambda_i m \quad \rightarrow \quad \sum_{i=1}^N \lambda_i = 1$$

On va donc, en utilisant la méthode des multiplicateurs de Lagrange, minimiser $\mathbb{E} [\hat{Z}_{OK}(s_0) - Z(s_0)]^2$ sous la contrainte $\sum_{i=1}^N \lambda_i = 1$.

On trouvera dans l'encadré 5.4.1 la façon d'introduire le variogramme et d'aboutir aux équations de krigeage.

$$\begin{aligned} \sum_{j=1}^N \lambda_j \gamma(s_i - s_j) + m &= \gamma(s_0 - s_i) \\ \sum_{i=1}^N \lambda_i &= 1 \end{aligned} \quad (5.26)$$

La valeur de $\hat{Z}_{OK}(s_0)$ est déterminée par des points qui dépendent de la corrélation entre le point d'estimation et les points d'observation, mais aussi des corrélations entre les points d'observation. On convient en général d'écrire ces équations de krigeage sous forme matricielle :

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ m \end{bmatrix} = \begin{bmatrix} \gamma(s_1 - s_1) & \dots & \gamma(s_1 - s_N) & 1 \\ \gamma(s_2 - s_1) & \dots & \gamma(s_2 - s_N) & 1 \\ \vdots & \dots & \vdots & \vdots \\ \gamma(s_N - s_1) & \dots & \gamma(s_N - s_N) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(s_0 - s_1) \\ \gamma(s_0 - s_2) \\ \vdots \\ \gamma(s_0 - s_N) \\ 1 \end{bmatrix} \quad (5.27)$$

ou de façon plus ramassée :

$$\lambda_0 = \Gamma_{ij}^{-1} \gamma_0. \quad (5.28)$$

La matrice Γ ne dépend pas du point d'estimation, et n'a donc pas à être recalculée à chaque fois. Les valeurs de tous les $\gamma(s_i - s_j)$ et $\gamma(s_0 - s_i)$ sont calculées à partir des valeurs du variogramme estimé. L'erreur quadratique moyenne de prédiction, connue sous le nom de *variance de krigeage*, est égale à $\lambda_0^t \gamma_0$. En résumé, le krigeage fournit un estimateur sans biais, de variance minimale, qui est aussi un interpolateur exact puisqu'il redonne pour chaque point connu une valeur estimée égale à la valeur observée.

Encadré 5.4.1 — Estimation des équations de krigeage. En utilisant la méthode des multiplicateurs de Lagrange, on minimise :

$$\mathbb{E} [\hat{Z}_{OK}(s_0) - Z(s_0)]^2 \text{ sous la contrainte } \sum_{i=1}^N \lambda_i = 1.$$

On va chercher les $\lambda_1, \dots, \lambda_N$ et le multiplicateur m qui permettent d'introduire la contrainte. La fonction d'objectif s'écrit donc :

$$\mathbb{E} \left[\left(\sum_{i=1}^N \lambda_i Z(s_i) - Z(s_0) \right)^2 \right] - 2m \left(\sum_{i=1}^N \lambda_i - 1 \right). \quad (5.29)$$

Du fait de la contrainte, on peut écrire :

$$\left[\sum_{i=1}^N \lambda_i Z(s_i) - Z(s_0) \right]^2 = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N [(Z(s_i) - Z(s_j))^2] + \sum_{i=1}^N [(Z(s_i) - Z(s_0))^2]. \quad (5.30)$$

En prenant l'espérance des expressions, on a :

$$\mathbb{E} \left[\sum_{i=1}^N \lambda_i Z(s_i) - Z(s_0) \right]^2 = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \mathbb{E} [(Z(s_i) - Z(s_j))^2] + \sum_{i=1}^N \lambda_i \mathbb{E} [(Z(s_i) - Z(s_0))^2]. \quad (5.31)$$

Cette expression permet de faire apparaître le variogramme et on peut réécrire la contrainte comme :

$$-\sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \gamma(s_i - s_j) + 2 \sum_{i=1}^N \gamma(s_0 - s_i) - 2m \left(\sum_{i=1}^N \lambda_i - 1 \right). \quad (5.32)$$

On va minimiser cette expression en dérivant par rapport à $\lambda_1, \dots, \lambda_N$ et m , ce qui conduit aux équations de krigeage :

$$\begin{aligned} \sum_{j=1}^N \lambda_j \gamma(s_i - s_j) + m &= \gamma(s_0 - s_i) \\ \sum_{i=1}^N \lambda_i &= 1. \end{aligned} \quad (5.33)$$

5.4.2 Application aux données de précipitations

Données brutes

Un premier krigeage a été réalisé à partir des données brutes en utilisant un modèle sphérique pour le variogramme. On trouve en figure 5.12 le variogramme sphérique utilisé, après estimation des paramètres par le maximum de vraisemblance comparé au variogramme expérimental.

```
library(geoR)
vario.ex <- variog(sic.100, bin.cloud=TRUE)
plot(vario.ex, main="")
lines.variomodel(cov.model="spher", cov.pars=c(15000, 50),
nug=0, max.dist=300)
```

On peut ainsi calculer des valeurs krigées sur un carroyage, ainsi que les variances de krigeages qui sont représentées dans la figure 5.13. Les valeurs sont représentées selon les conventions de sémiologie graphique, les couleurs chaudes correspondant aux valeurs élevées.

```
library(geoR)
pred.grid <- expand.grid(seq(0, 350, l=51), seq(0, 220, l=51))
rgb.palette <- colorRampPalette(c("blue", "lightblue",
"orange", "red"), space = "rgb")
kc <- krige.conv(sic.100, loc = pred.grid,
krige=krige.control(cov.model="spherical", cov.pars=c(15000, 50)))
```

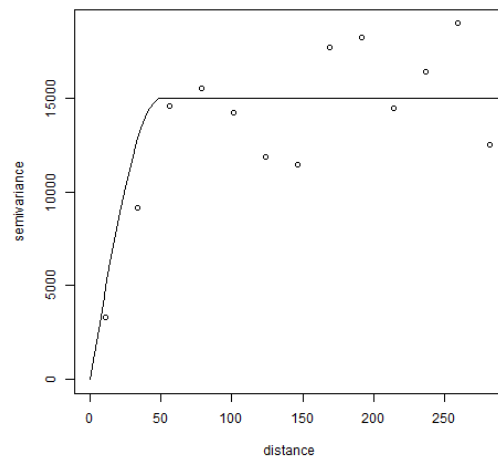


FIGURE 5.12 – Variogramme sphérique pour les données brutes
Source : *Swiss rainfall* du package *geoR*

