Université d'Angers

Master Mathématiques et Applications

M2 DATA SCIENCE

Année 2021-2022

Projet annuel

Krigeage

Etudiantes :
Alexandre Belnou
Linh Nguyen Phuong

Enseignants :
Frédéric Proïa

Table des matières

Table des figures

Introduction

L'utilisation des données spatiales est en pleine expansion dans de nombreux domaines, tels qu'en sciences de l'environnement et de la terre, météorologie, économétrie, épidémiologie, démographie. L'analyse mathématique associée est donnée par la statistique spatiale, qui permet de décrire, modéliser ces donnés géolocalisées. La particularité de ce type de données est qu'elles sont rarement indépendantes entre elles : l'hypothèse d'indépendance et de distribution identique n'est pas vérifiée. Plusieurs méthodes adaptées ont donc été développées, notamment le krigeage, développé en 1962 par Matheron. La problématique principale du traitement de données spatiales est l'impossibilité à mesurer la valeur d'un phénomène en chaque point d'un champ en deux dimensions. L'idée est donc de trouver une méthode permettant de prédire la valeur en chaque point en fonction d'un nombre fini de mesures. Le krigeage permet de réaliser cette interpolation grâce à son étude de la structure de dépendance spatiale du champ étudié.

Dans ce travail de recherche seront présentés le krigeage sous ses formes simple et ordinaire, l'analyse variographique permettant l'étude des structures de dépendance spatiale, ainsi que les fondements théoriques du krigeage. Enfin, une application directe sera réalisée sur des données simulées grâce au langage *Python*.

Chapitre 1

Présentation du Krigeage

1.1 Définitions et notations

Nous définissons ci-dessous le vocabulaire et les notations utilisées pour la suite des recherches.

Definition 1 *Champ* : région spatiale utilisée pour l'étude en question. On le note D.

Definition 2 *Variable régionalisée* : mesure localisée dans le champ D représentant le phénomène étudié. Elle est notée $\{z(s), s \in D\}$, où $s = (x, y)$ est un point du champ localisé par ses coordonnées.

Definition 3 *Valeur régionalisée* : lorsque cette variable est appliquée en un point particulier $s_i \in D$, on parle de valeur régionalisée. On la note $z(s_i)$.

Definition 4 *Sites d'observation* : comme introduit précédemment, les valeurs régionalisées sont discrétisées dans le champ D, il n'est généralement pas possible de mesurer le phénomène étudié en chaque point. Ainsi, seul un nombre fini de sites est observé. On note ces sites d'observations $s_i, i \in \{1, ..., n\}, n \in \mathbf{N}^*$, et $S = \{s_1, ..., s_n\}$ l'ensemble des sites d'observation.

Definition 5 *Prédiciton d'une valeur régionalisée* : soit un site $s_0 \in D$ dont la valeur n'est pas mesurée. Le krigeage permet d'interpoler cette valeur grâce aux sites d'observation :
$\hat{z}(s_0) = f(z(s_1), ..., z(s_n))$.

Definition 6 *Processus stochastique* : la détermination de la valeur d'un site s_0 peut être déterministe mais également stochastique. C'est le cas pour le krigeage. Dans ce cas, on redéfinit notre variable généralisée comme une fonction aléatoire $\{Z(s), s \in D\}$.

1.2 Krigeage

Différentes méthodes d'interpolation spatiale ont été développées, qu'elles soient déterministes ou stochastiques, comme par exemple les méthodes barycentriques, les splines, la régression classique. Nous nous concentrons ici sur le krigeage.

1.2.1 Estimation

L'intérêt principal du krigeage est que cette méthode prend en compte la structure de dépendance spatiale dans l'interpolation des points géographiques. L'idée est d'estimer la valeur d'une variable régionalisée en un site s_0 par une combinaison linéaire des données mesurées en les i sites d'observation.

$$\hat{z}(s_0) = a + \sum_{i=1}^{n} \lambda_i z(s_i) \tag{1.1}$$

L'objectif est alors de déterminer les paramètres a et $\lambda_i, i \in \{1, ..., n\}$, afin de minimiser la variance des erreurs sous la contrainte de non-biais, en fonction de la structure de dépendance spatiale.

1.2.2 Modèle

Soit le modèle de base du krigeage :

$$Z(s) = \mu(s) + \delta(s), s \in D \tag{1.2}$$

où :

- $\mu(s)$ est la structure déterministe pour l'espérance de $Z(.)$;
- $\delta(s)$ est la structure aléatoire stationnaire, d'espérance nulle et de structure de dépendance connue.

Il existe plusieurs types de krigeage :

- le krigeage simple : $Z(s) = m + \delta(s)$, où m est une constante connue ;
- le krigeage ordinaire : $Z(s) = \mu + \delta(s)$, où μ est une constante inconnue ;
- le krigeage universel : $Z(s) = \sum_{j=0}^{p} f_j(s)\beta_j + \delta(s)$, où la tendance $\mu(.)$ est une combinaison linéaire de fonctions de s (ce type de krigeage ne sera pas développé dans ce travail de recherche).

Le krigeage consiste donc à étudier dans un premier temps le comportement de dépendance spatiale du modèle, inclus dans la fonction aléatoire $\delta(.)$ (voir Chapitre 2), puis de déterminer les paramètres a et λ_i afin de respecter les contraintes de non-biais et de minimisation de la

variance de prédiction, en fonction du modèle de krigeage (voir Chapitre 3). En réinjectant dans (1.1), on peut alors prédire les valeurs de n'importe quel site $s_0 \in D$.

Le krigeage est une méthode d'interpolation exacte, les valeurs prédites aux sites observés sont égales à la valeur mesurée. Ainsi, afin de tester la qualité de prévisions réalisées, il est nécessaire de séparer les sites d'observation en un échantillon d'entraînement (sur lequel les recherches de paramètres seront réalisées) et un échantillon test (sur lequel les prédictions seront réalisées et une mesure d'erreur sera calculée).

Chapitre 2

Analyse Variographique

La première étape du krigeage consiste à rechercher un modèle pouvant décrire la structure de dépendance du champ D grâce aux valeurs des sites d'observation s_i.

2.1 Stationnarité

L'unicité de la réalisation de notre phénomène naturel régionalisé entraîne l'unicité de la fonction aléatoire $Z(.)$ observable, rendant l'inférence statistique impossible. Ainsi, Matheron introduit de nouvelles hypothèses sur la fonction aléatoire $Z(.)$, afin de « réduire le nombre de paramètres dont dépend sa loi »[2]. Ci-dessous les différents types de stationnarité.

Definition 7 Stationnarité stricte

La loi de probabilité de la fonction aléatoire $Z(.)$ est invariante par translation, c'est-à-dire que la distribution jointe de $Z(s_i)$ est la même que celle de $Z(s_i + h)$, h étant un vecteur de translation par rapport à l'origine s_i. Toutes les caractéristiques de la fonction $Z(.)$ restent les mêmes par translation.

Cette définition de la stationnarité est cependant très restrictive et ne représente que rarement des situations réelles. Matheron introduit donc une stationnarité plus faible.

Definition 8 Stationnarité de second ordre

Les conditions relatives à la loi de probabilité de $Z(.)$ ne concernent que les moments d'ordre 1 et 2, dont les indicateurs sont invariants par translation.

- L'espérance de $\delta(.)$ existe et est la même en tout site : $\mathbb{E}[\delta(s)] = u, \forall s \in D$, où u est constante. Dans le cas du krigeage, l'espérance est même nulle.

- L'invariance par translation de l'espérance implique la constance de $\mu(.)$:
$\mathbb{E}[Z(s)] = \mathbb{E}[\mu(s) + \delta(s)] = \mathbb{E}[\mu(s)] = \mu(s)$.
Ainsi, $\mu(s + h) = \mu(s) = \mu, \forall s$.

- La variance est constante : $\mathbb{V}(\delta(s)) = \mathbb{E}[\delta^2(s)] = \sigma^2$.

- La covariance de $\delta(.)$ entre toute paire de sites distancés par un vecteur de translation h existe et dépend uniquement de h, et ce, pour tout $h \in \mathbb{R}^+$: $Cov(\delta(s), \delta(s + h)) = C(h)$. La fonction $C(h)$ est appelée **covariogramme**.

En pratique, ce niveau de stationnarité reste encore trop fort, la moyenne n'étant généralement pas strictement constante sur le champ complet et la variance pouvant ne pas être bornée. On a alors la stationnarité suivante, encore plus faible :

Definition 9 Stationnarité intrinsèque

Les conditions d'invariance par translation sont vérifiées sur les accroissements de la fonction aléatoire. Les accroissements sont stationnaires sans que le processus lui-même le soit.

- L'espérance des accroissements existe et est nulle :

$$\mathbb{E}[\delta(s+h) - \delta(s)] = 0 \quad \forall s \in D$$

.

- La variance de tout accroissement existe et dépend uniquement du vecteur de translation h :

$$\mathbb{V}(\delta(s+h) - \delta(s)) = 2\gamma(h) \quad \forall s, s+h \in D$$

La fonction $\gamma(h)$ est alors appelée **semi-variogramme**.

La stationnarité du second ordre entraîne la stationnarité intrinsèque.

$$\begin{aligned}\mathbb{V}(\delta(s+h) - \delta(s)) &= \mathbb{V}(\delta(s+h)) + \mathbb{V}(\delta(s)) - 2Cov(\delta(s+h), \delta(s)) \\ &= 2C(0) - 2C(h) \\ &= 2(C(0) - C(h)) \\ &= 2\gamma(h)\end{aligned}$$

En revanche l'implication inverse n'est vraie que si $\gamma(.)$ est bornée (résultat que nous admettrons). Ainsi, l'hypothèse de stationnarité intrinsèque est plus générale.

