Université de Lille

Travail d'étude et de recherche

Filière : Statistiques

Statistique Spaciale

Réalisé par : **M. Mohamed Ben Yassine**

Sous la direction de : **Pr. Sophie Dabo**

Année universitaire

2019/2020

Table des matières

Remerciements

Tout d'abord, j'adresse mes remerciements à **Mme Sophie DABO** Responsable pédagogique du master MIASHS à l'Université de Lille, m'avoir encadré, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je désire aussi remercier les professeurs de l'Université de Lille, qui m'ont fourni les outils nécessaires à la réussite de mes études universitaires.

Introduction

Ce sujet consiste à éclaircir la notion de **statistique spatiale**, il existe beaucoup 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.
La statistique spatiale concerne l'étude de ces phénomènes observés dans un domaine spatial.
On note $s \in S$ la localisation d'un site de mesure et $X = \{X_s, s \in S\}$ le phénomène étudié, où X est une variable aléatoire indexée par l'ensemble S, une variable qui prend des valeurs en fonction de sa localisation spatiale est connue comme une variable régionalisée.

Les domaines d'application de **la statistique spatiale** sont nombreux : géologie, écologie, sciences du sol, météorologie, épidémiologie..

Notre but est de caractériser la dépendance spatiale entres différentes observations à l'aide des outils statistiques spatiales.
La géostatistique est une branche des statistiques spatial visant à donner une description de quantités distribuées spatialement ou encore spatiotemporellement.

On s'intéressera au cadre géostatistique où la variable d'étude se déploie continûment sur le domaine S.
Pour cela nous allons étudier les processus aléatoires, 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.

Le travail est tourné vers la pratique, la deuxième partie est une application numérique, on va utiliser la méthode de **Krigeage** sur le **langage R** pour interpoler des données spatial et on va comparer cette méthode avec deux autres méthodes d'interpolations spatiale à savoir :
La Pondération par l'Inverse de la Distance et la méthode de spline.

Chapitre 1

Type de structures spatiales

Les méthodes de statistique spatiale servent à décrire, modéliser des données géo-référencées ou localisées, elle étudie des phénomènes dont l'observation est un processus aléatoire $\{X_s, s \in S\}$ indexé par un ensemble spatial S, X_s appartenant à un espace d'états E avec s localisation d'un site d'observations.

Avant toute analyse ,il faut prendre en mains les données brutes afin de les adapter au processus spatial qui leurs correspondent. Cela permet non seulement dans la démarche de l'analyse, de voir des problèmes particuliers dans les données(données manquantes, problème lors de la collecte, ou valeurs aberrantes...) et de laisser des discussions les hypothèses nécessaires des méthodes économétriques.

On distingue quatre grands types de données géo-référencées

1.1 Données de Type géostatistiques

Les données géostatistiques : les données de type géostatistique sont des données continues, interpolables.
L'outil de modélisation des données géostatistiques est le champ aléatoire. Lorsqu'une caractéristique $X(s; w)$ d'une unité statistique est mesurée en la position s pour la réalisation w, on notera X_s la variable aléatoire associée , où l'indice s varie dans une partie $\mathbb{R}^d$.
Les données sont en général des observations du champ en des points discrets et déterministes de D.

Exemple en météorologique : le champ "vitesse du vent" peut être déni en tout point d'une zone géographique mais est mesuré en un nombre ni de stations météo.

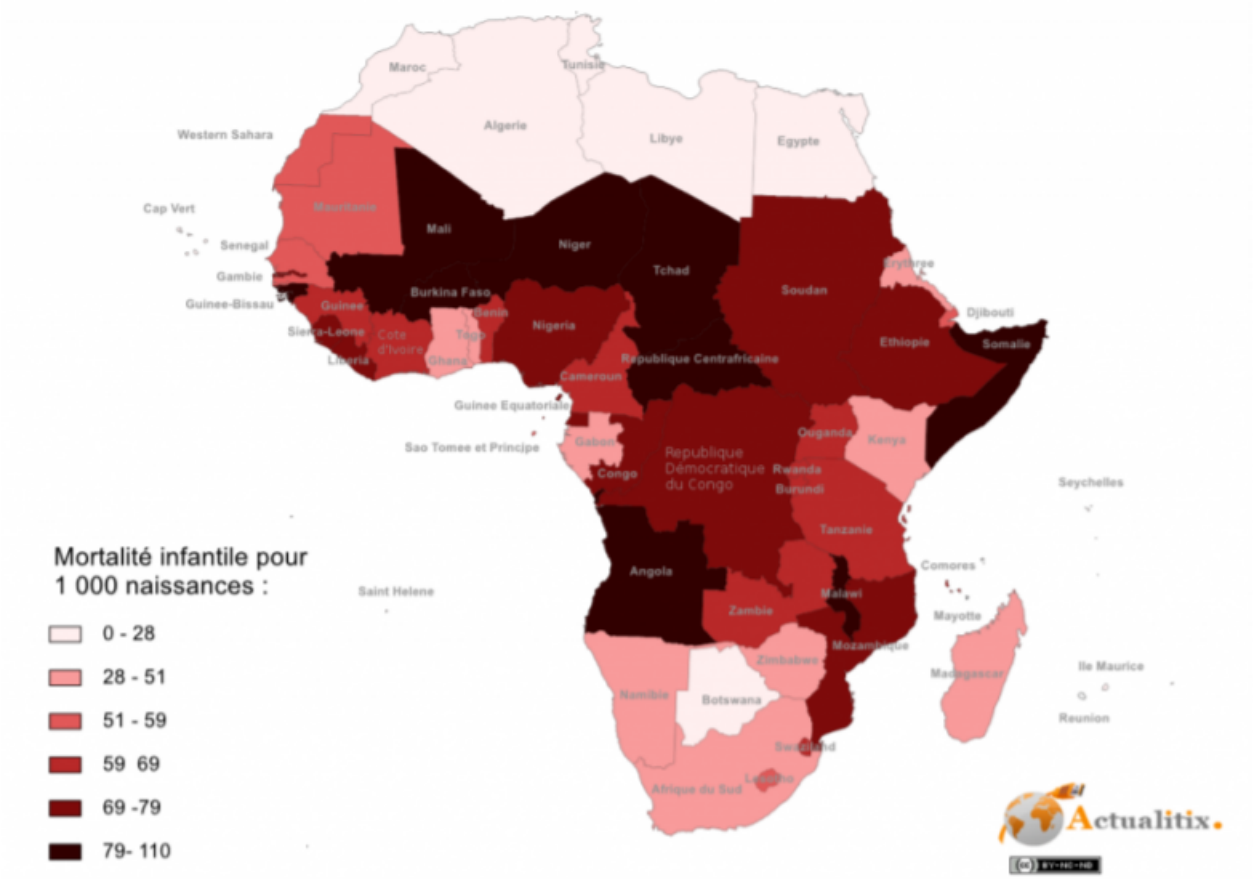
Les objectifs étudiés de la géostatistique :

- Prédiction de X sur tout S en un site non observé (Krigeage).
- Modélisation de la loi du processus continu de $\{X_s, s \in S\}$.

1.2 Données de Type laticielles ou surfacing

Les données laticielles ou données sur réseau S fixé : Les données sont mesurées en un ensemble discret xé non aléatoire, en générale $S \in \mathbf{R}^d$ et l'espace d'état **E** est réel ou non.
Les sites s représentent en général des unités géographiques, repérées par un graphe de voisinage.
X_s n'a de sens que sur une collection dénombrable de zones. par exemple l'exemple suivant chercherr à étudier mortalité infantile en Afrique.