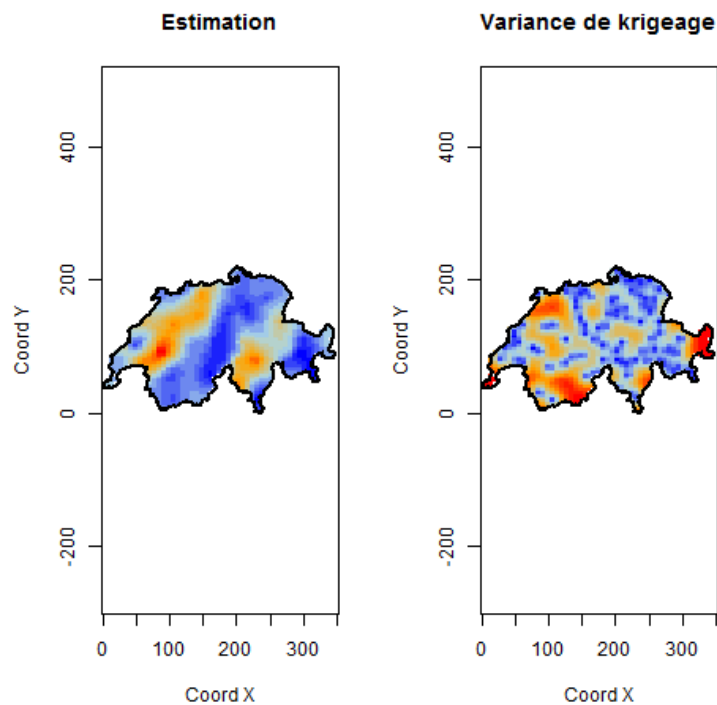


FIGURE 5.13 – Estimations et variances de krigeage pour les données brutes
Source : *Swiss rainfall* du package *geoR*

```

image(kc, loc = pred.grid,col =rgb.palette(20) ,xlab="Coord X",
ylab="Coord Y",borders=sic.borders,main="Estimation")
image(kc, krige.var,loc = pred.grid,col=rgb.palette(20),
xlab="Coord X",ylab="Coord Y",borders=sic.borders,
main="Variance de krigeage")

```

- L'estimation a été effectuée en faisant intervenir les 100 points de l'échantillon. On peut :
- vérifier que le krigeage est sans biais sur les points observés;
 - mesurer les écarts entre valeurs estimées et valeurs observées.

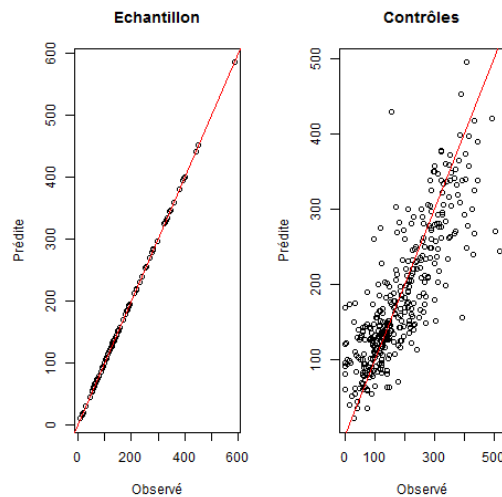


FIGURE 5.14 – Valeurs estimées et observées
Source : *Swiss rainfall* du package *geoR*

```

library(geoR)
kc1<- krige.conv(sic.100, loc = sic.100$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47)))
kc2<- krige.conv(sic.100, loc = sic.367$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47)))
plot(sic.100$data,kc1$predict,xlab="Observ\'e",ylab="Pr\'edite",
main="Echantillon")
abline(a=0,b=1,col="red")
plot(sic.367$data,kc2$predict, ,xlab="Observ\'e",ylab="Pr\'edite",
main="Contr\^oles")
abline(a=0,b=1,col="red")

```

Données après transformations

L'histogramme de la figure 5.3 indiquait que la distribution des données de précipitations s'écartait d'une distribution gaussienne. Une première possibilité, classique en statistique, est la modification des variables. En effet, il peut être avantageux de travailler avec des données qui suivent une loi normale. Ce n'est pas absolument requis dans le modèle de krigeage, mais l'hypothèse de linéarité du krigeage n'est vraiment performante que lorsque les données sont gaussiennes. Les transformations classiques sont les transformations logarithmiques, ou de façon plus générale les

transformations de Box-Cox. La fonction `boxcoxfit` de *geoR* suggère dans le cas des données un coefficient proche de 0.5. Les données exploratoires sont représentées dans la figure 5.15.

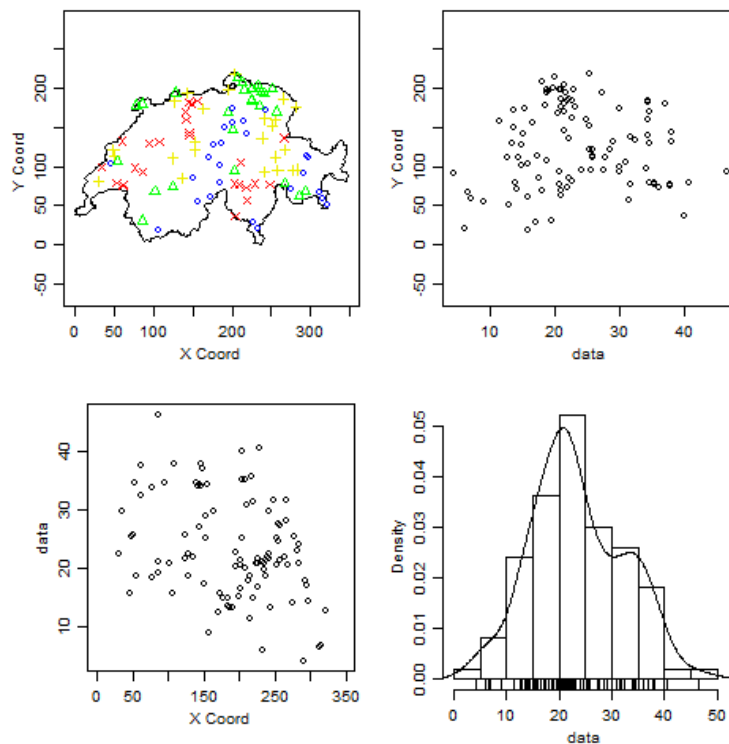


FIGURE 5.15 – Données exploratoires, après transformation Box-Cox

Source : *Swiss rainfall* du package *geoR*

```
library(geoR)
plot.geodata(sic.100, bor=sic.borders, lambda=0.5)
```