2.2 Propriétés du semi-variogramme

Proposition 1 *Une fonction $f(.)$ est de type négatif conditionnel si elle vérifie la condition suivante :*

$$\sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j f(s_i - s_j) \leq 0$$

pour tout $\{s_i : i = 1, ...l\}$ et pour tout $\{a_i : i = 1, ..., l\}$ tel que $\sum_{i=1}^{l} a_i = 0$.

Theorem 1 *Une fonction continue représente un semi-variogramme si et seulement si elle est de type négatif conditionnel.*

Soit $\gamma(h)$ le semi-variogramme vérifiant la condition précédente. On a alors les propriétés suivantes :

- $\gamma(h) = \gamma(-h) \rightarrow$ la fonction est paire ;
- $\gamma(0) = 0$;
- $\gamma(h) > 0, \forall h \neq 0$.

2.2.1 Isotropie

Un semi-variogramme est dit isotrope s'il ne dépend que de la norme du vecteur h et non de sa direction. S'il prend la direction en compte, il est dit anisotrope. Par la suite, on ne considèrera que les semi-variogrammes isotropes. Ainsi, le semi-variogramme devient :

$$\gamma(r) = \tfrac{1}{2}\mathbb{V}(\delta(s+h) - \delta(s)),$$

où $r = |h| = \sqrt{x_h^2 + y_h^2}$.

2.2.2 Effet de pépite

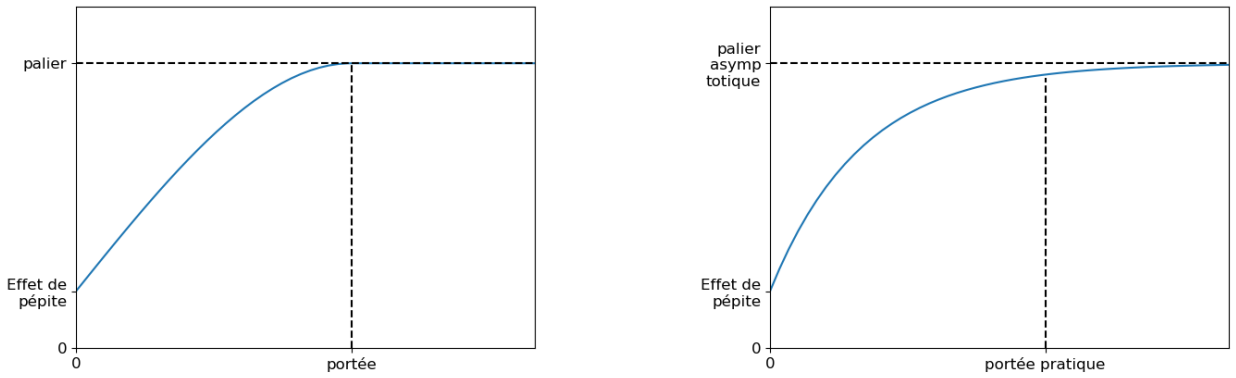
On appelle effet de pépite la valeur c_0 telle que $\lim_{r\to 0^+} \gamma(r) = c_0 > 0$.

Si $c_0 = 0$, il y a absence d'effet de pépite. Cela signifie qu'aucune variabilité n'est détectée à très petite échelle. En revanche, si le saut est trop brusque à l'origine, cela marque une absence de similarité entre deux sites très proches : on parle alors de pépite pure. Attention donc à choisir un modèle de semi-variogramme dont l'effet de pépite n'est ni nul ni pur.

2.2.3 Seuil et portée

On appelle seuil, ou palier, la valeur γ_s telle que $\lim_{r\to+\infty} \gamma(r) = \gamma_s$. La portée indique alors la distance à laquelle il n'y a plus de dépendance spatiale significative, soit la distance minimale telle que le seuil est atteint. Lorsque la valeur du palier est atteinte de manière asymptotique, la portée n'est pas définie, on introduit donc la notion de portée pratique correspondant à la distance minimale telle que le semi-variogramme soit supérieur à 95% de la valeur du seuil.

Attention, lorsque le semi-variogramme n'est pas borné, il ne possède ni portée ni seuil. C'est le cas pour les fonctions stationnaires intrinsèques non stationnaires du second ordre.



(a) (b)

FIGURE 2.1 – (a) Palier et portée réels (b) Portée pratique et palier asymptotique

2.3 Estimation variographique

Dans le cas des krigeages simple et ordinaire stationnaires du second ordre ou intrinsèques, le semi-variogramme peut être développé de la façon suivante :

$$\begin{aligned}
\gamma(r) &= \frac{1}{2}\mathbb{V}(\delta(s+h) - \delta(s)) \\
&= \frac{1}{2}\mathbb{V}(\delta(s) - \delta(s+h)) \\
&= \frac{1}{2}\mathbb{V}(Z(s) - \mu(s) - Z(s+h) + \mu(s+h)) \\
&= \frac{1}{2}\mathbb{V}(Z(s) - Z(s+h)) \\
&= \frac{1}{2}\mathbb{E}[(Z(s) - Z(s+h))^2] - \frac{1}{2}\underbrace{\mathbb{E}[Z(s) - Z(s+h)]^2}_{=0}
\end{aligned}$$

Le semi-variogramme peut alors être estimé de manière empirique par :

$$\hat{\gamma}(r) = \frac{1}{2|N(r)|} \sum_{N(r)} (z(s_i) - z(s_j))^2$$

où $N(r) = \{(i,j) \text{ tel que } |s_i - s_j| = r\}$ et $|N(r)|$ est le nombre de couples distincts de l'ensemble $N(r)$.

Dans la pratique, l'estimation $\hat{\gamma}(r)$ perd cependant en précision à partir du moment où l'on se place dans un champ où les sites d'observation sont répartis de façon irrégulière, augmentant le nombre de distances r différentes, chacune représentée par un faible nombre de couples de données. Ainsi, une tolérance peut être attribuée aux distances ente deux sites d'observation, de telle sorte que l'on ait $N(r_k) = \{(i,j) \text{ tel que } |s_i - s_j| = r_k \pm b_k\}$, où b_k est un paramètre de tolérance à déterminer. L'optimisation de ce paramètre est également très coûteuse en calcul, et est donc pratiquement irréalisable en pratique. Nous ne retiendrons donc pas cette méthode pour la suite.

2.4 Modélisation du semi-variogramme

Le calcul du semi-variogramme empirique repose sur un nombre fini de variables et donc de distances. Afin de déterminer la valeur du semi-variogramme pour l'ensemble des distances $r \in \mathbb{R}$, il est donc nécessaire de modéliser $\gamma(.)$ à l'aide d'une fonction conditionnellement négative. Cependant, déterminer si une fonction est conditionnelle négative n'est pas évident. Par la suite, nous chercherons le modèle fournissant la meilleure approximation parmi un ensemble de modèles vérifiant cette condition. Nous retiendrons les modèles suivants :

- **Modèle linéaire avec palier**, de portée exacte a, d'effet de pépite c_0, et de palier $c + c_0$:

$$\gamma(r) = \begin{cases} c_0 + \frac{c}{a} \times r \ si \ 0 \leq r \leq a \\ c_0 + c \ si \ r \geq a \end{cases}$$

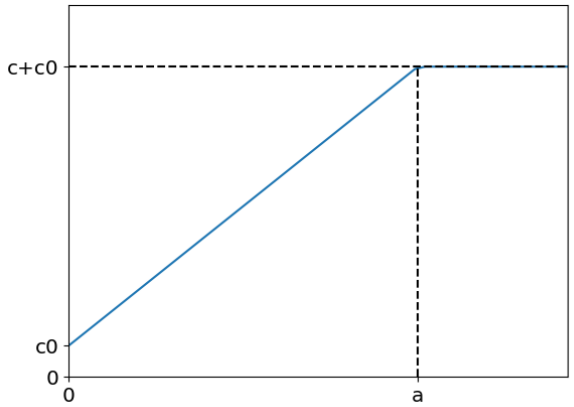


FIGURE 2.2 – Modèle linéaire avec palier

- **Modèle linéaire sans palier**, d'effet de pépite c_0, et de pente m :

$$\gamma(r) = c_0 + mr \ pour \ r \geq 0$$

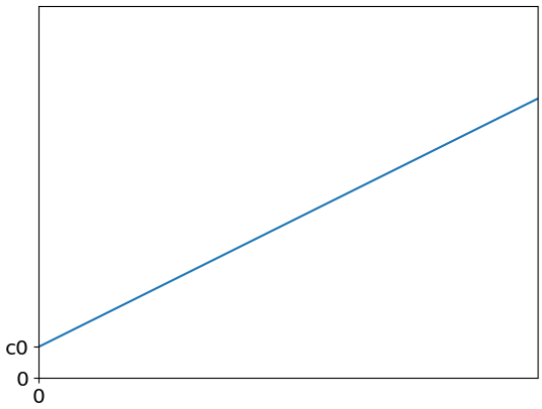


FIGURE 2.3 – Modèle linéaire sans palier

- **Modèle puisssance sans palier**, d'effet de pépite c_0, d'exposant ν et de facteur d'échelle m :

$$\gamma(r) = c_0 + mr^{\nu} \; pour \; r \geq 0, \; 0 \leq \nu \leq 2$$

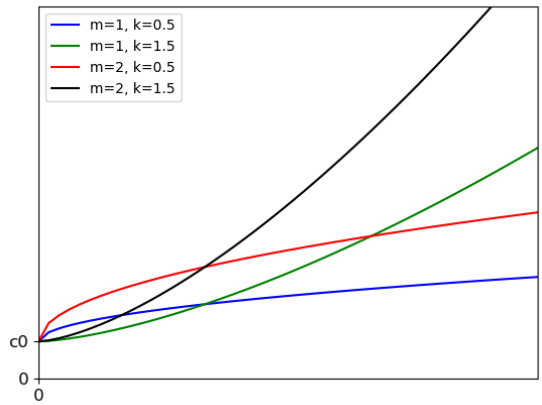


FIGURE 2.4 – Modèle puissance sans palier

- **Modèle exponentiel** d'effet de pépite c_0, de palier $c + c_0$ et portée pratique $3a$:

$$\gamma(r) = c_0 + c(1 - \exp(-\frac{r}{a})) \; si \; r \geq 0$$

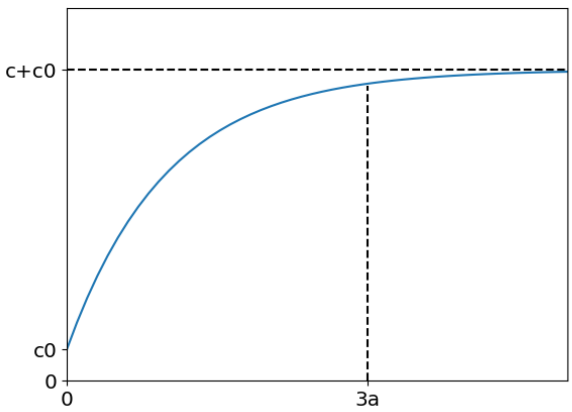


FIGURE 2.5 – Modèle exponentiel

- **Modèle gaussien** d'effet de pépite c_0, de palier $c + c_0$ et portée pratique $\sqrt{3}a$:

$$\gamma(r) = c_0 + c(1 - \exp(-(\frac{r}{a})^2)) \; si \; r \geq 0$$

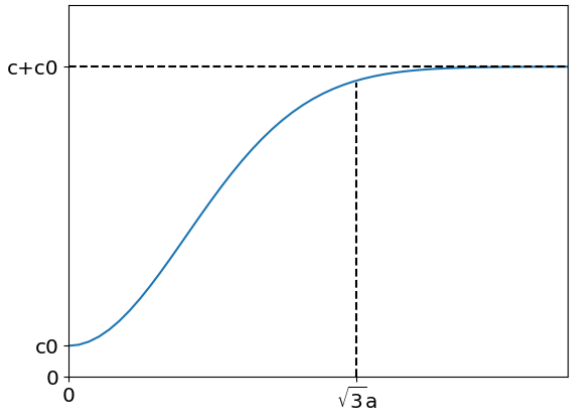


FIGURE 2.6 – Modèle gaussien

Remarques :

— Toute somme de modèles cités ci-dessus est également conditionnellement négative et donc représente un modèle de semi-variogramme valide.

— Les modèles ne présentant pas de palier et étant donc non bornés seront traités sous l'hypothèse de stationnarité intrinsèque.

— La valeur de la portée pratique dans le cadre des modèles gaussien et exponentiel est issue de la littérature.

Chapitre 3

Théorie du krigeage

3.1 Equations du Krigeage

Dans cette partie, nous étudierons les cas des krigeages simple et ordinaire, soit les cas où la valeur de la variable aléatoire au site s est de la forme :

$$Z(s) = m + \delta(s)$$

où m est une constante, connue dans le cas du krigeage simple et inconnue dans le cas du krigeage ordinaire.

Notons $(Z(s_1), ..., Z(s_n))$ l'ensemble des valeurs observées aux sites $s_1, ..., s_n$. L'estimation de la valeur en un site non observé s_0 par la méthode du krigeage s'effectue en respectant les contraintes suivantes :

1. Contrainte de linéarité :
La valeur $\hat{Z}(s_0)$ doit être une combinaison linéaire des observations :

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

où λ_i est le poids associé à la variable $Z(s_i)$ et a est une constante.

2. Contrainte d'autorisation :
L'espérance et la variance de la prédiction existent.

3. Contrainte de non biais :
La prédiction doit être non biaisée.

4.Contrainte d'Optimalité :
La variance de l'erreur de prédiction doit être minimale.

Ces quatre contraintes nous permettent d'aboutir aux équations du krigeage.
Afin de simplifier les calculs qui vont suivre, nous posons les notations matricielles suivantes :
- $\mathbf{Z}$ le vecteur de taille $n \times 1$ contenant les variables aléatoires $Z(s_{[1]}), ..., Z(s_{[n]})$;
- $\boldsymbol{\lambda}$ le vecteur de taille $n \times 1$ contenant les poids associés aux variables aléatoires ;
- $\boldsymbol{\delta}$ le vecteur de taille $n \times 1$ contenant les erreurs associées aux variables aléatoires.

3.1.1 Stationnarité de second ordre

Plaçons-nous d'abord dans le cadre où le semi-variogramme $\gamma(.)$ est supposé stationnaire de second ordre et posons les notations suivantes :
- $\boldsymbol{\Sigma}$ la matrice de variances-covariances de $\delta(.)$ de taille $n \times n$;
- $\mathbf{c_0}$ la matrice des covariances entre les $\delta(\mathbf{s_i})$ et $\delta(\mathbf{s_0})$, de taille $n \times 1$.

3.1.1.1 Contrainte de linéarité

L'erreur de prédiction au point s_0 peut s'écrire de la façon suivante :

$$\hat{Z}(s_0) - Z(s_0) = \hat{Z}(s_0) - m - \delta(s_0)$$

En appliquant la contrainte selon laquelle la prédiction est une combinaison linéaire des estimations, cette quantité s'écrit :

$$\begin{aligned}
\hat{Z}(s_0) - Z(s_0) &= a + \sum_{i=1}^{n} \lambda_i Z(s_i) - m - \delta(s_0) \\
&= a + \sum_{i=1}^{n} \lambda_i (m + \delta(s_i)) - m - \delta(s_0) \\
&= a - m \left(1 - \sum_{i=1}^{n} \lambda_i\right) + \sum_{i=1}^{n} \lambda_i \delta(s_i) - \delta(s_0)
\end{aligned}$$

3.1.1.2 Contrainte d'autorisation

L'espérance et la variance des erreurs de prédiction peuvent s'écrire de la façon suivante :

$$\begin{aligned}\mathbb{E}\left[\hat{Z}(s_0) - Z(s_0)\right] &= \mathbb{E}\left[a - m\left(1 - \sum_{i=1}^{n}\lambda_i\right) + \sum_{i=1}^{n}\lambda_i\delta(s_i) - \delta(s_0)\right] \\ &= a - m\left(1 - \sum_{i=1}^{n}\lambda_i\right) + \sum_{i=1}^{n}\lambda_i\mathbb{E}\left[\delta(s_i)\right] - \mathbb{E}\left[\delta(s_0)\right] \\ &= a - m\left(1 - \sum_{i=1}^{n}\lambda_i\right)\end{aligned}$$

car a, m et λ_i sont non aléatoires et, par hypothèse, les variables $\delta(s_i)$ sont centrées.

$$\begin{aligned}\mathbb{V}\left[\hat{Z}(s_0) - Z(s_0)\right] &= \mathbb{V}\left[a - m\left(1 - \sum_{i=1}^{n}\lambda_i\right) + \sum_{i=1}^{n}\lambda_i\delta(s_i) - \delta(s_0)\right] \\ &= \mathbb{V}\left[\sum_{i=1}^{n}\lambda_i\delta(s_i) - \delta(s_0)\right]\end{aligned}$$

De plus, la fonction $\delta(.)$ étant stationnaire de second ordre, ces quantités sont bien définies : la condition d'autorisation est donc elle aussi vérifiée.

3.1.1.3 Contrainte de non-biais

La contrainte de non biais implique que l'espérance de l'erreur de prédiction soit nulle, soit :

$$\begin{aligned}\mathbb{E}[\hat{Z}(s_0) - Z(s_0)] &= 0 \\ a - m\left(1 - \sum_{i=1}^{n}\lambda_i\right) &= 0 \\ a &= m\left(1 - \sum_{i=1}^{n}\lambda_i\right) \\ a &= m\left(1 - \boldsymbol{\lambda}^t\mathbf{1}_n\right)\end{aligned}$$

3.1.1.4 Contrainte d'optimalité

Minimisons maintenant la variance de l'erreur de prédiction.

$$\begin{aligned}\mathbb{V}\left[\hat{Z}(s_0) - Z(s_0)\right] &= \mathbb{V}\left[\sum_{i=1}^{n}\lambda_i\delta(s_i) - \delta(s_0)\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n}\lambda_i\delta(s_i) - \delta(s_0)\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n}\lambda_i\delta(s_i)\right)^2\right] - 2\sum_{i=1}^{n}\lambda_i\mathbb{E}\left[\delta(s_i)\delta(s_0)\right] + \mathbb{E}\left[\delta(s_0)^2\right] \\ &= \boldsymbol{\lambda}^t\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^t\mathbf{c_0} + \sigma^2\end{aligned}$$

Posons la fonction f telle que $f(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t \boldsymbol{\Sigma} \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^t \mathbf{c_0} + \sigma^2$. Minimiser l'erreur de prédiction consiste à annuler la dérivée de f par rapport à $\boldsymbol{\lambda}$.