Une carte présentant la mortalité infantile en Afrique. Plus la couleur est rouge foncé plus la mortalité infantile du pays est importante. Inversement plus la couleur est clair moins la mortalité du pays est importante.



1.3 Les données ponctuelles

Les données ponctuelles : Dans ce cas La localisation est elle même l'objective de l'étude.

Le nombre de réalisations ponctuelles et leur localisation sont aléatoire $n = n(s)$

Exemple : Etude de la répartition spatiale d'une espèce d'arbres dans une forêt. Les objectifs étudiés :

— Homogénéité de la localisation des sites est elle plutôt régulière ?

1.4 Spécicité des données spatiales

Spécicité des données spatiales : Les données analysées sont dépendantes.

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.

Chapitre 2

Modélisation Géostatistique

La géostatistique, ou théorie des variables régionalisées, fait l'hypothèse d'une variable, représentative d'un phénomène, qui se déploie dans l'espace, ou du moins dans un domaine D de l'espace, une telle variable s'appelle une variable régionalisée. Il peut s'agir de la concentration en polluaant dans le sol ou dans l'air, de la densité d'arbres d'une forêt, de la densité d'une population, etc. par la suite on va adopter la notation suivante : s un point dans l'espace. En géostatistique, on cherche à exprimer espérance et variance de combinaisons linéaires.

$$\sum_i \lambda_i X(s_i)$$

en des points s_i données qui appartiennent à S qu'on appelle un ensemble spatial, X est une variable régionalisée. De telles quantités seraient connues si l'on disposait d'un modèle d'ordre 2, déni par les 2 premiers moments : — L'espérance en chaque point s :

$$m(s) = E(X_s)$$

— La covariance entre deux points s et t :

$$\begin{aligned} Cov[X_s, X_t] &= E[X_s - m(s)][X_t - m(t)] \\ &= E[X_s X_s] - m(s)m(t) = C(s,t) \end{aligned}$$

Ce chapitre est consacré à l'étude des modèles du second ordre.
On étudiera également la classe plus large des modèles intrinsèques qui sont à accroissements de variances finies.

2.1 Rappels sur les processus stochastiques

Dénition 2.1.1 Un processus (ou champ) stochastique (ou aléatoire) est une famille de variables aléatoires dénies sur le même espace de probabilité (Ω, F, P) indexée par S et à valeurs dans E.
Un processus stochastique est noté par $\{X_s\}_{s\in S}$. La valeur de la variable aléatoire X_s en un certain $\omega \in \Omega$ est désignée par $X_s(\omega)$.

- S est souvent appelé ensemble des indices, souvent, on aura $S = \mathbb{R}$,
Par exemple pour la modélisation de la pluie, nous dénirons donc un processus stochastique déni sur $\mathbb{R}^2$ et à valeur dans $\mathbb{R}_+$.
- Si l'espace d'état E est de la forme $\mathbb{R}^d$, on parle de champ aléatoire.

2.2 Processus stationnaire au second ordre

Pour ce modèle on suppose la stationnarité des 2 moments, c'est à dire leur invariance par translation ce qui permet l'inférence des moments par déplacement dans l'espace.

Dénition 2.2.1 $X_s, s \in S$ est **un processus au second ordre** si $\forall s \in S, E(X_s^2) < \infty$ La moyenne de X est la fonction $m : S \to \mathbb{R}$ dénie par $m(t) = E_t$, et la covariance la fonction $c : S \times S \to R$ dénie par $c(s,t) = Cov(X_s, X_t)$. Le processus est dit centré si sa moyenne est nulle en tout point.

Dénition 2.2.2 $X_s, s \in S$ est dit **strictement stationnaire** si pour $\forall n \geqslant 1$ et $(s_1, s_2, ..., s_n) \in S^n$

$$P\{X(s_1) \in B_1, X(s_2) \in B_2, ...\} = P\{X(s_{1+h}) \in B_1, X(s_{2+h} \in B_2), ...\}$$

C'est généralement une propriété bien trop forte qui font que l'on considère souvent une version assouplie.

Dénition 2.2.3 $X_s, s \in S$ est **faiblement stationnaire au second ordre** si X est un processus au second ordre

1. de moyenne constante

$$\forall s \in S, E(X_s) = m$$

2. de covariance invariante par translation, c'est à dire :

$$\forall t \in S, c(s,t) = Cov(X_s, X_t) = r(t-s)$$

Propriétés de la covariance Soit $X_s, s \in S$ un processus faibelemnt stationnaire. Sa fonction de covariance $Cov(.)$ vérie

-

$$Cov(0) = VarX(s) \geqslant 0 \forall s \in S.$$

-

$$Cov(h) = Cov(h);$$

- Pour tout $n \geqslant 1$ et $\lambda_1, ..., \lambda_n \in \mathbb{R}$ et $s_1, ..., s_n \in S$

$$\sum_{i,j=1}^{n} \lambda_i \lambda_j Cov(s_i - s_j) \geqslant 0$$

Cette dernière propriété revient à dire que $Cov(.)$ est (semi) dénie positive. Dans un esprit de modélisation, il sera souvent agréable de supposer l'isotropie. L'isotropie, signie que la structure spatiale de la variable est uniforme dans toutes les directions. Ceci n'est pas souvent le cas, et la variable montre une anisotropie, ce qui signie qu'il y a une tendance de dépendance à la direction dans les données. Si un paramètre présente différentes variations avec la direction, ceci implique qu' il y a une anisotropie géométrique. Par exemple, dans un dépôt de dunes, le plus large éventail coïncide avec la direction du vent comparé à celui qui lui est perpendiculaire.

Chapitre 3

Variabilité spaciale

3.1 Le variogramme

Le variogramme est un outil descriptif puissant utilisable dans une multitude de domaines, il explore la structure spatiale des données. Le variogramme décrit l'évolution de la semi-variance en fonction de la distance entre les mesures et permet ainsi d'étudier le lien spatial entre les données. Il est dénit de la manière suivante :

$$\frac{1}{2}E(X_{s+h} - X_s)^2$$

En pratique, l'analyse variographique va se dérouler en deux étapes :

1. Estimation du variogramme.

2. Modélisation du variogramme.

3.1.1 Processus intrinsèques

Lorsque la notion de stationnarité devient trop forte, une de ces est souvent quand la moyenne change sur le territoire d'intérêt et que la variance ne peut pas toujours être borné lorsque cette région d'intérêt s'agrandit . Georges Matheron établit alors après avoir tiré les conséquences de ces limites de la stationnarité au second ordre en proposant une encore plus faible : la stationnarité intrinsèque .

En pratique la condition de la moyenne constante du modèle précédent est souvent trop contraignate. Un modèle plus général, très employé, c'est le modèle intrinsèque, déni par des hypothèses de stationnarité sur ses accroissements $X(s+h) - X(s)$

Définition Un processus est dit **intrinsèque** si ses accroissements sont :

- D'espérance nulle :

$$E(X_s + h - X_s) = 0$$

et de variance ne dépendant que de la distance h entre les points :

$$\frac{1}{2}E(X_{s+h} - X_s)^2 = \gamma(h)$$

Le modèle est donc entièrement spécié par la fonction $\gamma(h)$, appelée variogramme, qui mesure direcetement la variabilité spatiale du phénomène.