Les données sur les précipitations ont souvent été étudiées. RIBEIRO JR et al. 2004 préconisent la même transformation des variables. Pour le variogramme, ils proposent l'utilisation d'un modèle de Matern pour lequel $K = 1$. Les paramètres du modèle sont déterminés à l'aide du maximum de vraisemblance. Les variogrammes expérimentaux et théoriques pour les données transformées sont présentés en figure 5.16.

```
library(geoR)
vario.ext<- vario(sic.100, option="bin", lambda=0.5)
plot(vario.ext)
lines.variomodel(cov.m = "mat", cov.p = c(105, 36), nug = 6.9,
                max.dist = 300, kappa = 1, lty = 1)
```

Comme pour les données brutes, on peut fournir une cartographie des estimations et des valeurs de krigeage (figure 5.17).

```
library(geoR)
kct<- krige.conv(sic.100, loc = pred.grid,
krige=krige.control(cov.model="matern", cov.pars=c(105, 36),
kappa=1, nugget=6.9, lambda=0.5))
```

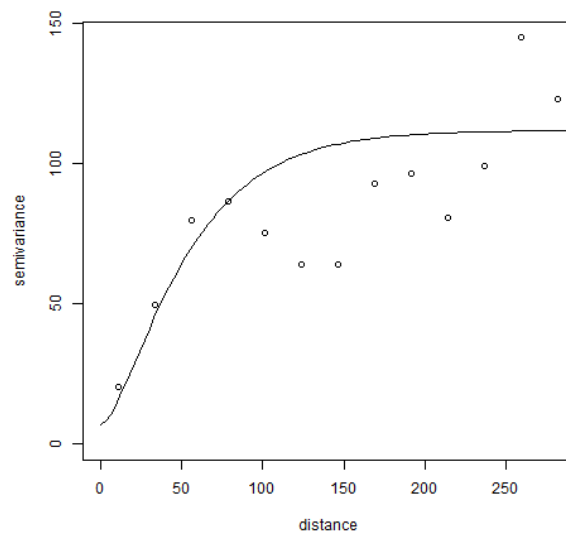


FIGURE 5.16 – Variogramme expérimental et variogramme théorique après transformation
Source : *Swiss rainfall du package geoR*

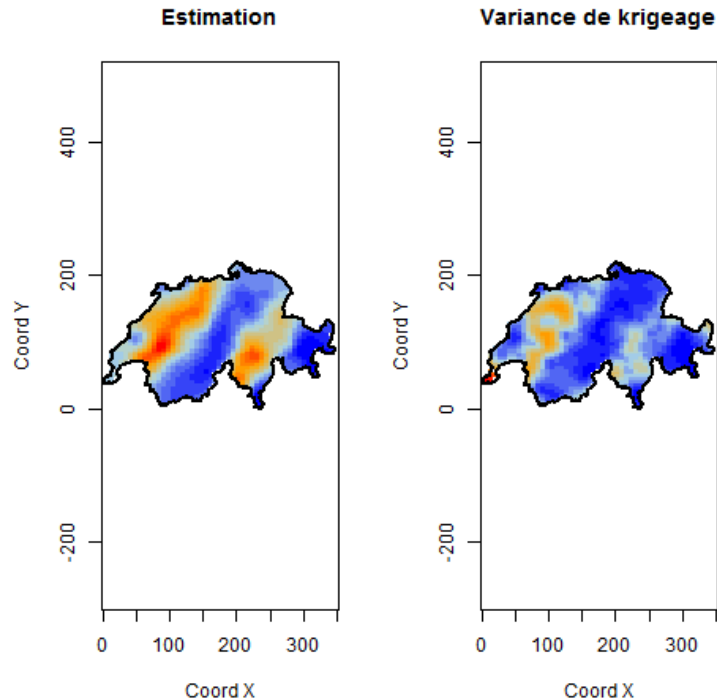


FIGURE 5.17 – Estimations et variance de krigeage des précipitations en Suisse après transformation
Source : *Swiss rainfall du package geoR*

```

pred.grid <- expand.grid(seq(0,350, l=51),seq (0,220, l=51))
rgb.palette <- colorRampPalette(c("blue", "lightblue",
"orange", "red"),space = "rgb")
image(kct, loc = pred.grid,col =rgb.palette(20) , xlab="Coord X",
ylab="Coord Y",borders=sic.borders,main="Estimation")
image(kct, krige.var,loc = pred.grid,col =rgb.palette(20) ,
xlab="Coord X",ylab="Coord Y",borders=sic.borders,
main="Variance de krigeage")

```

On peut comparer les valeurs estimées et les valeurs observées (5.18).

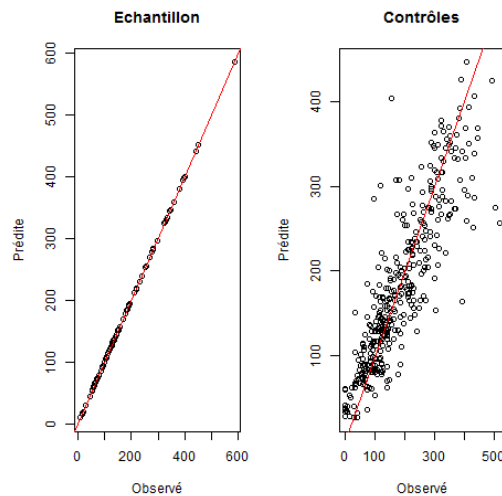


FIGURE 5.18 – Valeurs estimées et observées

Source : *Swiss rainfall* du package *geoR*

```

library(geoR)
kct1<- krige.conv(sic.100, loc = sic.100$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47),
kappa=1,nugget=6.9,lambda=0.5))
kct2<- krige.conv(sic.100, loc = sic.367$coords,
krige=krige.control(cov.model="spherical",cov.pars=c(16000,47),
kappa=1,nugget=6.9,lambda=0.5))
plot(sic.100$data,kct1$predict,xlab="Observ\'e",ylab="Pr\'edite",
main="Echantillon")
abline(a=0,b=1,col="red")
plot(sic.367$data,kct2$predict,,xlab="Observ\'e",ylab="Pr\'edite",
main="Contrôles")
abline(a=0,b=1,col="red")

```

Si on prend comme critère la racine carrée de l'écart quadratique moyen

$$RMSE(y) = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}},$$

on constate une amélioration de la prédiction si on utilise les données transformées modélisées par

le modèle de Matern. Dans le cas des données brutes la RMSE vaut 62.3, dans le cas des données transformées 55.2.

5.5 Support et changement de support

La question des échelles d'analyse est de première importance en analyse spatiale. Les géographes, depuis Openshaw notamment utilisent le terme de *Modifiable areal unit problem* (MAUP), présenté dans le chapitre 1 : "Analyse spatiale descriptive". On peut résumer le MAUP comme la superposition d'un effet de zonage et d'un effet d'agrégation. On peut l'illustrer de façon simple à partir de données surfaciques, en montrant la variabilité des résultats obtenus selon le découpage territorial utilisé.