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\lambda}} f &= 0 \\ 2\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\mathbf{c_0} &= 0 \\ \hat{\boldsymbol{\lambda}} &= \boldsymbol{\Sigma}^{-1}\mathbf{c_0} \end{aligned}$$

où $\hat{\boldsymbol{\lambda}}$ est bien défini car $\boldsymbol{\Sigma}$ est symétrique définie positive et donc inversible. De plus, $\hat{\boldsymbol{\lambda}}$ est bien un minimum, la dérivée seconde de f étant le produit d'une matrice semi-définie positive et d'une constante postive.
Dans le cadre de la stationnarité de second ordre, on a donc l'estimateur $\hat{Z}(s_0)$ donné par :

$$\begin{aligned} \hat{Z}(s_0) &= a + \boldsymbol{\lambda}^t \mathbf{Z} \\ &= m\left(1 - \boldsymbol{\lambda}^t \mathbf{1}_n\right) + \boldsymbol{\lambda}^t \mathbf{Z} \\ &= m + \boldsymbol{\lambda}^t \left(\mathbf{Z} - m\mathbf{1}_n\right) \\ &= m + \mathbf{c_0}^t \boldsymbol{\Sigma}^{-1} \left(\mathbf{Z} - m\mathbf{1}_n\right) \end{aligned}$$

La variance de l'erreur de prédiction vaut alors :

$$\begin{aligned} \mathbb{V}\left[\hat{Z}(s_0) - Z(s_0)\right] &= \boldsymbol{\lambda}^t \boldsymbol{\Sigma} \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^t \mathbf{c_0} + \sigma^2 \\ &= \sigma^2 - \mathbf{c_0}^t \boldsymbol{\Sigma}^{-1} \mathbf{c_0} \end{aligned}$$

En krigeage simple, la constante m étant connue, on peut calculer directement la variable $\hat{Z}(s_0)$. En revanche, dans le cas du krigeage ordinaire, m est inconnue, il faut donc d'abord l'estimer, par exemple en calculant le moment d'ordre 1 sur les sites observés.

3.1.2 Stationnarité intrinsèque

Etudions maintenant le cas où la fonction $\delta(.)$ est supposée staionnaire intrinsèque.
Introduisons les notations suivantes :
- $\boldsymbol{\Gamma}$ la matrice de taille $n \times n$, dont l'élément (i, j) correspond au semi-variogramme de l'accroissement entre $\delta(s_i)$ et $\delta(s_j)$;
- $\boldsymbol{\gamma}_0$ le vecteur de taille $n \times 1$, dont le i-ème élément correspond au semi-variogramme de l'accroissement entre $\delta(s_i)$ et $\delta(s_0)$.

3.1.2.1 Contraintes de linéarité et de non-biais

Comme dans le cas de la stationnarité de second ordre, les contraintes de linéarité et de non biais nous donnent les équations suivantes :

$$\hat{Z}(s_0) = m + \boldsymbol{\lambda}^t (\mathbf{Z} - m\mathbf{1}_n)$$

et

$$\mathbb{V}\left[\hat{Z}(s_0) - Z(s_0)\right] = \mathbb{V}\left[\sum_{i=1}^{n} \lambda_i \delta(s_i) - \delta(s_0)\right]$$

.

3.1.2.2 Contrainte d'autorisation

Contrairement au cas précédent, la contrainte d'autorisation n'est pas nécessairement vérifiée. En effet, l'hypothèse de stationnairité intrinsèque permet seulement d'affirmer que les combinaisons linéaires d'accroissement de $\delta(.)$ possèdent une variance finie et donc respectent la contrainte d'autorisation. Afin de respecter cette condition, il est donc nécessaire de poser $\sum_{i=1}^{n} \lambda_i = 1$. Dans ce cas, on a :

$$\begin{aligned}\hat{Z}(s_0) &= m + \boldsymbol{\lambda}^t \mathbf{Z} - m\boldsymbol{\lambda}^t \mathbf{1}_n \\ &= \boldsymbol{\lambda}^t \mathbf{Z}\end{aligned}$$

et donc :

$$\begin{aligned}\hat{Z}(s_0) - Z(s_0) &= \sum_{i=1}^{n} \lambda_i \delta(s_i) - \delta(s_0) \\ &= \sum_{i=1}^{n} \lambda_i \delta(s_i) - \sum_{i=1}^{n} \lambda_i \delta(s_0) \\ &= \sum_{i=1}^{n} \lambda_i (\delta(s_i) - \delta(s_0))\end{aligned}$$

L'erreur de prédiction est donc une combinaison linéaire d'accroissements de la fonction $\delta(.)$, assurant la bonne définition de la variance.

Nous pouvons donc minimiser cette variance sous contrainte.

3.1.2.3 Contrainte d'optimalité

$$\begin{aligned}\mathbb{V}\left[\hat{Z}(s_0) - Z(s_0)\right] &= \mathbb{V}\left[\sum_{i=1}^{n} \lambda_i \delta(s_i) - \delta(s_0)\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \lambda_i \delta(s_i) - \delta(s_0)\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \lambda_i \delta(s_i)\right)^2 - 2\delta(s_0)\sum_{i=1}^{n} \lambda_i \delta(s_i) + \delta(s_0)^2\right] \\ &= \mathbb{E}\left[\underbrace{\left(\sum_{i=1}^{n} \lambda_i \delta(s_i)\right)^2 - \sum_{i=1}^{n} \lambda_i \delta(s_i)^2}_{1} + \underbrace{\sum_{i=1}^{n} \lambda_i \delta(s_i)^2 - 2\delta(s_0)\sum_{i=1}^{n} \lambda_i \delta(s_i) + \delta(s_0)^2}_{2}\right]\end{aligned}$$

En notant T_n le terme 1, on obtient :

$$\begin{aligned}
T_n &= \sum_{j=1}^{n} \lambda_j \delta(s_j) \sum_{i=1}^{n} \lambda_i \delta(s_i) - \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{n} \lambda_i \delta(s_i)^2 \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i \delta(s_j)\delta(s_i) - \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i \delta(s_i)^2 \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i [\delta(s_j)\delta(s_i) - \delta(s_i)^2] \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i [-\frac{1}{2}[\delta(s_i)^2 - 2\delta(s_j)\delta(s_i) + \delta(s_j)^2 - \delta(s_j)^2 + \delta(s_i)^2]] \\
&= -\frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i [[\delta(s_i) - \delta(s_j)]^2 - \delta(s_j)^2 + \delta(s_i)^2] \\
&= -\frac{1}{2} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i (\delta(s_i) - \delta(s_j))^2 - \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i \delta(s_j)^2 + \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i \delta(s_i)^2] \right) \\
&= -\frac{1}{2} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i (\delta(s_i) - \delta(s_j))^2 - \underbrace{\sum_{i=1}^{n} \lambda_i}_{=1} \sum_{j=1}^{n} \lambda_j \delta(s_j)^2 + \underbrace{\sum_{j=1}^{n} \lambda_j}_{=1} \sum_{i=1}^{n} \lambda_i \delta(s_i)^2 \right) \\
&= -\frac{1}{2} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i (\delta(s_i) - \delta(s_j))^2 - \sum_{j=1}^{n} \lambda_j \delta(s_j)^2 + \sum_{i=1}^{n} \lambda_i \delta(s_i)^2 \right) \\
&= -\frac{1}{2} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i (\delta(s_i) - \delta(s_j))^2 \right)
\end{aligned}$$

En isolant le terme 2, on obtient :

$$\begin{aligned}
\sum_{i=1}^{n} \lambda_i \delta(s_i)^2 - 2\delta(s_0) \sum_{i=1}^{n} \lambda_i \delta(s_i) + \delta(s_0)^2 &= \sum_{i=1}^{n} \lambda_i \delta(s_i)^2 - 2 \sum_{i=1}^{n} \lambda_i \delta(s_0)\delta(s_i) + \sum_{i=1}^{n} \lambda_i \delta(s_0)^2 \\
&= \sum_{i=1}^{n} \lambda_i (\delta(s_i) - \delta(s_0))^2
\end{aligned}$$

On a donc :

$$\begin{aligned}
\mathbb{V}\left[\hat{Z}(s_0) - Z(s_0)\right] &= \mathbb{E}\left[-\frac{1}{2} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i (\delta(s_i) - \delta(s_j))^2 \right) + \sum_{i=1}^{n} \lambda_i (\delta(s_i) - \delta(s_0))^2 \right] \\
&= \sum_{i=1}^{n} \lambda_i \mathbb{E}[(\delta(s_i) - \delta(s_0))^2] - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i \mathbb{E}[(\delta(s_i) - \delta(s_j))^2] \\
&= \sum_{i=1}^{n} \lambda_i \mathbb{V}[\delta(s_i) - \delta(s_0)] - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i \mathbb{V}[\delta(s_i) - \delta(s_j)] \\
&= 2 \sum_{i=1}^{n} \lambda_i \gamma(\delta(s_i) - \delta(s_0)) - \sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_j \lambda_i \gamma(\delta(s_i) - \delta(s_j)) \\
&= 2\boldsymbol{\lambda}^t \boldsymbol{\gamma}_0 - \boldsymbol{\lambda}^t \boldsymbol{\Gamma} \boldsymbol{\lambda}
\end{aligned}$$

On minimise alors l'erreur de prédiction sous la contrainte $\sum_{i=1}^{n} \lambda_i = 1$ en annulant la derivée du Lagrangien $L(\boldsymbol{\lambda}, l) = 2\boldsymbol{\lambda}^t\boldsymbol{\gamma}_0 - \boldsymbol{\lambda}^t\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2l(\boldsymbol{\lambda}^t\mathbf{1}_n - 1)$ par rapport à $\boldsymbol{\lambda}$.

