Modélisation statistique spatiale :
Géostatistique et autres modèles spatiaux ...

Pascal MONESTIEZ
Biostatistique et Processus Spatiaux,
INRA, Avignon

Journées ”Modélisation en Environnement”
22 février au 26 février 2010
Europôle Méditerranéen de l’Arbois

ECCOREV, Fédération de Recherche CNRS 3098
Ecosystèmes Continentaux et Risques Environnementaux

Plan :

- Les types de données spatiales et
  le domaine d'application de la Géostatistique,
  Exploration préalable de données positionnées dans l'espace
- Le cadre théorique, hypothèses et outils nécessaires,
  exemples de simulation de ces modèles théoriques.

Plan :

- Les types de données spatiales et
  le domaine d'application de la Géostatistique,
  Exploration préalable de données positionnées dans l'espace
- Le cadre théorique, hypothèses et outils nécessaires,
  exemples de simulation de ces modèles théoriques.
- Caractériser les variations spatiales : variogramme expérimental,
  modélisation et ajustement (estimation).
- Interpolation (cartographie) optimale par krigeage,

Plan :

- Les types de données spatiales et
  le domaine d'application de la Géostatistique,
  Exploration préalable de données positionnées dans l'espace
- Le cadre théorique, hypothèses et outils nécessaires,
  exemples de simulation de ces modèles théoriques.
- Caractériser les variations spatiales : variogramme expérimental,
  modélisation et ajustement (estimation).
- Interpolation (cartographie) optimale par krigeage,
- Caractériser les relations entre variables spatiales
  (Géostatistique Multivariable)
- Interpoler à partir de plusieurs variables par cokrigeage ou krigeage
  à dérive externe.

Partie 1 :
Les différents types de données spatiales

Ils se distinguent par

- leur support de mesure et leur nature
- les modèles et les méthodes statistiques qui seront utilisés
- les questions que l'on posera au travers de ces données

Partie 1 :
Les différents types de données spatiales

Ils se distinguent par

- leur support de mesure et leur nature
- les modèles et les méthodes statistiques qui seront utilisés
- les questions que l'on posera au travers de ces données

Grossièrement, on peut les regrouper en trois grand types :

A : Données définies en tout point de l'espace (support continu) et mesurées sur un nombre fini de sites échantillonnés

- Exemples : la température, les précipitations, les propriétés des sols, les caractéristiques géologiques, et à la limite, une concentration ou une densité de quelque chose ou un comptage d'animaux.

A : Données définies en tout point de l'espace (support continu) et mesurées sur un nombre fini de sites échantillonnés

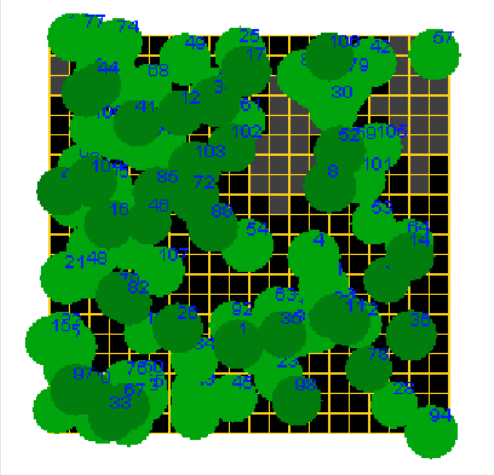
- Exemples : la température, les précipitations, les propriétés des sols, les caractéristiques géologiques, et à la limite, une concentration ou une densité de quelque chose ou un comptage d'animaux.
- Questions typiques :
  - caractériser la variabilité spatiale en fonction de la distance (géographique) entre deux lieux
  - interpoler (cartographier) la variable entre les points mesurés
  - simuler des variations spatiales du même type
  - évaluer l'erreur d'interpolation et la qualité de l'échantillonnage