En géostatistique, on parle de *change of support problem* (COSP). Ce problème est de fait plus général que le MAUP, puisqu'il renvoie à la taille, à la forme et aux orientations. Le COSP en géostatistique prend son origine dans des préoccupations très pratiques, liées à la recherche minière dont on peut trouver les origines dans les travaux de Krige, que développera Matheron. Les articles de Krige sont contemporains de ceux de Yule et Kendall en statistique classique, qui anticipent ceux d'Openshaw en géographie. Dans une perspective pratique, les géostatisticiens avaient rapidement perçu qu'il était plus important de prédire une valeur sur un bloc important que sur un point, quand bien même la première prédiction dérivait de la seconde.

Le support peut être un point, un bloc plus ou moins gros, une réunion de points dans une configuration géométrique donnée. Il y a des relations entre les valeurs prises sur des volumes différents. Dans le cadre de variables additives, pour un volume V partitionné en unités v_i de même support v (V étant un multiple de v) on a :

$$z(V) = \frac{1}{n} \sum_{i=1}^n z(v_i) \quad (5.34)$$

ou si v est ponctuel :

$$z(V) = \frac{1}{V} \int_V z(x) dx. \quad (5.35)$$

$Z(V)$ est qualifiée de variable *régularisée*, parce qu'elle accroît la régularité statistique. En fait, c'est une forme particulière de variable régularisée, la forme plus générale étant une *convoluée de Z* (CHILES et al. 2009).

5.5.1 Variance de dispersion empirique et relations d'additivité de Krige

Définition 5.5.1 — Variance de dispersion empirique.

$$s^2(v|V) = \frac{1}{n} \sum_{i=1}^n [z(v_i) - z(V)]^2 \quad (5.36)$$

Si V est inclus dans un domaine plus large noté D , on peut démontrer la relation 5.37, appelée **relation d'additivité de Krige** :

$$s^2(v|D) = s^2(v|V) + s^2(V|D). \quad (5.37)$$

ARMSTRONG 1998 donne un exemple pédagogique à partir des rendements en millet sur 16 blocs de 2m de côté, partagés en 64 parcelles de 1m de côté. La moyenne vaut 201 dans le cas des blocs comme dans le cas des parcelles. La variance de dispersion des blocs dans le champ vaut 16.64, celle des parcelles 27.59, ce qui veut dire que la dispersion des parcelles dans les blocs vaut 10.85 (voir table 5.1).

735	325	45	140	125	175	167	485
540	420	260	128	20	30	105	70
450	200	337	190	95	260	245	278
180	250	380	405	250	80	515	605
124	120	430	175	230	120	460	260
40	135	240	35	190	135	160	170
75	95	20	35	32	95	20	450
200	35	100	59	2	45	58	90

505	143	88	207
270	328	171	411
102	220	154	263
101	54	44	155

TABLE 5.1 – Valeurs observées sur les parcelles (en haut) et les blocs (en bas)

Source : ARMSTRONG 1998.

5.5.2 Variogramme de la variable régularisée

On peut définir la variance par bloc à partir des informations sur les données ponctuelles (fonction de covariance).

$$\text{Var}[Z(V)] = \bar{C}(V, V) = \frac{1}{|V|^2} \int_V \int_V C(x-y) dx dy \quad (5.38)$$

où C désigne la fonction de covariance pour les données ponctuelles, et \bar{C} la covariance des données par bloc. La covariance s'écrit :

$$\text{Cov}[Z(V), Z(V')] = \bar{C}(V, V') = \frac{1}{|V||V'|} \int_V \int_{V'} C(x-y) dx dy. \quad (5.39)$$

La fonction de covariance C_V est définie en introduisant V_h qui est le translaté du support V par le vecteur h :

$$C_V(h) = \text{Cov}[Z(V), Z(V_h)] = \bar{C}(V, V_h). \quad (5.40)$$

On peut également déduire du variogramme ponctuel le variogramme de la variable régularisée :

$$\gamma_V(h) = \bar{\gamma}(V, V_h) - \bar{\gamma}(V, V) \quad (5.41)$$

avec $\gamma(V, V) = \frac{1}{|V|^2} \int_V \int_V \gamma(x-y) dx dy$ et $\bar{\gamma}(V, V_h) = \frac{1}{|V|^2} \int_V \int_{V_h} \gamma(x-y) dx dy$ où γ est le variogramme calculé à partir des observations ponctuelles.

On peut alors montrer que $\gamma_V(h) \sim \gamma(h) - \bar{\gamma}(V, V)$ ce qui se traduit graphiquement par la figure 5.19.

Pour passer du variogramme ponctuel au variogramme régularisé, on conserve le même type de modèle théorique en corrigeant le palier et la portée.

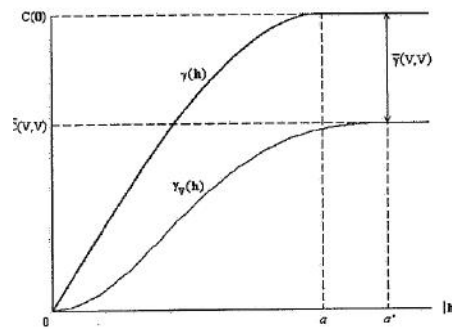


FIGURE 5.19 – Passage du variogramme ponctuel au variogramme régularisé

5.5.3 Krigeage par bloc

On peut déterminer des équations de krigeage, dans la logique du krigeage ordinaire. De $\mathbb{E}[Z(x)] = m$, on en déduit que $\mathbb{E}[Z_V] = m$.

On cherche comme dans le cas du krigeage ordinaire un estimateur qui soit une combinaison linéaire des observations collectées.

$$Z_V = \sum_{i=1}^n \lambda_i Z(x_i) \quad (5.42)$$

La démarche est la même que celle qui est exposée en encadré 5.4.1 (minimisation sous contrainte) et on aboutit aux équations de krigeage :

$$\begin{aligned} \sum_{j=1}^n \lambda_j \gamma(x_i, x_j) + \mu &= \gamma(x_i, V) \quad \text{pour } i = 1, 2, \dots, n \\ \sum_{i=1}^n \lambda_i &= 1 \end{aligned} \quad (5.43)$$

$$\text{avec } \bar{\gamma}(x_i, V) = \frac{1}{|V|} \int_V \gamma(x_i - x) dx.$$

5.6 Extensions

Le krigeage ordinaire constitue la méthode de base de la géostatistique. Mais elle n'en constitue au final qu'une partie, et de nombreux développements ont vu le jour, notamment à l'école des Mines de Fontainebleau.