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\lambda}} L(\boldsymbol{\lambda}, l) &= 0 \\ 2\boldsymbol{\gamma}_0 - 2\boldsymbol{\Gamma}\boldsymbol{\lambda} + 2l\mathbf{1}_n &= 0 \\ \boldsymbol{\lambda} &= \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 + l\mathbf{1}_n) \end{aligned}$$

De plus on a :

$$\begin{aligned} \mathbf{1}_n^t\boldsymbol{\lambda} &= 1 \\ \mathbf{1}_n^t\Gamma^{-1}(\boldsymbol{\gamma}_0 + l\mathbf{1}_n) &= 1 \\ l &= \frac{1 - \mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0}{\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_n} \end{aligned}$$

Alors :

$$\boldsymbol{\lambda}^* = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 + \frac{1 - \mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0}{\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_n}\mathbf{1}_n)$$

La prévision d'un site s_0 est alors :

$$\hat{Z}(s_0) = (\boldsymbol{\gamma}_0 + \frac{1 - \mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0}{\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_n}\mathbf{1}_n)^t\boldsymbol{\Gamma}^{-1}\mathbf{Z}$$

Posons $r = \frac{1-\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0}{\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_n} \in \mathbb{R}$. Alors la variance de l'erreur de prédiction est égale à :

$$\begin{aligned} \mathbb{V}[\hat{Z}(s_0) - Z(s_0)] &= 2\left(\boldsymbol{\gamma}_0 + r\mathbf{1}_n\right)^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0 - \left(\boldsymbol{\gamma}_0 + r\mathbf{1}_n\right)^t\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{\gamma}_0 + r\mathbf{1}_n\right) \\ &= (\boldsymbol{\gamma}_0^t + r\mathbf{1}_n^t)\boldsymbol{\Gamma}^{-1}(\boldsymbol{\gamma}_0 - r\mathbf{1}_n) \\ &= \boldsymbol{\gamma}_0^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0 - r^2\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_n \\ &= \boldsymbol{\gamma}_0^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0 - \frac{(1 - \mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0)^2}{(\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_n)^2}(\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_n) \\ &= \boldsymbol{\gamma}_0^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0 - \frac{(1 - \mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0)^2}{\mathbf{1}_n^t\boldsymbol{\Gamma}^{-1}\mathbf{1}_n} \end{aligned}$$

3.2 Transformation des données

Dans certains cas, les valeurs d'un phénomène peuvent être bornées, par exemple dans le cadre d'une étude épidémiologique où le nombre de personnes malades ne peut être négatif, ou dans le cadre d'études géologiques où la concentration d'un minerai doit être comprise entre 0 et 1. Cependant, dans les méthodes présentées précédemment, rien n'oblige les prévisions du krigeage à respecter ces contraintes. Pour tenir compte des bornes, il est possible d'introduire des contraintes sur les poids λ_i en les obligeant à être positifs, ou bien d'appliquer une transformation sur les données observées avant d'effectuer l'estimation via le krigeage. On applique ensuite la transformation inverse sur les prévisions obtenues. Il faut néanmoins prendre en compte que la prévision issue de données transformées peut introduire un facteur de biais qu'il faudra corriger lors de l'application de la transformation inverse des estimations.

Chapitre 4

Simulations

4.1 Simulations des données

Afin de tester nos algorithmes de krigeage, des données spatiales ont été simulées en deux dimensions. Pour cela, nous utilisons un modèle de diffusion de chaleur, en sélectionnant aléatoirement et de manière uniforme 50 points sources notés $\{x_1, ..., x_{50}\}$ dans un rectangle R de dimensions $[0, 10] \times [0, 10]$. Il leur est attribué à chacun une valeur z_k selon une loi uniforme de paramètres $[-10, 10]$. Une grille de taille 100×100 est alors construite dans le rectangle R et il est attribué à chacun des noeuds de la grille une valeur d'après la formule suivante, en fonction des points sources :

$$z_i = \sum_{k=1}^{50} z_k \times \exp(-\frac{h_k^2}{|z_k|})$$

où $h_k = ||x_k - x_i||$ est la distance euclidienne entre les points x_i et x_k.

En deux dimensions nous obtenons le graphique ci-dessus, où les courbes bleu, verte et rouge sont des fonctions de la forme $f(x_i, z_k) = z_k \times \exp(-\frac{h_k^2}{|z_k|})$ et la fonction noir est la somme de ces trois fonctions.

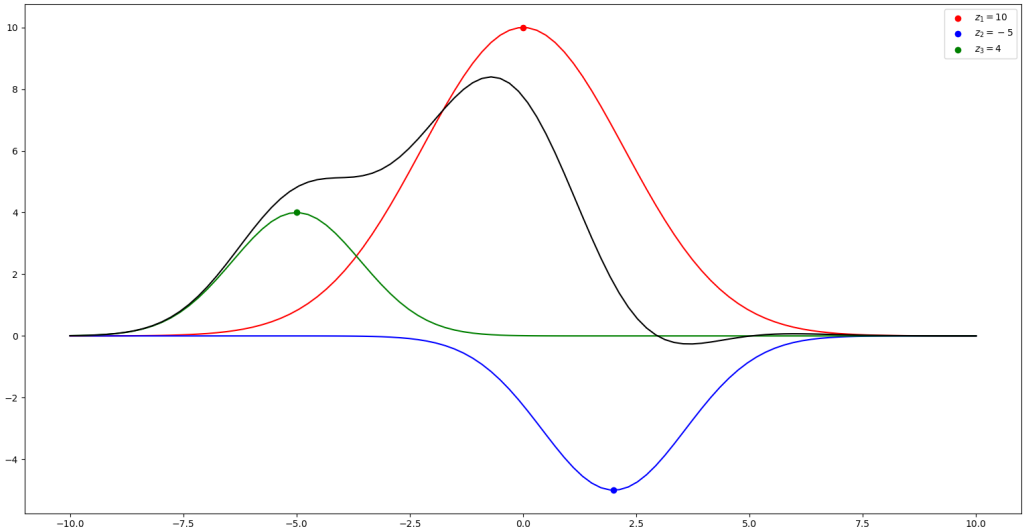


FIGURE 4.1 – Diffusion de chaleur en deux dimensions

On obtiens ainsi le graphe suivant :

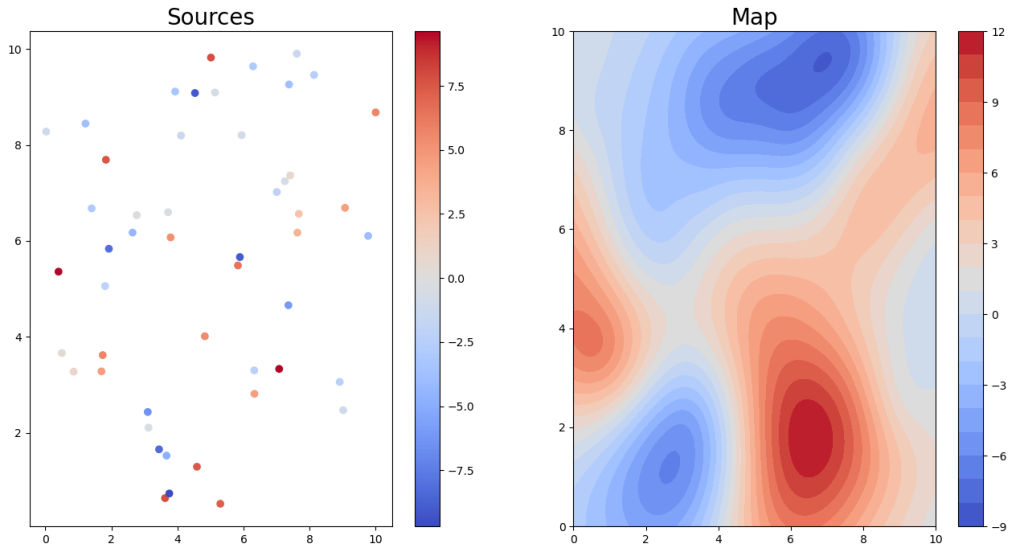


FIGURE 4.2 – Simulation d'une carte en fonction du tirage au sort de 50 sources

4.2 Sélection du modèle

On se place dans un premier temps dans le cas où l'échantillon des sites d'observation est égal à 100, soit 1% de nos données, et tiré aléatoirement dans le jeu de données. La manière dont les sites d'observation sont répartis ainsi que leur nombre jouent un rôle très important dans la qualité de l'estimation, en plus de celui joué par les choix du semi-variogramme et de ses paramètres. Ainsi, pour chacune des méthodes qui suivent, nous estimons le semi-variogramme en divisant l'ensemble de nos sites d'observation en un échantillon d'entraînement, contenant 80% de ces sites, et un échantillon test, contenant les 20% restants. Un affinage sera également réalisé par validation croisée. Nous verrons plus tard l'impact d'un échantillon tiré aléatoirement ou de manière déterministe, ainsi que l'impact du nombre n de sites d'observation. Le krigeage étant une méthode d'interpolation exacte, l'erreur quadratique moyenne entre l'échantillon test et l'estimation est un bon indice de qualité de notre modèle.

4.2.1 Recherche du meilleur semi-variogramme

Avant de pouvoir kriger notre carte en fonction d'un ensemble de n sites d'observation, il faut passer par l'estimation du semi-variogramme de nos données, traduisant la structure de dépendance de la carte. Cette estimation est réalisée par un krigeage sur les données d'entraînement et une minimisation de l'erreur sur nos données test. Ainsi, on souhaite comparer l'erreur quadratique moyenne de notre échantillon test en fonction du modèle de semi-variogramme choisi parmi ceux cités en 2.4.

En ce qui concerne les paramètres de nos modèles, nous traitons les situations suivantes :
- recherche simultanée de deux paramètres sur trois, le troisième étant fixé :
  - pour le modèle puissance :
    - c_0 est fixé à 0, k varie dans $[0, 2]$ et m varie dans $[0, 10]$;
    - k est fixé à 1, c_0 varie dans $[0, 10]$ et m varie dans $[0, 10]$;
    - m est fixé à 1, k varie dans $[0, 2]$ et c_0 varie dans $[0, 10]$;
  - pour les autrees modèles :

- c_0 est fixé à 0, a varie dans $[1, 10]$ et c varie dans $[0, 10]$;
- a est fixé à 2, c_0 varie dans $[0, 10]$ et c varie dans $[0, 10]$;
- c est fixé à 5, a varie dans $[1, 10]$ et c_0 varie dans $[0, 10]$.