3.2 Analyse du corrélogramme

Après avoir détaillé le choix du modèle probabiliste, nous nous sommes focalisés sur un outil indispensable pour la modélisation spatiale des données : le variogramme.
Le variogramme reflète la structure de la régionalisation.
On appelle analyse variographique l'inférence du variogramme à partir de données expérimentales.
Elle permet de restituer des informations quant à la distribution spatiale de la variable régionalisée, c'est-à-dire de la réalisation de la fonction aléatoire d'intérêt.
On s'attache usuellement à l'étude du variogramme plutôt que de la covariance car, comme nous l'avons vu précédemment, celle-ci n'étant pas nécessairement définie dans le cadre intrinsèque.
L'analyse variographique constitue une étape cruciale dans une étude géostatistique : elle permet postérieurement d'estimer les valeurs inconnues de la variable régionalisée et d'assortir leur estimation d'une précision.

3.2.1 Estimation du variogramme

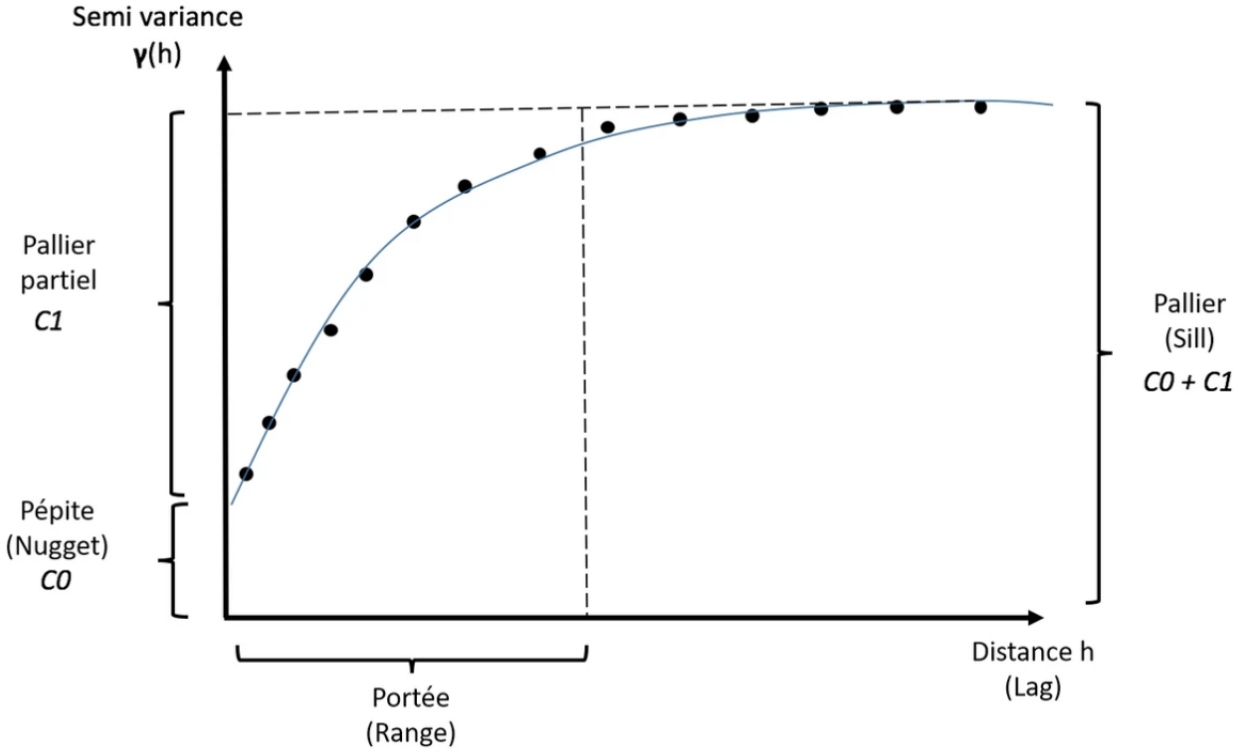
Objectif : Fournir des informations sur la nature et la structure de dépendance spatiale dans un espace aléatoire, le variogramme doit être estimée à partir des données. Pour une valeur de $\gamma(h)$ donnée, on obtient une estimation empirique de de la manière suivante

$$\frac{1}{2}\frac{1}{N(h)}\sum_{s_i - s_j \approx h}(X(s_i) - X(s_j))^2$$

où $N(h)$ est le nombre de paires des points (s_i, s_j) séparés par la distance h. En pratique, le variogramme expérimental est calculé pour des classes de distances, et par classes de direction si nécessaire.
Dénition 3.1.1 le variogramme est isotrope, c'est-à-dire que la variabilité est la même dans les toutes les directions du plan. Ce n'est généralement pas le cas. Lorsque cela se produit, nous devons dénir un variogramme pour

chacun des axes du plan.

3.3 Les composantes du variogramme



— **Effet de pépite (Nugget) C_0 :** L'eet de pépite correspond à la limite du variogramme en zéro. Elle représente donc la variation entre deux mesures très proches et peut donc provenir de :
- 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.

— **Portée a :** Distance où deux observations ne se ressemblent plus du tout en moyenne, elles ne sont plus liées (covariance nulle) linéairement (non corrélés). A cette distance, la valeur du variogramme correspond à la variance de la variable aléatoire.

— **Palier $C - C_0 = \alpha^2$** C'est la variance de la variable aléatoire

- Lorsque $h = 0+$, alors $\gamma(h) = C_0$. L'effet de pépite se présente donc comme une discontinuité à l'origine du variogramme.
- Lorsque les variogrammes montrent un palier alors on peut facilement établir le lien entre la valeur du variogramme pour la distance h et la covariance pour deux observations séparées de h.

$$\gamma(h)\frac{1}{2}Var[X_s - X_S + h]$$

$$\frac{1}{2}[Var[X_s] + Var[X_S + h] - 2Cov(X_s, X_S + h)]$$

$$\sigma^2 - C(h)$$

3.4 Modélisation du variogramme

Le variogramme estimé n'est pas prédictif et ne respecte le plus souvent pas les contraintes de krigeage. C'est pourquoi les méthodes géostatistiques modélisent le variogramme estimé par une fonction continue soumise à certaines contraintes (fonction conditionnellement dénie négative). Cette étape s'appelle la modélisation ou l'ajustement du variogramme. La modélisation est la partie essentielle du krigeage.

Condition d'admissibilité des modèles : Toute fonction ne peut être utilisée comme modèle. Soit une somme quelconque de variables aléatoires (plus généralement, une combinaison linéaire de telles v.a.), la variance de cette combinaison est nécessairement positive (une variance est, par dénition, toujours positive). Or cette variance peut s'exprimer en fonction du covariogramme (modèles avec palier) ou du variogramme (modèles avec palier ou sans palier pourvu que la somme des poids de la combinaison linéaire donne 0). Il faut donc que le covariogramme ou le variogramme assure des variances positives quelle que soit la combinaison des v.a. considérée.