Les hypothèses de stationnarité intrinsèque restent assez restrictives, notamment la constance de la moyenne. De nombreuses méthodes ont été élaborées pour permettre d'introduire des hypothèses moins contraignantes, ou pour utiliser des informations auxiliaires. Dans ce paragraphe, on présentera quelques modifications et variantes, à partir essentiellement des données suisses sur les précipitations, et d'un autre jeu de données fréquemment mis à contribution, sur les teneurs en divers minéraux dans une boucle de la Meuse.

5.6.1 Le cokrigeage

La géostatistique a développé depuis longtemps des méthodes multivariées. L'une d'entre elle est le cokrigeage, qui va prendre en compte plusieurs variables. La définition qu'en donne WALLER et al. 2004 est la suivante :

Définition 5.6.1 — cokrigeage. "Le cokrigeage est une extension du krigeage au cas de deux variables spatiales ou plus. Il a été développé à l'origine comme une technique pour améliorer la prédiction d'une variable pour laquelle seuls quelques échantillons pourraient être prélevés, en utilisant sa corrélation spatiale avec d'autres variables plus facilement mesurées. Le cokrigeage diffère du krigeage avec dérive externe en ce sens que les variables explicatives ne sont pas supposées indiquer la nature d'une tendance dans la variable primaire, mais sont elles-mêmes des variables aléatoires spatiales pour lesquelles on peut utiliser l'analyse variographique".

On peut définir un covariogramme croisé :

$$\gamma_{ZY}(h) = \frac{1}{2p(h)} \sum_{i=1}^{p(h)} (z(s_i) - z(s_i + h))(y(s_i) - y(s_i + h)) \quad (5.44)$$

avec $p(h) = \text{Card} \{(s_i, s_j) \mid |s_i - s_j| \approx h\}$

Comme pour le krigeage, il y aura plusieurs versions pour le cokrigeage. On ne présentera que le cokrigeage ordinaire. On se limitera au cas où l'on n'introduit qu'une variable auxiliaire, que l'on notera Y . L'estimateur que l'on calcule est de la forme :

$$Z(s_0) = \sum_{i=1}^{n_Z} \lambda_i Z(s_i) + \sum_{i=1}^{n_Y} \alpha_i Y(s_i) \quad (5.45)$$

avec les contraintes d'absence de biais :

$$\begin{aligned} \sum_{i=1}^{n_Z} \lambda_i &= 1 \\ \sum_{i=1}^{n_Y} \alpha_i &= 0. \end{aligned} \quad (5.46)$$

Sous forme matricielle, les équations de cokrigeage s'écrivent :

$$\begin{bmatrix} \Gamma_{ZZ} & \Gamma_{ZY} & 1 & 0 \\ \Gamma_{YZ} & \Gamma_{YY} & 0 & 1 \\ 1' & 0' & 0 & 0 \\ 0 & 1' & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \alpha \\ \mu_Z \\ \mu_Y \end{bmatrix} = \begin{bmatrix} \gamma_{ZZ} \\ \gamma_{YZ} \\ 1 \\ 0 \end{bmatrix} \quad (5.47)$$

On illustrera cette méthode à l'aide des données *Meuse* fournies dans le package *sp* et étudiées notamment par Pebesma (PEBESMA 2001) et Rossiter (ROSSITER 2017). Ce dernier fournit sur son site des programmes R.

■ **Exemple 5.1** Analyse par cokrigeage des données *Meuse*

Les données *Meuse* fournissent des mesures localisées de teneur en plomb, zinc, cadmium, mais aussi d'autres variables comme l'altitude, la teneur en matière organique du sol. Le package *sp* contient une table nommée *Meuse*, pouvant être chargée avec la fonction : `data(meuse)`. Il fournit également une grille de 40×40 m : *meuse.grid* et les limites du département : *meuse.riv*.

Les données sont constituées de 155 observations sur des supports de 15×15 m, sur les 20 cm supérieurs des sols alluviaux de la rive droite de la Meuse. Les données associées fournissent les coordonnées géographiques des observations, leur altitude, et les concentrations en cadmium, cuivre, plomb, zinc, matière organique. On trouve aussi la distance à la Meuse et la fréquence

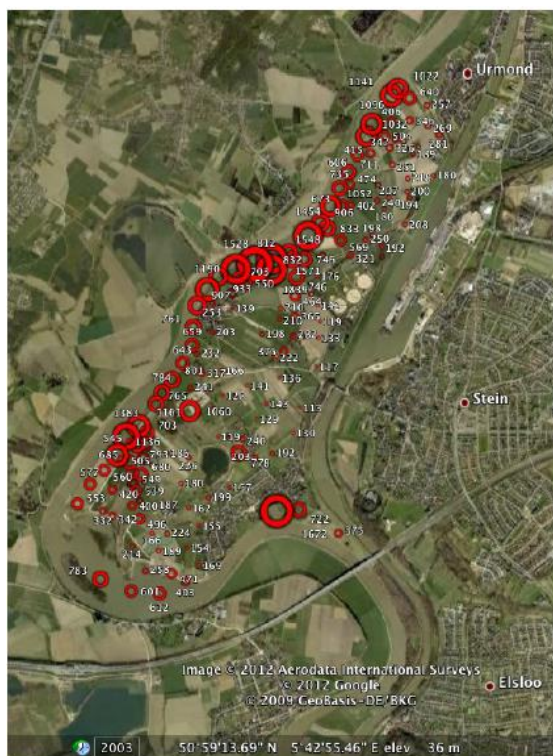


FIGURE 5.20 – Géographie et position des prélèvements

Source : Données Meuse du package *sp*

des inondations. Dans l'exemple fourni par ROSSITER 2017, c'est la teneur en plomb qui est étudiée (après une transformation logarithmique) et la teneur en matière organique qui va servir de covariable.

L'analyse variographique dans le cas du cokrigage repose sur l'étude des deux variogrammes simples et du variogramme croisé.

Le même exercice est effectué en prenant comme covariable le logarithme de la teneur en zinc. La figure ci-dessous présente les résultats de krigeage de la teneur en plomb, par krigeage ordinaire et par cokrigage en utilisant les deux variables mentionnées ci-dessus (teneur en zinc et matière organique). On trouve dans la figure 5.22 les valeurs krigées (colonne de gauche) et les valeurs résiduelles (colonne de droite). On trouve successivement le krigeage ordinaire, le cokrigage avec la matière organique comme covariable, puis le cokrigage avec la teneur en zinc.

Le code nécessaire pour ce traitement est assez long. Le lien ci-dessous permet d'accéder au programme R mis à disposition par Rossiter (ROSSITER 2007) :

http://www.css.cornell.edu/faculty/dgr2/teach/R/ck_plotfns.R.