- recherche simultanée des trois paramètres.

Aucune différence n'étant visible à l'oeil nu pour chacun des résultats des méthodes précédentes, on ne présente ci-dessus qu'un seul exemple de résultat :

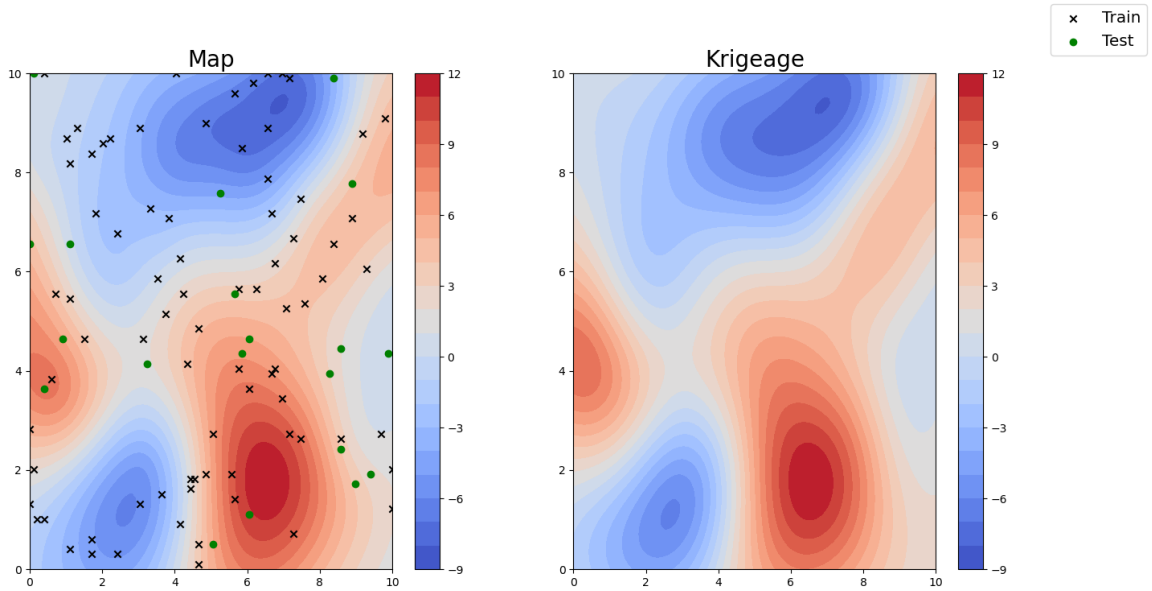


FIGURE 4.3 – $a = 2$ fixé : $c = 0.526$ et $c_0 = 8.94$ après minimisation du MSE : $MSE = 0.0456$

Ci-dessous les résultats et variations observés lors des différentes situations proposées :

- Cas où $c_0 = 0$:

f	c0	c	a	MSE
Lineaire	0.0	4.737	2.421	0.443
Gaussien	0.0	3.684	1.947	0.141
Exponentielle	0.0	3.684	10.0	0.826

- Cas où $c = 5$:

f	c	c0	a	MSE
Lineaire	5.0	4.211	2.421	0.443
Exponentielle	5.0	5.263	10.0	0.826
Gaussien	5.0	6.842	1.947	0.141

- Cas où $a = 2$:

f	a	c0	c	MSE
Lineaire	2.0	10.0	0.526	0.915
Exponentielle	2.0	7.895	1.579	1.106
Gaussien	2.0	8.947	1.053	0.148

- Modèle puissance :

f	c0	m	k	MSE
Puissance	0	0.526	1.895	0.153
Puissance	8.421	1	1.895	0.153
Puissance	7.895	2.105	1	0.806

- Recherche simultanée des trois paramètres : L'estimation du meilleur triplet de paramètres pour chacun de nos modèles étant plus coûteuse en calculs, le nombre de valeurs testées par intervalle d'étude est diminué, passant de 20 à 10. On obtient les résultats suivants :

f	a	c0	c	MSE
Lineaire	2.33	8.89	1.11	0.847
Exponentielle	5.0	8.89	1.11	0.873
Gaussien	1.89	5.56	1.11	0.134

f	k	c0	m	MSE
Puissance	1.78	1.11	1.11	0.195

Le meilleur modèle est donc le modèle gaussien qui nous fournit l'estimation suivante :

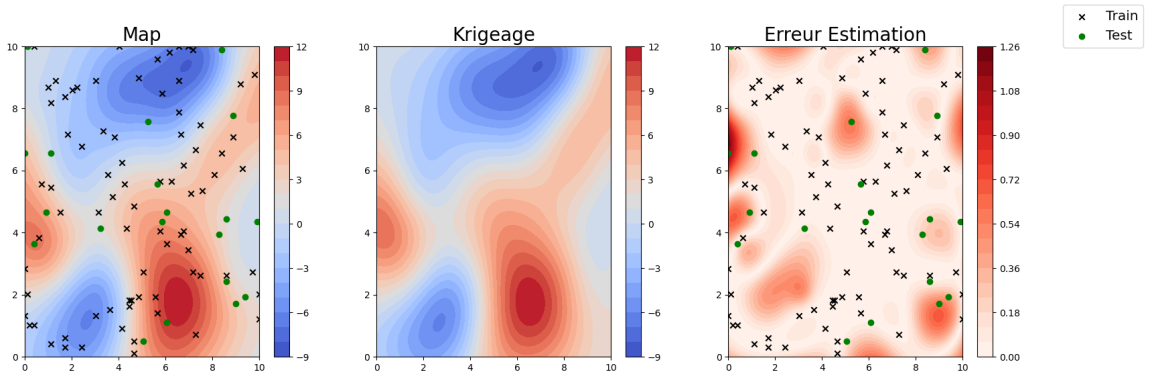


FIGURE 4.4 – Minimisation du MSE : $MSE = 0.042$

Il est également intéressant d'observer l'influence des variations des différents paramètres, en fixant deux d'entre eux et en observant le krigeage réalisé pour un ensemble du troisième. Ci-dessous, nous nous plaçons dans le cas du semi-variogramme gaussien.

- Influence du paramètre a :

On fixe ici nos paramètres a et c_0 à des valeurs proches de celles trouvées dans nos recherches précédentes, puis l'on observe les résultats du krigeage en fonction de différentes valeurs de a.

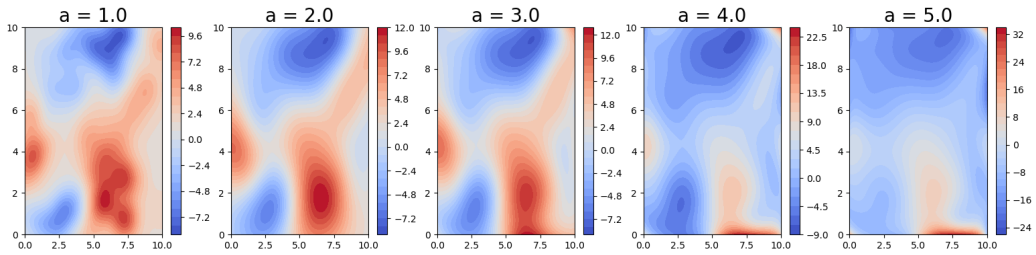


FIGURE 4.5 – Modèle gaussien : $c_0 = 1.5$, $c = 5$

On peut voir que le krigeage est très sensible, à un pas de 1, au paramètre a.
N.B. : Pour des valeurs de a comprises entre 5 et 10, les tendances de températures sont totalement inversées et les prédictions chaotiques.

- Influence du paramètre c : On fixe cette fois-ci nos paramètres c_0 et a afin d'observer l'influence de c.

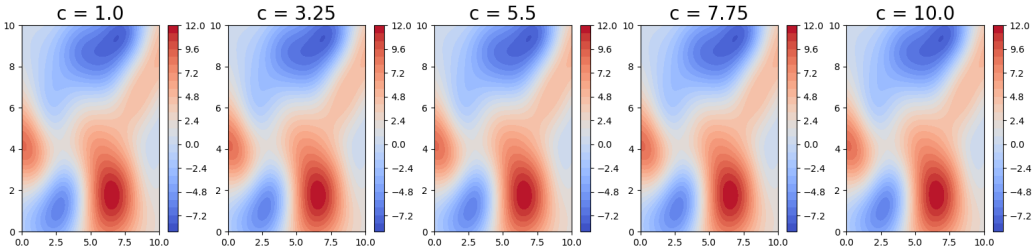


FIGURE 4.6 – Modèle gaussien : $c_0 = 1.5$, $a = 2$

On peut voir qu'à des valeurs de c_0 et a proches de celles trouvées précédemment par la minimisation du MSE, le paramètre c n'a aucun impact visible à l'œil nu sur le krigeage. Après recherche, le paramètre c n'a d'influence visible sur le krigeage que lorsque $a \geq 5$, comme ci-dessous.

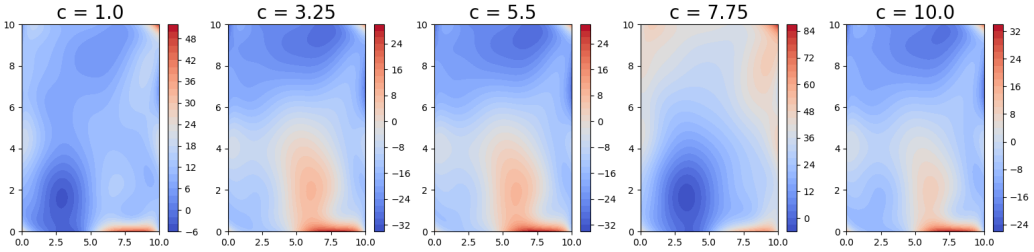


FIGURE 4.7 – Modèle gaussien : $c_0 = 1.5$, $a = 5$

- Variations du paramètre c_0 :

On fixe finalement nos paramètres c et a afin d'observer l'influence de c_0.