Modèles de variogramme

- Le modèle pépitique :

$$\gamma(h) = 1, \forall h \neq 0$$

- Le modèle sphérique de portée a :

$$\gamma(h) = \frac{1}{3}\frac{|h|}{a} - \frac{1}{2}(\frac{|h|}{a})^3$$

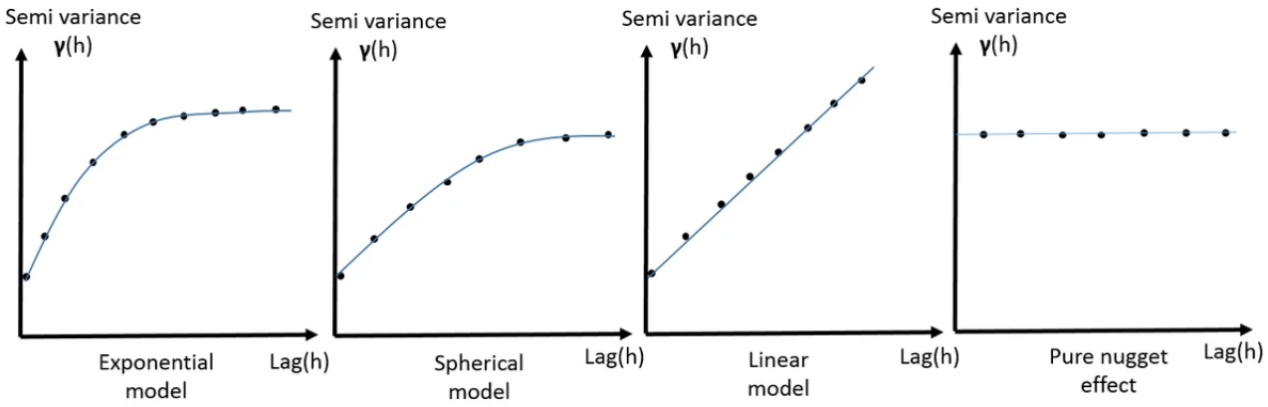
pour $|h| < a$ et pour $|h| \geq a$ - Le modèle exponentiel, de paramètre d'échelle a et d eportée pratique $3a$

$$\gamma(h) = 1 - exp(-\frac{|h|}{a})$$

- Le modèle gaussien, de paramètre d'échelle a et de portée pratique $1.73a$

$$\gamma(h) = 1 - \exp(-(\frac{|h|}{a})^2)$$

- Le modèle gigones : Chacun des modèles précédents peut être utilisé isolément, mais on peut aussi les additionner, en faisant varier le palier, portée et éventuellement anisotropie.



Chapitre 4

Krigeage

4.1 Principe et origine du krigeage

L'interpolation spatiale est un problème classique d'estimation d'une fonction $F(s)$, ou $s = (x, y)$, en un point x_s du plan à partir de valeurs connues de F en un certain nombre, n, de points environnants x_i.

$$F(s) = \sum_{i=1}^{n} \lambda_i X(s_i)$$

Le problème consiste à déterminer la pondération,i.e. les λ_i, de chacun des points environnants. Il existe plusieurs fa¸cons de choisir ces poids. Dans ce chapitre, on s'intéresse à la méthode de krigeage.

La méthode du krigeage Le Krigeage est une technique d'interpolation spatiale basée sur des semi-variogrammes où on cherche à estimer la valeur d'une variable régionalisée X en un point s_0 quelconque du champ à partir des mesures observées $X(s_i), i = 1, .., n$ (n : nombre de points observés).

Le krigeage est un interpolateur exact (la valeur estimée sur un point de mesure est égale à la valeur du point de mesure) et optimal (il est sans biais et minimise la variance sur l'erreur d'estimation).

Contraintes de construction de Krigeage Soit $\hat{X}(s_0)$ la prédiction en un site s_0. On veut une estimation non biaisée et de variance minimale. Ces conditions se résument en quatre contraintes qui vont permettre de construire le système de krigeage :
- contrainte de linéarité : L'estimateur est une combinaison linéaire pondérée des données.

$$\hat{X}(s_0) = \sum_{i=1}^{n} \lambda_i \hat{X}(s_i)$$

- Contrainte d'autorisation : L'erreur d'estimation doit être une combinaison linéaire autorisée, c'est-à-dire que son espérance $E[\hat{X}(s_0) - X(s_0)]$ et sa variance $Var[\hat{X}(s_0) - X(s_0)]$ doivent exister.
- Contrainte de non-biais : L'espérance de l'erreur d'estimation doit être nulle.

$$E[\hat{X}(s_0) - X(s_0)] = 0$$

- Contrainte d'otpimalité : Les poids λ_i sont déterminés de façon à minimiser la variance des erreurs $Var[\hat{X}(s_0) - X(s_0)]$.

4.2 Types de Krigeage

Le modèle du Krigeage Le modèle du krigeage s'énonce comme suit :

$$X(s) = \mu(s) + \delta(s), s \in S$$

- $\delta(s)$ est une fonction aléatoire stationnaire, d'espérance nulle et de structure de dépendance connue.

- $\mu(s)$ est la structure déterministe pour l'espérance de $X(.)$ Selon μ on distingue trois types de Krigeage

- krigeage simple : Variable à interpoler stationnaire $\mu = m$ est une constante connue.

- krigeage ordinaire : Variable à interpoler stationnaire $\mu(s) = \mu$ est une constante inconnue.
- Le krigeage universel : Variable non-stationnaire.

4.2.1 krigeage simple

Le modèle de Krigeage simple s'écrit de la façon suivante :

$$X(s) = \mu(s) + \delta(s), s \in S$$

avec m constante connue et $\delta(.)$ fonction aléatoire stationnaire de second ordre d'espérance nulle et de structure de dépendance connue. La stationnarité de second ordre implique que le semi-

variogramme de $\delta(.)$ atteigne un palier. La prévision avec le krigeage simple s'écrit sous la forme :

$$\hat{X}(s) = \mu(s) + \lambda^T(X - m1_n)$$

preuve 4.1.1 La contrainte de linéarité implique que la prévision b X doit être de la forme suivante

$$\hat{X}(s) = a + \sum \lambda_i X(s_i) = a + \lambda^T X$$

avec X : vecteur des variables aléatoires intervenant dans la prévision et λ vecteur des poids. La Contrainte de non-biais implique que

$$E[\hat{X}(s_0) - X(s_0)] = 0$$

$$E[a + \lambda^T X - X(s_0)] = 0$$

$$a + \lambda^T m1_n - m = 0$$

$$a - m(1 - \lambda^T 1_n) = 0$$

$$a = m(1 - \lambda^T 1_n)$$

La prévision peut donc être réecrite sous la forme :

$$\hat{X} = a + \Sigma\lambda_i X(s_i) = a + \lambda^T X$$

$$\hat{X} = m(1 - \lambda^T 1_n) + \lambda^T X$$

$$\hat{X} = m + \lambda^T + (X - m1_n)$$

Pour le krigeage simple, les λ_i s'écrivent sous la forme

$$\hat{\lambda} = \Sigma^{-1} c_0$$

et la variance du krigeage vaut :

$$\sigma^2(s_0) = \sigma^2 - c_0 \Sigma^{-1} c_0$$

avec :

- Σ :matrice de variance-covariance de δ.