Si l'on compare, en utilisant la RMSE, les trois modélisations, on trouve les résultats suivants :

- 0.166 pour le krigeage ordinaire ;
- 0.226 pour le cokrigage avec la teneur en matière organique ;
- 0.078 pour le cokrigage avec la teneur en zinc.

Le cokrigage faisant intervenir la teneur en zinc améliore donc les performances du krigeage ordinaire. Ce n'est pas le cas pour le cokrigage avec la teneur en matière organique. ■

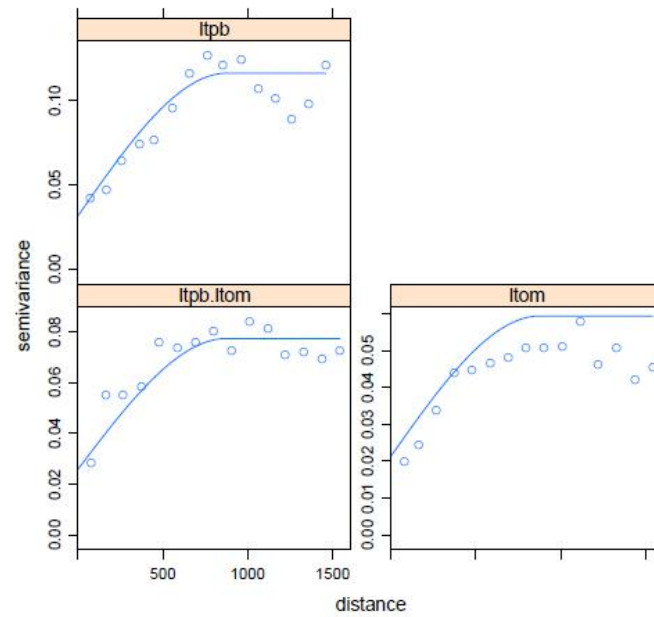


FIGURE 5.21 – Variogrammes simples et croisés

Source : Données Meuse du package *sp*

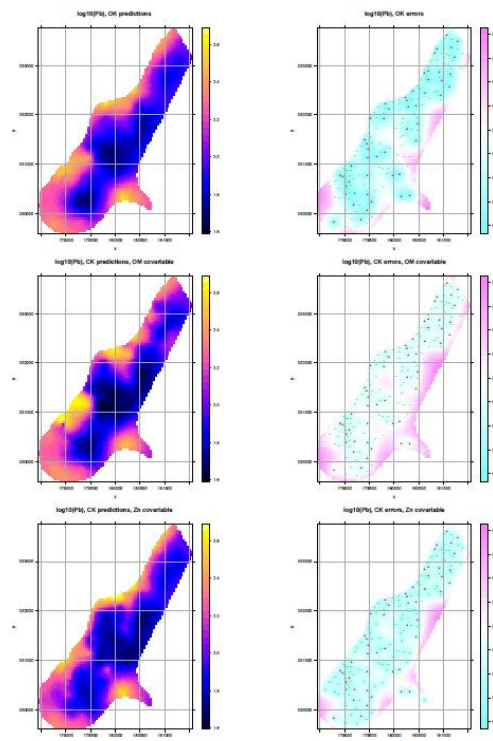


FIGURE 5.22 – Estimation et variance de krigeage et cokrigeage

Source : Données Meuse du package *sp*

5.6.2 Le krigeage universel

Dans de nombreux cas, la valeur moyenne n'est pas constante, et on ne peut pas utiliser le krigeage ordinaire. C'est le cas lorsque l'on observe des relations déterministes entre la valeur de la variable et sa position dans l'espace. La variable régionalisée peut alors s'écrire :

$$Z(s) = m(s) + Y(s) \quad (5.48)$$

où $m(s)$ représente la composante déterministe.

■ **Exemple 5.2 — Analyse par cokrigeage des données Meuse.** Les données *Meuse* fournissent deux variables susceptibles de construire une composante déterministe : la distance à la rivière et la fréquence des inondations (figure 5.23). Elles sont de nature différente des covariables utilisées précédemment dans le cokrigeage.

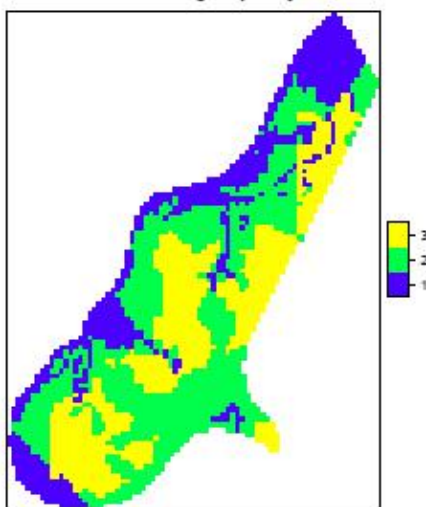


FIGURE 5.23 – Fréquence des inondations

Source : Données Meuse du package *sp*

Sur son site, Rossiter fournit des exemples permettant de comparer les prédictions du krigeage ordinaire et de deux modèles de krigeage universel.

Si l'on compare, en utilisant la RMSE, les trois modélisations, on trouve les résultats suivants :

- 0.173 pour le krigeage ordinaire ;
- 0.141 pour le modèle avec fréquence des inondations ;
- 0.145 pour le modèle avec fréquence des inondations et distance à la rivière.

L'introduction de la fréquence des inondations améliore les prédictions, mais l'introduction en plus de la distance le dégrade (probablement à cause de la corrélation des deux variables). Le code R utilisé est disponible à l'adresse ci-dessous. Il constitue un bon exemple d'illustration et d'approfondissement de ce qui a été présenté dans ce chapitre.

http://www.css.cornell.edu/faculty/dgr2/teach/R/gs_short_ex.pdf ■

5.7 Modèles mixtes avec variogramme

On peut trouver d'autres utilités aux outils développés en géostatistique. Lorsqu'on effectue des régressions sur des données spatialisées, il est fréquent que les résidus soient spatialement autocorrélés. Cette corrélation peut être mise en évidence par l'indicateur de Moran d'autocorrélation spatiale (voir chapitre 3 : "Indices d'autocorrélation spatiale"). La prise en compte de