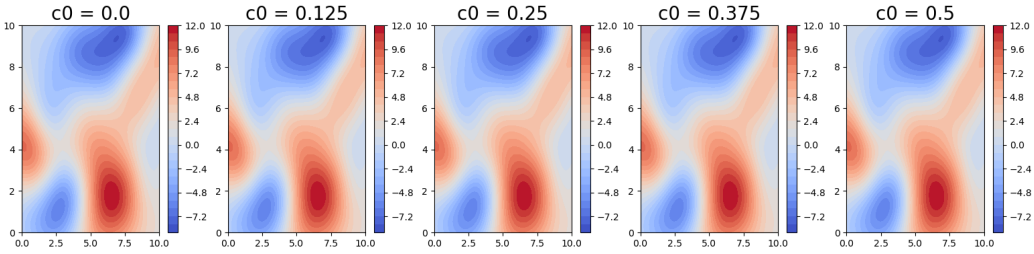


FIGURE 4.8 – Modèle gaussien : $a = 2$, $c = 5$

Encore une fois, le paramètre c_0 a très peu d'impact sur les prédictions du krigeage lorsque a et c sont proches des valeurs trouvées. Son influence ne devient visible qu'à partir du moment où $a \geq 5$.

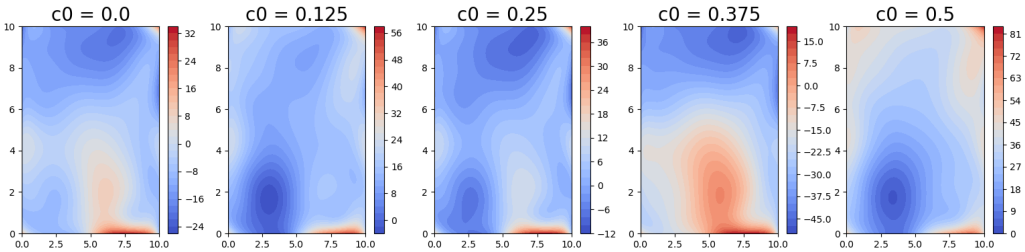


FIGURE 4.9 – Modèle gaussien : $a = 5$, $c = 5$

D'après nos recherches précédentes, le semi-variogramme gaussien semble être le plus adapté aux données de la map simulée. Afin d'optimiser au mieux les paramètres du modèle, il est alors possible de réaliser une recherche par validation croisée. On divise ainsi notre échantillon de sites d'observation en $k = 5$ (choix arbitraire) sous-échantillons. La recherche des meilleurs paramètres est alors réalisée en considérant consécutivement chacun des sous-échantillon comme échantillon test. De plus, on se place dans un cadre de recherche simultanée des trois paramètres optimaux c_0, a et c. On obtient alors le krigeage suivant :

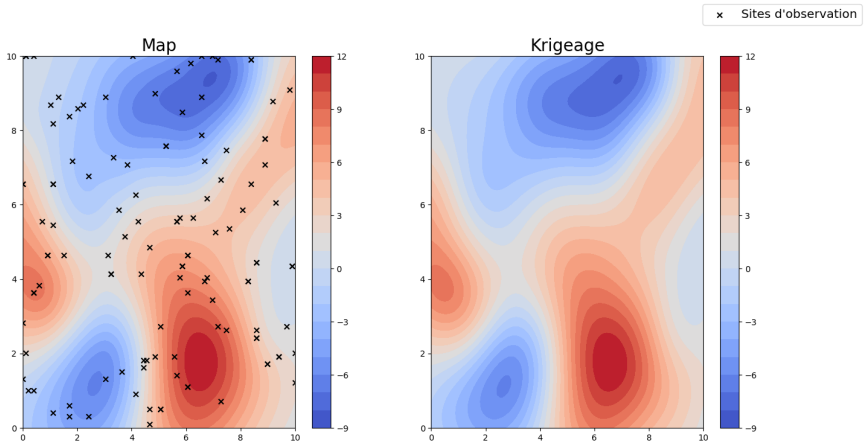


FIGURE 4.10 – Krigeage par validation croisée

où $c_0 = 1.6$, $c = 4.222$ et $a = 2.0$, nous donnant ainsi l'erreur totale suivante : $MSE = 0.0317$.

Afin de visualiser les différences à l'oeil nu, l'écart entre valeurs réelles et krigeage est représenté comme ci-dessous :

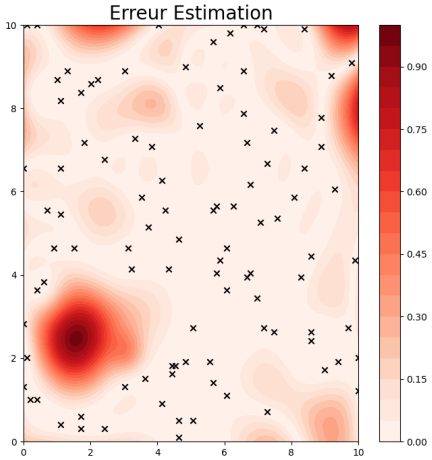


FIGURE 4.11 – Erreur d'estimation

On peut ainsi voir que les erreurs sont de plus en plus élevées dans les zones géographiques privées de sites d'observation : moins il y a de sites d'observation dans le voisinage proche d'un site s_0, plus l'erreur est élevée.

4.3 Méthodes d'échantillonnage

4.3.1 Echantillonnage aléatoire

La première estimation est réalisée en tirant de manière uniforme 100 points dans la carte construite précédemment. Le semi-variogramme minimisant la MSE parmi l'ensemble testé est alors $\hat{\gamma}(r) = 3.68 * (1 - \exp(-(\frac{r}{1.95})^2))$ avec une erreur test de $MSE_{test} = 0.141$. En appliquant ce semi-variogramme afin de prédire l'ensemble des valeurs de la carte, nous obtenons la carte krigée ci-dessous, avec une MSE totale de 0.043.

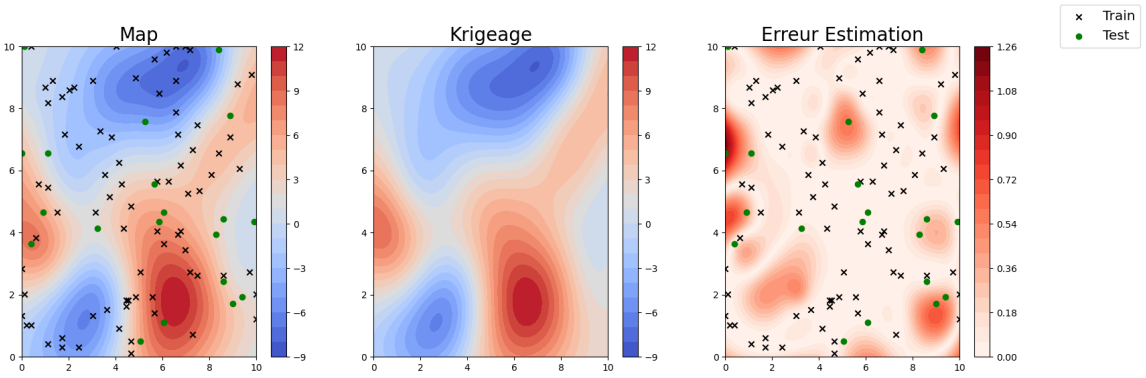


FIGURE 4.12 – Echantillonnage aléatoire : 100 points

On applique le même procédé de sélection aléatoire, cette fois-ci en augmentant le nombre de sites d'observation à 150. On a alors $\hat{\gamma}(r) = 6.84 * (1 - \exp(-(\frac{r}{1.95})^2))$. Les erreurs d'entrainement et totale sont alors : $MSE_{test} = 0.002 \quad MSE_{total} = 0.012$

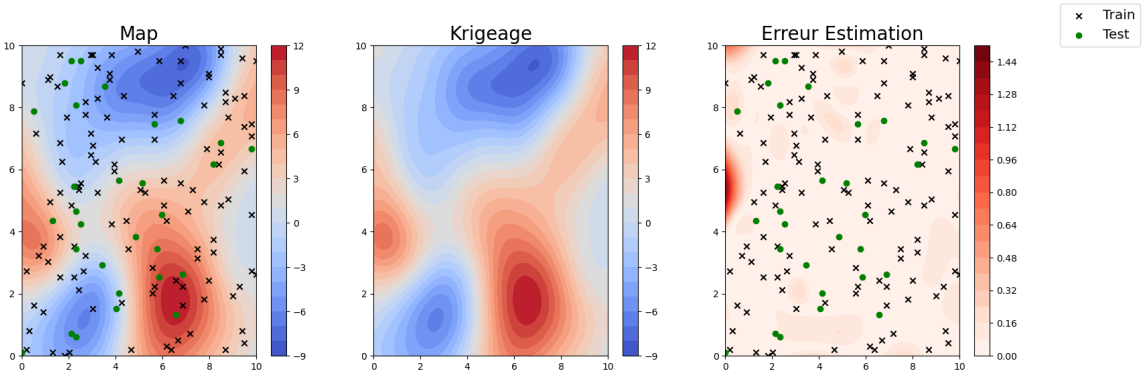


FIGURE 4.13 – Echantillonnage aléatoire : 150 points

Ainsi, l'augmentation du nombre de sites d'observation améliore la qualité de prédiction du krigeage.