- δ : vecteur des erreurs associés aux variables aléatoires $X(s_i)$ à $X(s_n)$

preuve 4.1.2

$$Var[\hat{X} - X(s_0)]$$

$$= Var[m + \lambda^T (X - m1_n) - X(s_0)]$$

Rappelons que $\hat{X} = m + \lambda^T (X - m1_n)$

$$= Var[m + \lambda^T \delta - m - \delta(s_0)]$$

$$= Var[\lambda^T \delta - \delta(s_0)]$$

$$= Var[\lambda^T \delta] + Var[\delta(s_0)] - 2Cov(\lambda^T \delta, \delta(s_0))$$

$$= \lambda^T Var[\delta] + Var[\delta(s_0)] - 2\lambda^T Cov(\delta, \delta(s_0))$$

$$= \lambda^T \Sigma \lambda + \sigma^2 \lambda^T c_0$$

$$= f(\lambda)$$

Le but est de minimiser $= f(\lambda)$, pour cela on calcule le gradient

de cette fonction et on chechre ces racines.

$$\frac{\partial f(\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda}(\lambda^T \Sigma \lambda + \sigma^2 \lambda^T c_0)$$

$$= 2\Sigma\lambda - 2c_0$$

$$\frac{\partial f(\lambda)}{\partial \lambda} = 0 \Rightarrow 2\Sigma\lambda - 2c_0 = 0$$

$$\Leftrightarrow 2\Sigma\lambda = 2c_0 \Leftrightarrow \Sigma\lambda = c_0$$

$$\lambda = \Sigma^{-1} c_0$$

La matrice hessien est égale à $2\sum$, cette matrice est semi dénie positive ce qui implique que $f(\lambda)$ est convexe et le point crtique $\lambda = \sum^{-1} c_0$ est un minimum global. Ainsi, la prévision (equation (4.2)) par le krigeage simple devient

$$\hat{X} = m + c_0^T \Sigma^{-1}(X - m1_n)$$

La valeur minimale de la variance de l'erreur de prevision s'appelle Variance de Krigeage elle vaut

$$\sigma^2(s_0) = Var[\hat{X} - X(s_0)]$$

$$= Var[\hat{X}] + Var[X(s_0)] - 2Cov(\hat{X}, X(s_0))$$

on a

$$Var[X(s_0)] = Var[m + \delta(s_0)] = \sigma^2$$

et

$$= Var[\hat{X}] = Var[m + c_0^T \Sigma^{-1}(X - m1_n)]$$

$$= Var[c_0^T \Sigma^{-1}(X - m1_n)] = Var[c_0^T \Sigma^{-1}\delta]$$

$$= c_0^T \Sigma^{-1} Var[\delta] \Sigma^{-1} c_0$$

$$= c_0^T \Sigma^{-1} \Sigma \Sigma^{-1} c_0$$

$$= c_0^T \Sigma^{-1} c_0$$

et

$$Cov(\hat{X}, X) = -2c_0^T \Sigma^{-1} c_0$$

On remplace dans (4.3) on trouve que la variance de Krigeage vaut

$$\sigma^2(s_0) = c_0^T \Sigma^{-1} c_0 + \sigma^2 - 2c_0^T \Sigma^{-1} c_0$$

$$\sigma^2(s_0) = \sigma^2 - c_0^T \Sigma^{-1} c_0$$

4.2.2 krigeage ordinaire

La méthode du Krigeage simple suppose que l'espérance de X est connue, cette hypothèse est rarement vérié, cette méthode a été généralisé. Dans le cas où l'espérance est inconnue, il s'agit du Krigeage Ordinaire. Le modèle de cette méthode est

$$X(s) = \mu(s) + \delta(s), s \in S$$

avec μ une quasi-constante (l'espérance n'est pas la même partout dans le champ S mais peut rester constante à l'intérieur de chaque voisinage de krigeage) et $\delta(.)$ est une fonction aléatoire intrinsèque d'espérance nulle. Le prédicteur du Krigeage ordinaire est donnée par

$$\hat{X} = (c_0 \frac{1 - 1_n \Sigma^{-1} c_0}{1_n \Sigma^{-1} 1_n} 1_n) \Sigma^{-1} X$$

preuve 4.1.3

1. La prévision $\hat{X}$ est une combinaison linéaires des variables $X(s_i)$

$$\hat{X} = a + \Sigma \lambda_i X(s_i) = a + \lambda^T X$$

2. $\delta(.)$est une fonction aléatoire intrinsèque, cela veut dire que les deux premiers moments de l'erreur de prévision existent. L'erreur de prévision est

$$\hat{X} - X(s_0) = a + \sum \lambda_i X(s_i) - X(s_0)$$

$$= a + \sum \lambda_i(\mu + \delta(s_i)) - (\mu + \delta(s_0))$$
$$= a + \sum \lambda_i(\mu + \delta(s_i)) - \mu - \delta(s_0))$$

$$= a + \mu \sum \lambda_i - \mu + \sum \lambda_i \delta(s_i) - \delta(s_0)$$

$$= a + \mu(\sum \lambda_i - 1) + \sum \lambda_i$$

Pour que l'erreur soit une combinaison linéaire d'accroissement de $\delta(.)$ il faut que $\sum \lambda_i - 1 = 0 \Leftrightarrow \sum \lambda_i = 1.$

Par la suite on travaille avec la contrainte $\sum \lambda_i = 1$ pour que l'espérance et la variance existent.

On sait que la prévision doit être non biaisé, il faut donc que

$$E[\hat{X} - X(s_0)] = E[a + \sum \lambda_i X(s_i) - X(s_0)] = a + \mu(\sum \lambda_i - 1) = 0$$

comme $\sum \lambda_i - 1$ et pour que $\hat{X}$ soit sans biais, il faut que a doit être égale à zéro.

La prévision se simplifie à $\hat{X} = \sum \lambda_i X(s_i)$. il reste à trouver les λ_i qui minimisent la variance de l'erreur de prévision. On pose :

- Γ : matrice dont l'élèment (i, j) et $\gamma(s_i, s_j)$, soit le semi-variogramme entre $\delta(s_i)$ et $\delta(s_j)$ les élèments i et j de δ.

- γ_0 : vecteur dont l'élèment i $\gamma(s_i, s_j)$, semi-variogramme entre $\delta(s_i)$ et $\delta(s_0)$ Dans un premier lieu, on va exprimer la variance de

l'erreur de prévision en fonction de Γ et γ_0.

$$Var[X(\hat{s}_0) - X(s_0)] = E[(X(\hat{s}_0) - X(s_0))^2]$$

$$= E[(\sum \lambda_i X(\hat{s}_i) - X(s_0))^2]$$

$$= E[(\sum \lambda_i \mu + \sum \lambda_i \delta(s_i) - \mu\delta(s_0))^2]$$
$$= E[(\mu(\sum \lambda_i - 1) + \sum \lambda_i \delta(s_i) - \delta(s_0))^2]$$

$$= E[(\sum \lambda_i \delta(s_i) - \delta(s_0))^2]$$

$$= E[(\sum \lambda_i \delta(s_i))^2 - 2\delta(s_0) \sum \lambda_i \delta(s_i) + \delta(s_0)^2]$$

$$\sum_i \sum_j \lambda_i \lambda_j \delta(s_i)\delta(s_j) - 2\delta(s_0) \sum \lambda_i \delta(s_i) + \delta(s_0)^2$$

$$\underbrace{\sum_i \sum_j \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \sum \lambda_i \delta(s_i)^2}_{\text{A}} + \underbrace{\sum \lambda_i \delta(s_i)^2 - 2\delta(s_0) \sum \lambda_i \delta(s_i) + \delta(s_0)^2}_{\text{B}}$$