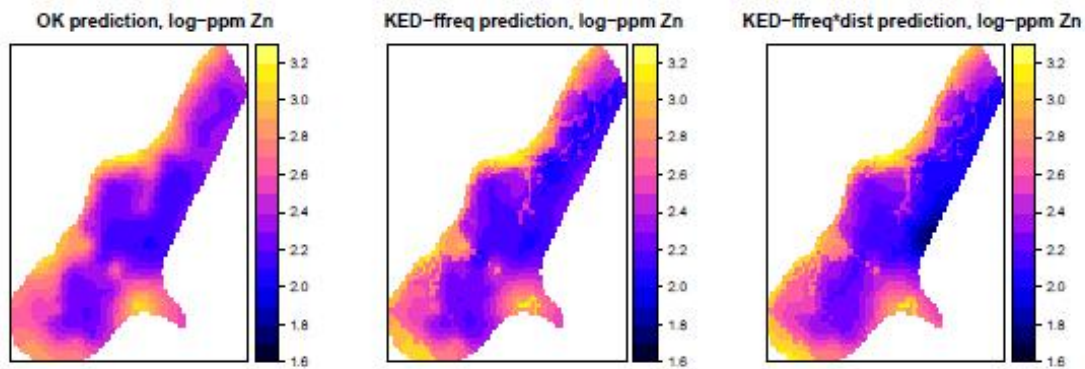


FIGURE 5.24 – Trois estimations

Source : *Données Meuse du package sp*

Note : krigeage ordinaire (à gauche), fréquence des inondations (au milieu), fréquence des inondations et distance à la rivière (à droite)

cette autocorrélation peut se faire en utilisant des modèles économétriques spatiaux (voir chapitre 6 : "économétrie spatiale : modèles courants") ou la régression géographiquement pondérée (voir chapitre 9 : "Régression géographiquement pondérée").

Lorsque les données s'y prêtent, le variogramme peut aussi être utilisé pour étudier la structure spatiale des résidus dans des modèles linéaires. Les exemples sont plus souvent présentés dans des ouvrages traitant d'écologie ou d'épidémiologie. On pourra en trouver des illustrations dans des manuels généraux de statistique spatiale comme WALLER et al. 2004 ou SCHABENBERGER et al. 2017, ainsi que des ouvrages traitant d'écologie, comme PLANT 2012 ou ZUUR et al. 2009, les deux derniers fournissant des exemples d'implémentation en R.

■ **Exemple 5.3 — Analyse de la structure spatiale des résidus avec un variogramme.** Un exemple d'utilisation sur des données écologiques est fournie par ZUUR et al. 2009². L'exemple provient de données collectées sur les forêts de la section Raifa de la biosphère naturelle d'état de Voljsko-Kamsky.

La variable d'intérêt est un indice de "boréalité" (*Bor*) défini comme la part des espèces spécifiquement boréales par rapport au nombre total des espèces sur un site. On dispose aussi de variables explicatives fournies par des images satellitaires :

1. l'indice normalisé de différence de végétation ;
2. la température ;
3. un indice d'humidité ;
4. un indice de verdure.

Du fait de la forte colinéarité entre ces variables, on n'a utilisé que l'humidité pour expliquer l'indice de boréalité. L'analyse de la variance réalisée avec les moindres carrés ordinaires fournit les résultats suivants :

Les coordonnées des sites permettent de fournir une vision exploratoire de la spatialisation des résidus du modèle MCO.

```
library(sp)
library(nlme)
```

2. https://github.com/James-Thorson/2016_class_CMR/tree/master/Other%20material/Zuur%20et%20al.%202007/ZuurDataMixedModelling

Variable	Valeur estimée	écart-type
Constante	27.63	0.981
Wet	429.609	27.45

TABLE 5.2 – Estimation par les moindres carrés

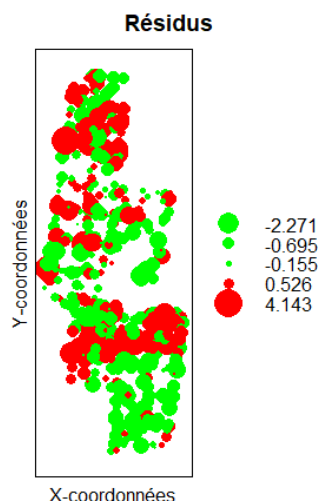


FIGURE 5.25 – Variogramme des résidus

Source : Données Meuse du package *sp*

```
Boreality<-read.table("C:/jmf/Boreality.txt",header=TRUE)
B.lm <- lm(boreal~ Wet, data = Boreality)
E <- rstandard(B.lm)
graphique <- data.frame(E, Boreality$x, Boreality$y)
library(sp)
coordinates(graphique) <- c("Boreality.x","Boreality.y")
bubble(graphique, "E", col= c ("green","red"),main = "R\'esidus",
xlab = "X-coordonn\'ees", ylab = "Y-coordonn\'ees")
```

Le modèle MCO ne permet pas d'introduire de structure spatiale sur les résidus. On ne peut l'introduire que dans un modèle linéaire généralisé. On l'estimera à l'aide de la fonction `gls` du package *nlme* de R. Ce package contient une fonction permettant d'estimer le variogramme. Le variogramme expérimental est représenté ci-dessous (figure 5.26)

```
mod<-gls(boreal~Wet,data=bor)
summary(mod)
plot(Variogram(mod,form=~x+y,maxdist=10000),xlim=c(0,10000))
```

L'estimation initiale du modèle, sans introduction de la structure spatiale, donne les mêmes résultats que les MCO. La commande `update` permet d'introduire une structure spatiale à l'aide d'une analyse variographique, et de réestimer le modèle.

On trouvera dans le tableau 5.3 le résultat de la comparaison entre les MCO et les modèles utilisant les variogrammes sphériques, gaussiens et exponentiels.

Les critères AIC et L montrent qu'on peut améliorer le modèle en utilisant un variogramme

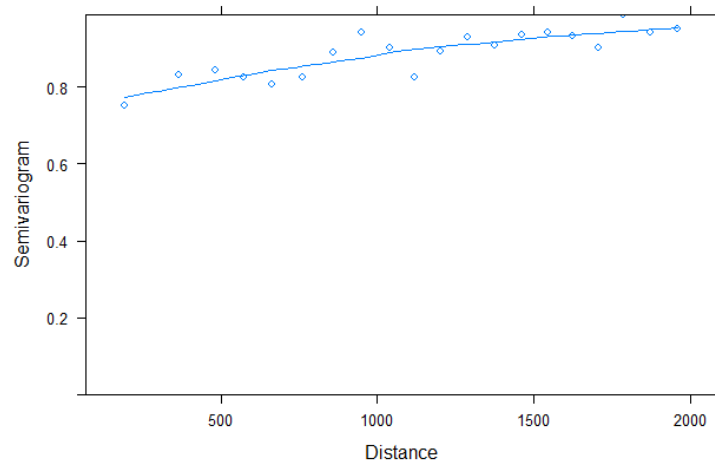


FIGURE 5.26 – Variogramme des résidus
Source : *Données Meuse du package sp*

Variogramme	AIC	Vraisemblance	L	Significativité
MCO	3855	-1924		
Sphérique	3859	-1924	1	1
Gaussien	3750	-1870	109	<0.001
Exponentiel	3740	-1865	119	<0.001

TABLE 5.3 – Critères AIC et L selon le type de variogramme

sphérique pour modéliser les résidus. Avec le variogramme exponentiel, les résultats de la régression sont présentés dans le tableau 5.4.