4.3.2 Echantillonnage sur grille

La seconde méthode d'échantillonnage utilisée consiste à répartir les sites d'observation sur une grille de taille $m \times m$, garantissant une répartition plus uniforme des sites de mesures

par rapport à la selection aléatoire. Cependant, cette représentation plus régulière des données réduit le nombre de distances sur lequel est estimé le semi-variogramme : les très faibles distances ne sont pas représentées, le modèle sera donc moins apte à détecter les variations à faible échelle. En utilisant une grille de taille 10×10, les résultats sont les suivants :

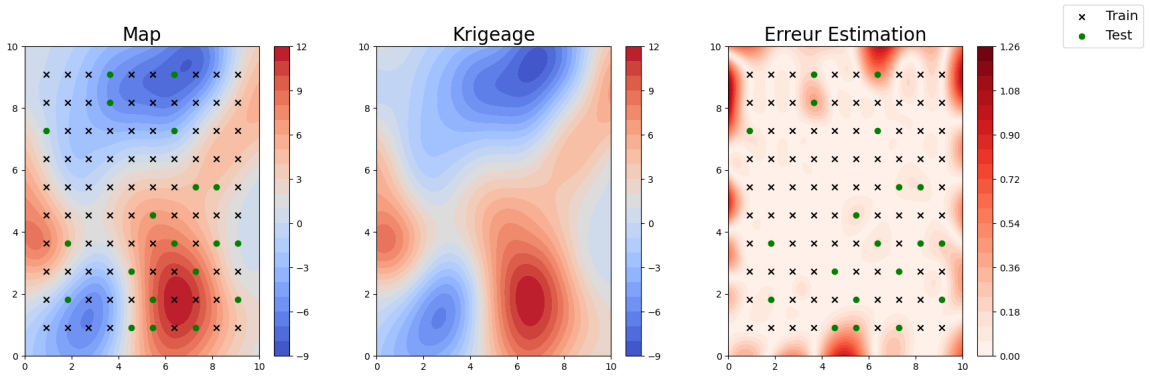


FIGURE 4.14 – Echantillonnage grille 10×10

Le semi-variogramme estimé est $\hat{\gamma}(r) = 2.64 * (1 - \exp(-(\frac{r}{1.95})^2))$ et les MSE de test et totale valent respectivement environ 0.038 et 0.039. La MSE de test est donc plus proche de la MSE totale que dans le cas de la sélection aléatoire, et la MSE totale est légèrement plus faible dans l'échantillonnage sur grille. De plus, cette méthode permet d'être sûr que chaque "source" de température, positive ou négative, soit détectée.

4.3.3 Taille d'échantillonnage

Afin d'étudier l'impact de l'échantillonnage dans l'estimation, on peut également faire varier à la fois le nombre de sites d'observation ainsi que leur répartition. Pour chacun des échantillons créés, on recherche alors le meilleur triplet de paramètres (a, c, c_0) de façon à obtenir le modèle de semi-variogramme gaussien qui minismise la MSE sur la partie test de l'échantillon. Par la suite, on effectue un krigeage de l'ensemble de la carte avec le semi-variogramme optimal obtenu.

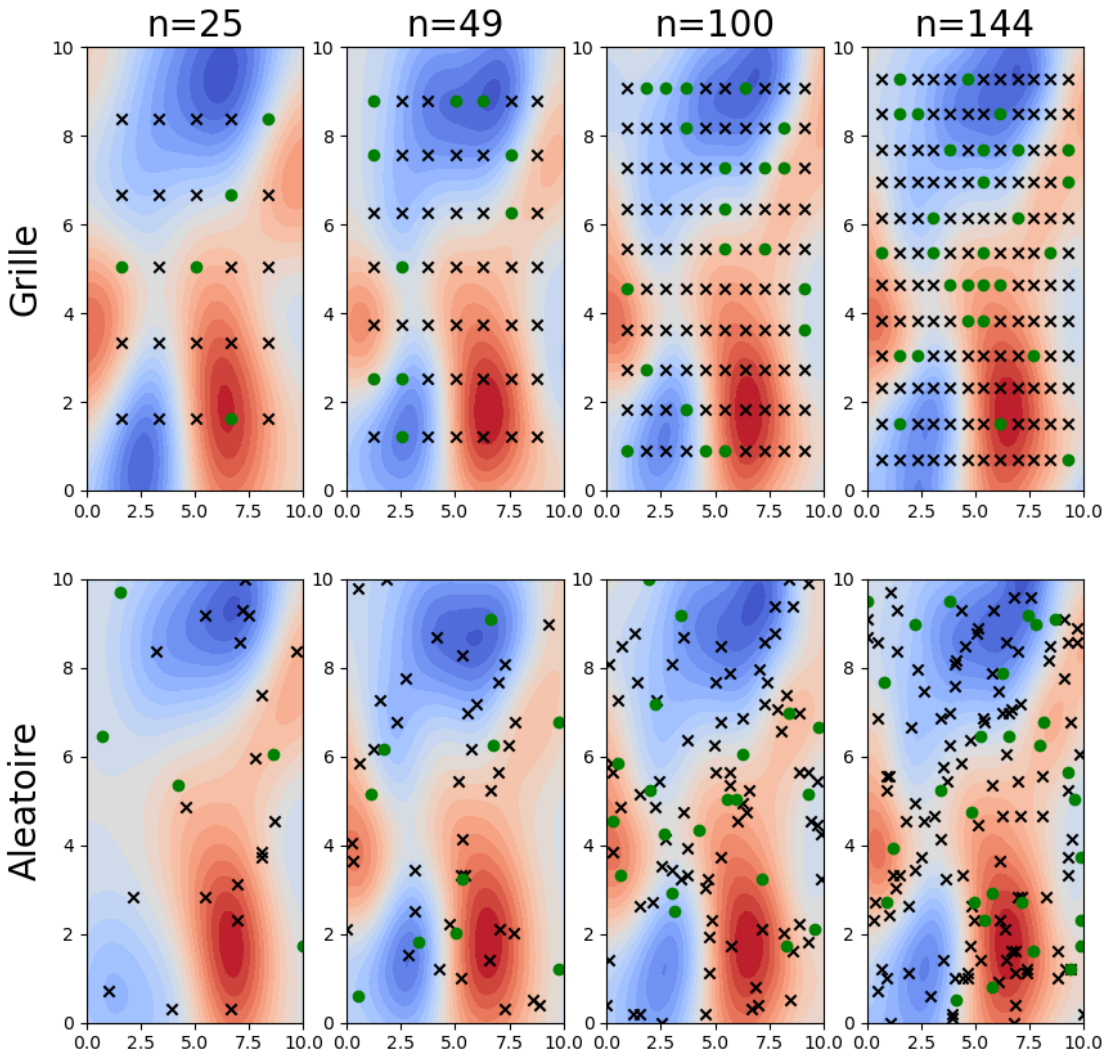


FIGURE 4.15 – Krigeages obtenus en fonction des tailles et méthodes d'échantillonnage

Méthode	nb point	MSE	Median
grille	10	0.048	0.001
grille	12	0.082	0.001
alea	144	0.121	0.004
alea	100	0.56	0.001
grille	7	0.623	0.032
alea	49	0.637	0.123
grille	5	1.19	0.222
alea	25	3.753	0.711

Conclusion

Le krigeage est une méthode d'interpolation spatiale permettant d'estimer la valeur d'un phénomène en chaque point d'un champ grâce à une combinaision linéaire sans biais et de variance des erreurs minimale grâce à un nombre fini de sites, tout en prenant en compte la structure de dépendance spatiale des données. L'estimation de cette structure de dépendance s'effectue au préalable à l'aide d'une analyse variographique sur les sites observés. Le nombre et la répartition de ces points de mesures va donc jouer un rôle essentiel dans la qualité de prévision du krigeage. En effet, un nombre trop faible d'observations nuira fortement au résultat obtenu. En revanche, un nombre plus important de sites entraînera l'augmentation de la précision des résultats mais également une estimation semi-variographique plus coûteuse en terme de temps de calcul. La localisation des sites d'observation joue également un rôle sur le résultat final : les sites d'observation doivent pouvoir capter un maximum de variabilité afin de pouvoir retrouver un maximum de fluctuations du phénomène mesuré.

De plus, l'estimation empirique de la méthode du krigeage étant difficilement utilisable dans la pratique, il est nécessaire de passer par une comparaison de modèles de semi-variogrammes, augmentant alors les temps de calcul. Enfin, il est important de rappeler que le krigeage est une méthode d'interpolation exacte en ces sites d'observation : ainsi, il est fondamental de diviser l'ensemble des sites d'observation en échantillons d'entraînement et de test afin de retrouver le modèle de semi-variogramme adapté.

Ce travail de recherche s'est porté sur le krigeage dans sa version simple et ordinaire. Il existe cependant d'autres modèles plus performants, permettant notamment de traiter des valeurs extrêmes, ou des cas où les hypothèses de stationnarité ne sont pas respectées, comme par exemple le krigeage universel.

Bibliographie

[1] BAILLARGEON Sophie, *Le krigeage : revue de la théorie et application à l'interpolation spatiale des données de précipitations*, 2005

[2] MATHERON, Georges et al. (1965). Les variables régionalisées et leur estimation. Masson et Cie

[3] BOUALLA Nabila, *Cours de géostatistique*, 2018-2019, https ://www.researchgate.net/publication/346083587_Geostatistique

[4] FLOCH Jean-Michel, *Géostatistique*, INSEE, https ://docplayer.fr/196430201-5-geostatistique-jean-michel-floch-insee-resume.html

[5] PROIA Frédéric, *Cours de Séries Chronologiques*, 2021-2022