Pour tous voisinage V du point s_0, (les points utilisé dans le krigeage) Le terme A peut s'écrire sous la forme

$$\sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \sum_{i \in V(s_0)} \lambda_i \delta(s_i)^2$$

$$= \sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \sum_{j \in V(s_0)} \lambda_j \sum_{i \in V(s_0)} \lambda_i \delta(s_i)^2$$

$$= \sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j \delta(s_i)\delta(s_j) - \frac{1}{2} \sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j \delta(s_i)^2 - \frac{1}{2} \sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j \delta(s_j)^2$$

$$= -\frac{1}{2} \sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j (-\delta(s_i)\delta(s_j) + \delta(s_i)^2 + \delta(s_j)^2)$$

$$= -\frac{1}{2} \sum_{i \in V(s_0)} \sum_{j \in V(s_0)} \lambda_i \lambda_j (\delta(s_i)\delta(s_j))^2$$

et le terme B :

$$\sum_{i \in V(s_0)} \lambda_i \delta(s_i)^2 - 2\delta(s_0) \sum_{i \in V(s_0)} \lambda_i \delta(s_i) + \delta(s_0)^2$$

$$= \sum_{i \in V(s_0)} \lambda_i \delta(s_i)^2 - 2\delta(s_0) \sum_{i \in V(s_0)} \lambda_i \delta(s_i) + \sum_{i \in V(s_0)} \lambda_i \delta(s_0)^2$$

$$= \sum_{i \in V(s_0)} \lambda_i (\delta(s_i)^2 - 2\delta(s_0)\delta(s_i) + \delta(s_0)^2)$$

$$= \sum_{i \in V(s_0)} \lambda_i (\delta(s_0) - \delta(s_i))^2$$

On remplace dans l'expréssion de la variance

$$-\frac{1}{2}\sum_i\sum_j E[(\delta(s_i)-\delta(s_j))^2]+\sum\lambda_i E[(\delta(s_0)-\delta(s_j))^2]$$

$$-\sum_i\sum_j\lambda_i\lambda_j(\delta(s_i)-\delta(s_j))+\sum\lambda_i\gamma(s_0-s_i)$$

$$-\lambda^T\Gamma\lambda+2\lambda^T\gamma_0=f(\lambda)$$

Le but est de minimiser $f(\lambda)$ sous contrainte $\sum\lambda_i-1=0$, à l'aide du Langrangien noté L on minimise

$$f(\lambda,L)=-\lambda^T\Gamma\lambda+2\lambda^T\gamma_0$$
$$\frac{\partial}{\partial\lambda}f(\lambda,L)=-2\Gamma\lambda+2\gamma_0+2L1_n$$

Le point critique est $\lambda=\Gamma^{-1}(\gamma_0+L1_n)$ on multiplie l'équation par 1_n^T on trouve que $L=\frac{1-1_n\Gamma^{-1}\gamma_0}{1_n\Gamma^{-1}1_n}$ l'expression de λ devient

$$\lambda=\Gamma^{-1}(\gamma_0+\frac{1-1_n\Gamma^{-1}\gamma_0}{1_n\Gamma^{-1}1_n})$$

Ainsi on trouve que la prévision avec le Krigeage ordinaire est

$$\hat{X}=(c_0+\frac{1-1_n\Sigma^{-1}c_0}{1_n\Sigma^{-1}1_n}1_n)\Sigma^{-1}X$$

Chapitre 5

Tests d'autocorrelation spaciale

L'auto-corrélation mesure la corrélation d'une variable avec elle-même, lorsque les observations sont considérées avec un décalage dans le temps (autocorrélation temporelle) ou dans l'espace (autocorrélation spatiale). Si la présence d'une qualité dans une partie d'un territoire rend sa présence dans les zones voisines plus ou moins probable, il existe un effet de contiguïté dans la structure spatiale, le phénomène montre une autocorrélation spatiale.

- L'autocorrélation spatiale est positive lorsque des valeurs similaires de la variable à étudier se regroupent géographiquement.
- L'autocorrélation spatiale est négative lorsque des valeurs dissemblables de la variable à étudier se regroupent géographiquement : des lieux proches sont plus différents que des lieux éloignés.
On retrouve généralement ce type de situation en présence de concurrence spatiale.
- En l'absence d'autocorrélation spatiale, on peut considérer que la répartition spatiale des observations est aléatoire.

5.1 Indice de Moran

L'indice de Moran (Moran 1950) permet de mesurer le niveau d'autocorrélation spatiale d'une variable et de tester sa significativité. Il est égal au ratio de la covariance entre observations contiguës (définies par la matrice d'interactions spatiales) à la variance totale de l'échantillon (Jayet, 2001).

L'indice a des valeurs comprises entre -1 (indiquant une dispersion parfaite) à 1 (corrélation parfaite).

Une valeur nulle signifie que la distribution spatiale de la variable étudiée est parfaitement aléatoire dans le territoire. Les valeurs négatives (positives) de l'indice indiquent une autocorrélation spatiale négative (positive).

Une interprétation facile de l'indice de Moran peut être faite avec l'aide du diagramme de Moran, qui représente, sous la forme d'un nuage de points, les couples de valeurs correspondant à la valeur de la variable dans chaque unité spatiale (en abscisse) et la moyenne des valeurs des zones contiguës (en ordonnée) définies par la matrice d'interactions spatiale (cette moyenne est appellée spatial lag ou décalage spatial).

La formule :

$$I_{MORAN} = \frac{n}{\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- i, j = unité spatiale ;
- n = nombre d'unités spatiales ;
- X_i est la valeur de la variable dans l'unité i ;
- $\bar{X}$ est la moyenne de x ;
- w_{ij} sont les éléments de la matrice d'interactions spatiales ;

5.2 Indice de Geary

Une autre alternative pour mesurer l'autocorrélation spatiale est l'indice de Geary qui est, à un facteur près, égal au ratio de la variance des écarts entre observations contigües à la variance totale (Jayet, 2001). Si l'indice de Moran est une mesure de l'autocorrélation spatiale globale, l'indice de Geary est plus sensible à l'autocorrélation spatiale locale.
L'indice de Geary varie de 0 à l'infini et vaut 1 s'il y a indépendance spatiale. Pour toute valeur inférieur à l'unité, il y a une autocorrélation spatiale positive, et inversement pour des valeurs supérieures à l'unité. L'indice de Geary varie en sens inverse de l'indice de Moran. Comme pour l'indice de Moran, l'indice de Geary peut être testé statistiquement, par une transformation en Z- scores et en déterminant son seuil de significativité.

La formulation de l'indice de Geary est

$$I_{GEARY} = \frac{(n-1)(X_i - \bar{X}) \sum_{j=1}^{n} w_{ij}(X_i - X_j)^2}{2(\sum_{i=1}^{n} \sum_{i=1}^{n} w_{ij}) \sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- i, j = unité spatiale ;
- n = nombre d'unités spatiales ;
- X_i est la valeur de la variable dans l'unité i ;
- $\bar{X}$ est la moyenne de x ;
- w_{ij} sont les éléments de la matrice d'interactions spatiales ;

Chapitre 6

Application numérique

6.1 Présentation des données

L'interpolation spatiale est le processus d'utilisation des points avec des valeurs connues pour des valeurs estimées à d'autres points inconnus.
Par exemple, pour faire une carte des précipitations (pluie) de votre pays, vous ne trouverez pas assez de stations météo réparties uniformément pour couvrir l'entier de la région.
L'interpolation spatiale peut estimer la température à des endroits sans données enregistrées en utilisant des relevés de températures connus dans des stations météo à proximité. Dans cette partie, nous allons appliquer la méthode du krigeage et nous l'avons comparons avec deux autres méthodes pour cela nous utilisons les données de la pollution de l'air en Californie.