Variable	Valeur estimée	écart-type
constante	18.099	2.333
Wet	180.247	34.932

TABLE 5.4 – Estimations

```
f1 <- formula(boreal ~ Wet)
B1.gls <- gls(f1, data = Boreality)
Vario.gls <- Variogram(B1.gls, form = ~ x + y, robust = TRUE,
maxDist = 2000, resType = "pearson")
B1A <- gls(f1, correlation = corSpher(form = ~ x + y, nugget = TRUE),
data = Boreality)
B1B <- gls(f1, correlation = corLin(form = ~ x + y, nugget = TRUE),
data = Boreality)
B1C <- gls(f1, correlation = corRatio(form = ~ x + y, nugget = TRUE),
data = Boreality)
B1D <- gls(f1, correlation = corGaus(form = ~ x + y, nugget = TRUE),
data = Boreality)
B1E <- gls(f1, correlation = corExp(form = ~ x + y, nugget = TRUE),
data = Boreality)
AIC(B1.gls, B1A, B1B, B1C, B1D, B1E)
B1 <- lm(f1, data = Boreality)
anova(B1.gls, B1A)
anova(B1.gls, B1D)
anova(B1.gls, B1E)
summary(B1E)
```

Les paramètres du modèle sont significatifs. L'influence de l'humidité est moins marquée lorsqu'on introduit l'autocorrélation spatiale dans le modèle. ■

Conclusion

Le premier chapitre de ce manuel présente les trois grands domaines de la statistique spatiale adaptés à l'analyse des données continues, surfaciques ou ponctuelles. La géostatistique, utilisée pour les données continues, est moins directement liée aux travaux de la statistique publique. Il semblait néanmoins utile d'en faire une présentation rapide dans le manuel. D'un point de vue pédagogique, les méthodes géostatistiques illustrent particulièrement bien comment la prise en compte de l'autocorrélation spatiale (à travers le variogramme) permet d'améliorer les estimateurs. D'un point de vue plus opérationnel, sans rentrer dans la complexité des travaux de recherche minière, la géostatistique, *via* les méthodes de krigeage, est utile pour modéliser des données continues plus simples (données climatiques par exemple). L'école des Mines de Fontainebleau, qui a joué un rôle décisif dans le développement de ces méthodes, a utilisé un langage un peu inhabituel pour les statisticiens, mais des échanges nombreux ont eu lieu, depuis les travaux de Cressie, pour faire le lien entre les approches des uns et des autres. Le livre classique de Chilès et Delfiner en est un bon exemple (CHILES et al. 2009). Dans le domaine de la santé, des travaux importants ont mobilisé

les méthodes géostatistiques pour la modélisation de données épidémiologiques, notamment ceux de Diggle (DIGGLE et al. 2003) davantage connu pour ses travaux sur les méthodes ponctuelles. Pour finir, on ne peut que recommander au statisticien appelé un jour à utiliser des modèles la lecture de l'article de réflexion du fondateur de la géostatistique (MATHERON 1978).

Annexes

Rappels mathématiques

Les expressions des variogrammes théoriques, notamment le variogramme de Matern font appel à des expressions mathématiques peu usuelles, dont on rappelle l'expression ci-dessous (CHILES et al. 2009).

La fonction Gamma

$$\Gamma(x) = \int_0^{\infty} e^{-u} u^{x-1} du$$

Dans le cas de valeurs entières :

$$\Gamma(n+1) = n!$$

Les fonctions de Bessel

La fonction de Bessel du premier type est la suivante :

$$J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k}$$

La fonction de Bessel modifiée du premier type est la suivante :

$$I_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{1}{k! \Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k}.$$

La fonction de Bessel modifiée du deuxième type est définie à partir de la précédente :

$$K_\nu(x) = \frac{\pi}{2} \frac{I_{-\nu}(x) + I_\nu(x)}{\sin \pi \nu}.$$

Références - Chapitre 5

- ARMSTRONG, Margaret (1998). *Basic linear geostatistics*. Springer Science & Business Media.
- CHAUVET, Pierre (2008). *Aide-mémoire de géostatistique linéaire*. Presses des MINES.
- CHILES, Jean-Paul et Pierre DELFINER (2009). *Geostatistics : modeling spatial uncertainty*. T. 497. John Wiley & Sons.
- DIGGLE, Peter J, Paulo J RIBEIRO JR et Ole F CHRISTENSEN (2003). « An introduction to model-based geostatistics ». *Spatial statistics and computational methods*. Springer, p. 43–86.
- JOLY, Daniel et al. (2009). « Interpolation par régressions locales : application aux précipitations en France ». *L'Espace géographique* 38.2, p. 157–170.
- LLOYD, Christopher D et Peter M ATKINSON (2004). « Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data ». *International Journal of Applied Earth Observation and Geoinformation* 5.4, p. 293–305.
- MATHERON, Georges et al. (1965). *Les variables régionalisées et leur estimation*. Masson et Cie.
- MATHERON, Georges (1978). *Estimer et choisir : essai sur la pratique des probabilités*. Ecole nationale supérieure des mines de Paris.
- PEBESMA, Edzer J (2001). « Gstat user's manual ». *Dept. of Physical Geography, Utrecht University, Utrecht, The Netherlands*.
- PLANT, Richard E (2012). *Spatial data analysis in ecology and agriculture using R*. cRc Press.
- RIBEIRO JR, Paulo J et Peter J DIGGLE (2006). « geoR : Package for Geostatistical Data Analysis An illustrative session ». *Artificial Intelligence* 1, p. 1–24.
- RIBEIRO JR, Paulo Justiniano et Peter J DIGGLE (2004). « Model Based Geostatistics ». *Springer Series in Statistics*.
- ROSSITER, David G (2017). « An introduction to geostatistics with R/gstat Version 3.7, 12-May-2017. »
- ROSSITER, DG (2007). « Co-kriging with the gstat package of the R environment for statistical computing ». *Web : [http://www. itc. nl/rossiter/teach/R/R ck. pdf](http://www.itc.nl/rossiter/teach/R/R ck. pdf)*.
- SCHABENBERGER, Oliver et Carol A GOTWAY (2017). *Statistical methods for spatial data analysis*. CRC press.
- WALLER, Lance A et Carol A GOTWAY (2004). *Applied spatial statistics for public health data*. T. 368. John Wiley & Sons.
- ZUUR, AF et al. (2009). « Mixed effects models and extensions in ecology with R. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors ». *New York, NY : Spring Science and Business Media*.