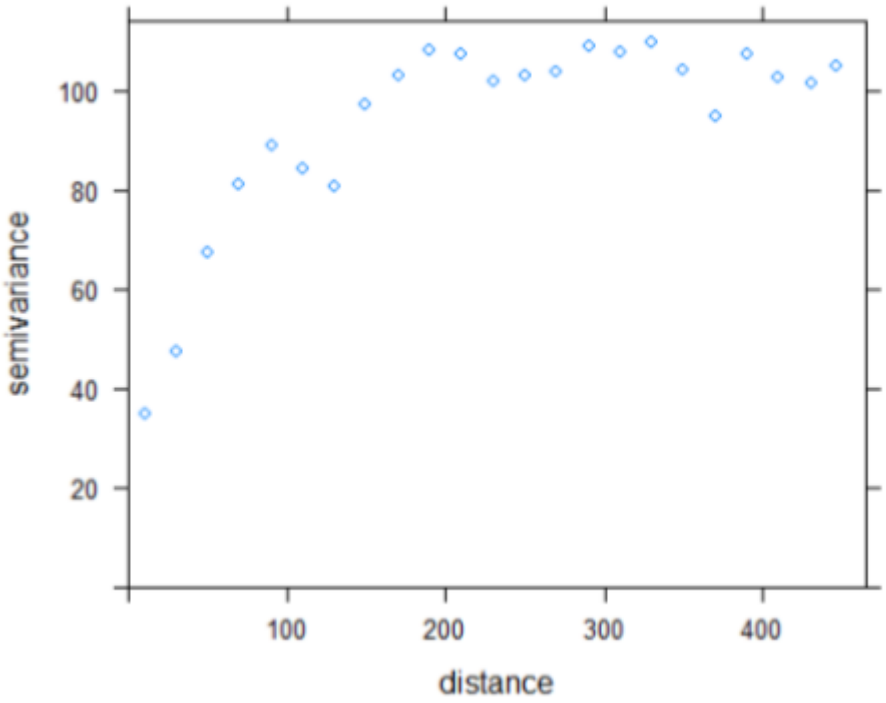
Cette base de données contient 36 variables

- Source : California air resources board.
- la variable SITENAME contient les noms des sites étudiés.
- Location numéro du site s.
- LATITUDE et LONGITUDE contienent les coordonnées des sites

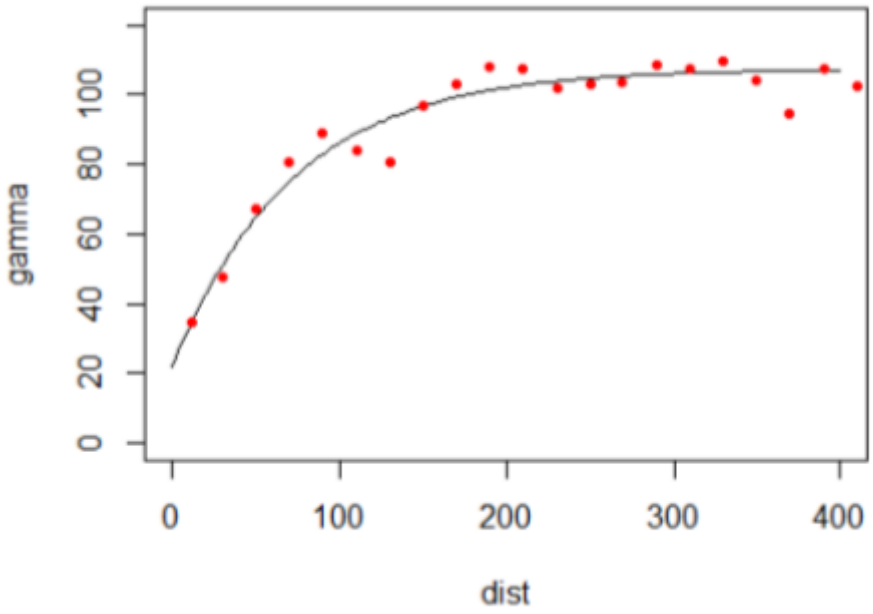
— Le reste des variables sont des composantes de l'air, on s'intéresse à la variable OZDLY AV qui présente la concentration moyenne d'ozone par jour.

6.2 Ajuster un variogramme

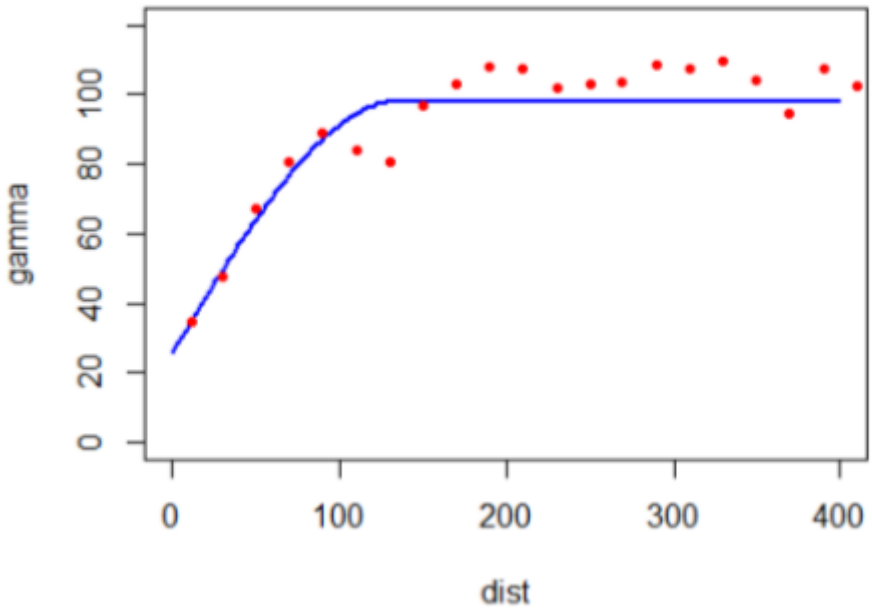
On commence par importer les données et les transformer en un spatialdatapoint.



ensuite on va choisir un modèle de variogramme. On utilise un modèle exponentiel



Puis un modèle sphérique

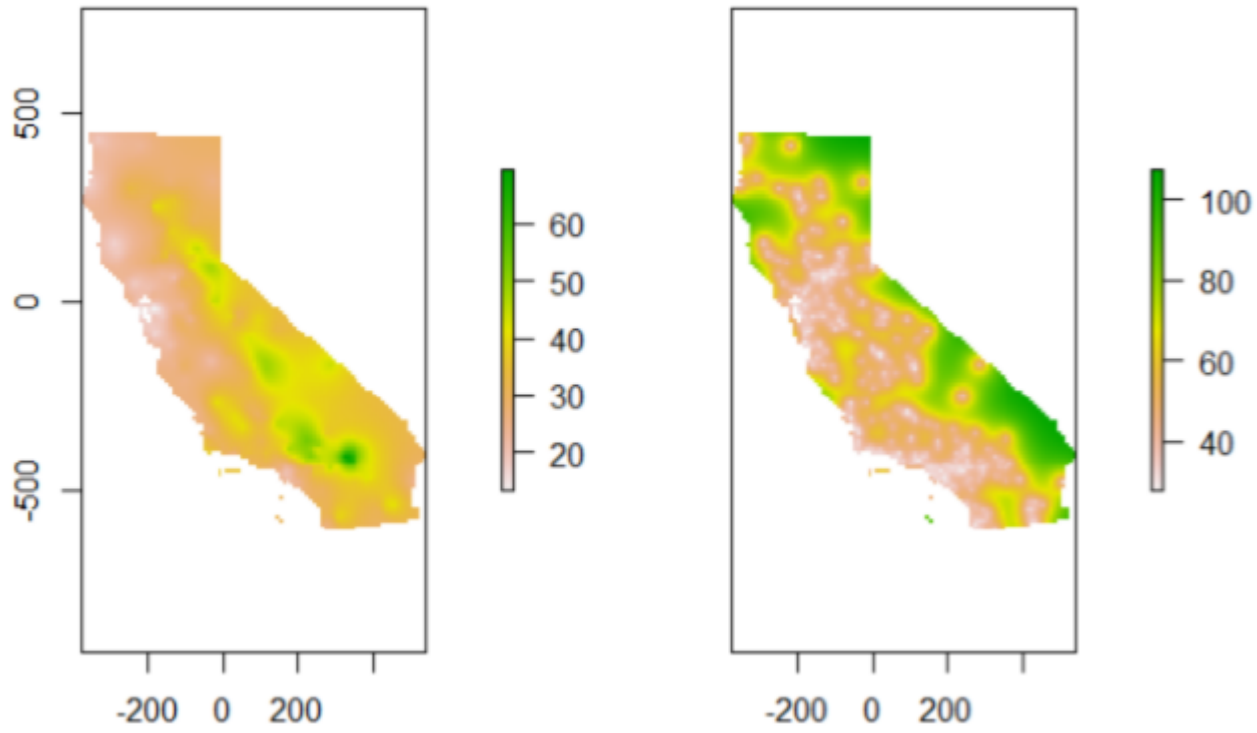


Les deux semblent plutôt bien dans ce cas.

6.3 Méthodes d'interpolations

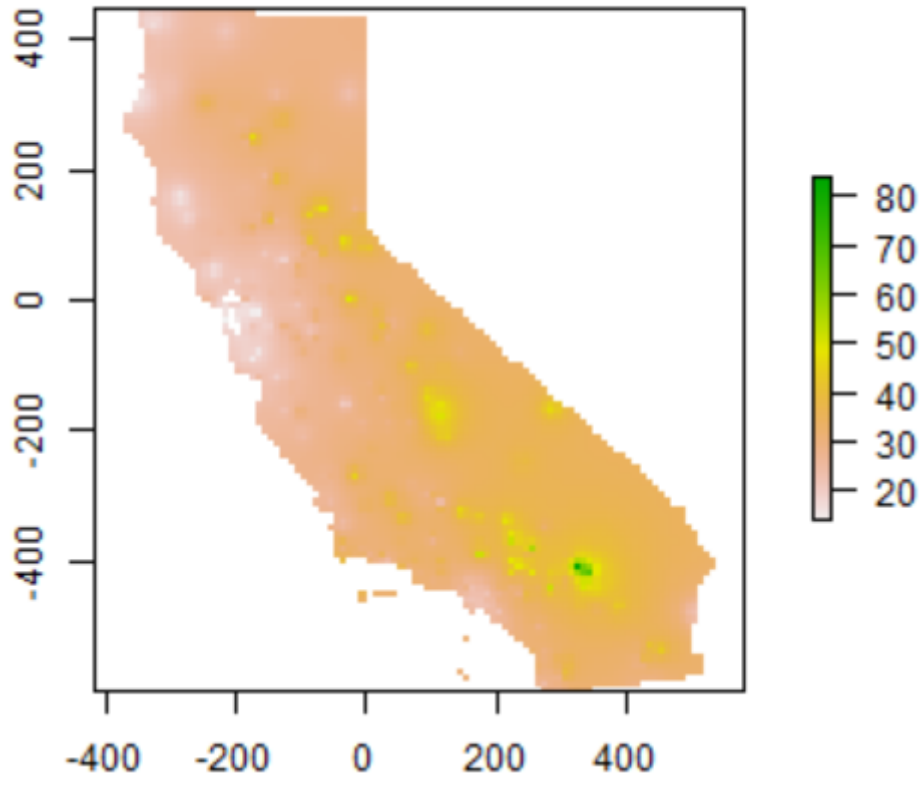
6.3.1 Krigeage

On compare le résultats du krigeage et la projection des vrais valeurs.

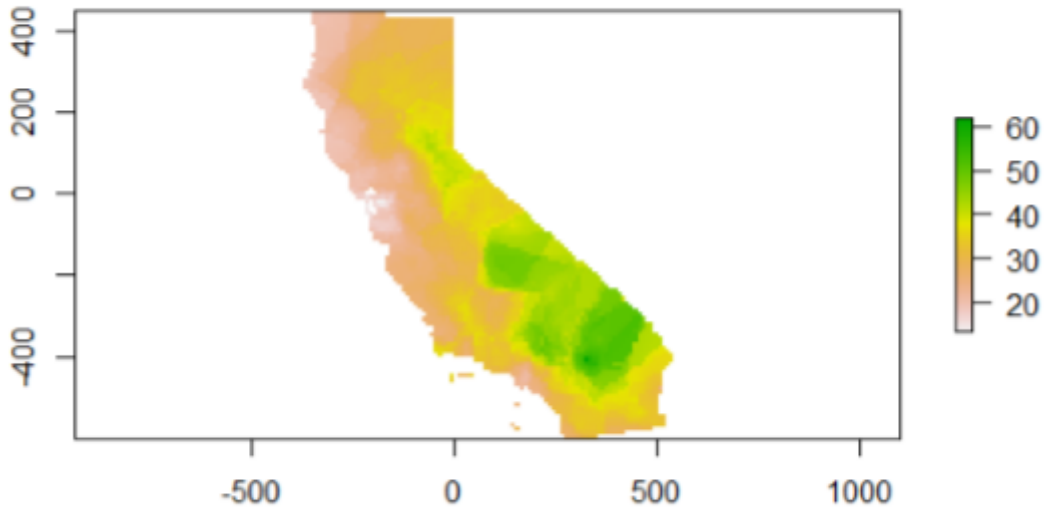


6.3.2 Pondération par l'Inverse de la Distance IDW

Dans la méthode d'interpolation IDW, les points d'échantillons sont pondérés durant l'interpolation de telle sorte que l'inuence d'un point par rapport à un autre décline avec la distance du point inconnu L'approche de base en premier



Nous pouvons trouver de bonnes valeurs pour les paramètres d'idw (décroissance de la distance et nombre de voisins) grâce à l'optimisation. On utilisera La fonction **optim**. Vous fournissez une fonction qui renvoie une valeur que vous souhaitez minimiser (ou maximiser) en fonction d'un certain nombre de paramètres inconnus. Vous fournissez des valeurs initiales pour ces paramètres et **optim** puis recherchez les valeurs optimales (pour lesquelles la fonction renvoie le nombre le plus bas).



6.4 Validation d'un modèle

En utilisant le code en annexes pour le calcul d'erreurs de chaque méthode, on trouve les résultats suivants

```
> rmi <- mean(idwrmse)
> rmk <- mean(krigrmse)
> rmt <- mean(tpsrmse)
> rms <- c(rmi, rmt, rmk)
> rms
[1] 7.925989 8.816963 7.588549
```

On choisit alors la méthode de krigeage.

Bibliographie

1. *Jean-jacques Droesbeke, Michel Lejeune, Gilbert Saporta* **Analyse statistique des données spatiales.**

2. Modélisation et statistique spatiales (Mathématiques et Applications) - *Springer (2008), Carlo Gaetan, Xavier Guyon.*

3. *Edith Gabriel* **Introduction à la statistique spatiale**.

4. *Gilles Guillot* **Introduction à la géostatistique.**