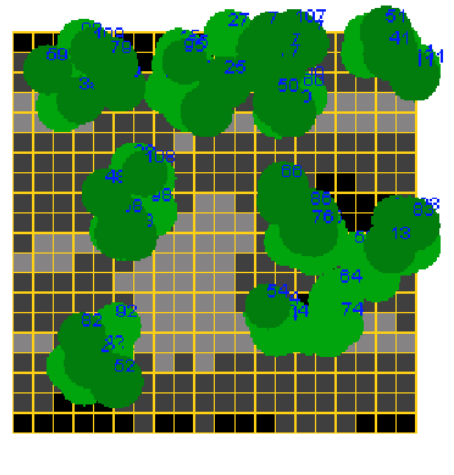
A : Données définies en tout point de l'espace (support continu) et mesurées sur un nombre fini de sites échantillonnés

- Exemples : la température, les précipitations, les propriétés des sols, les caractéristiques géologiques, et à la limite, une concentration ou une densité de quelque chose ou un comptage d'animaux.
- Questions typiques :
  - caractériser la variabilité spatiale en fonction de la distance (géographique) entre deux lieux
  - interpoler (cartographier) la variable entre les points mesurés
  - simuler des variations spatiales du même type
  - évaluer l'erreur d'interpolation et la qualité de l'échantillonnage
- C'est le domaine de la géostatistique

B : points ou objets positionnés aléatoirement dans l'espace et observés dans de petits domaines spatiaux ou des transects.

- Exemples : les arbres d'une forêt, la répartition d'espèces végétales, d'animaux. Chaque individus ou objet peut être caractérisé par des variables





B : points ou objets positionnés aléatoirement dans l'espace et observés dans de petits domaines spatiaux ou des transects.

- Questions typiques :
  - caractériser la distribution spatiale des objets : indépendance, régularités, agrégation ?
  - expliquer la distribution des caractéristiques des objets en fonction de leurs positionnements relatifs.
  - simuler des distributions spatiales du même type

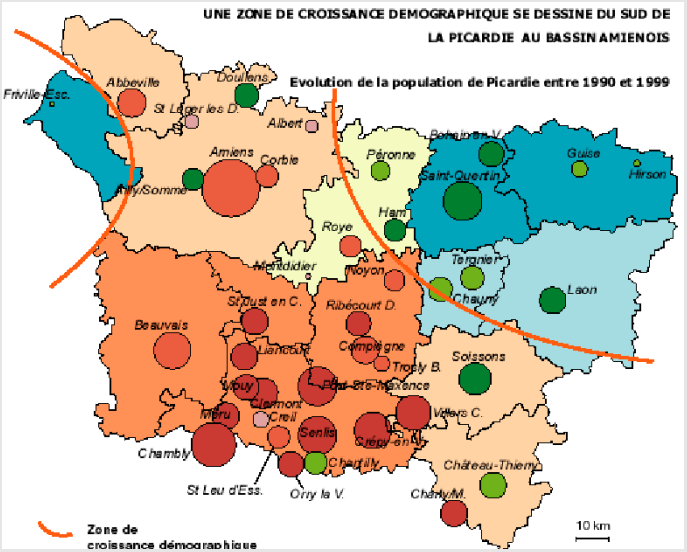
B : points ou objets positionnés aléatoirement dans l'espace et observés dans de petits domaines spatiaux ou des transects.

- Questions typiques :
  - caractériser la distribution spatiale des objets : indépendance, régularités, agrégation ?
  - expliquer la distribution des caractéristiques des objets en fonction de leurs positionnements relatifs.
  - simuler des distributions spatiales du même type

- C'est le domaine des Processus Ponctuels
  (ou processus ponctuels marqués).

C : données sur ”lattices” ou sur réseaux.

- Exemples : Données de populations, données épidémiologiques sur un ensemble d’entités administratives (communes, départements). Caractéristiques de villes reliées par un réseau de transport. Données parcellaires dans un paysage (le découpage foncier est donné)



C : données sur ”lattices” ou sur réseaux.

- Questions typiques :
  - caractériser la variabilité spatiale : indépendance entre voisins, régularités, agrégation ?
  - expliquer la distribution des caractéristiques en fonction des distributions dans un voisinage.
  - simuler des distributions spatiales du même type

C : données sur ”lattices” ou sur réseaux.

- Questions typiques :
  - caractériser la variabilité spatiale : indépendance entre voisins, régularités, agrégation ?
  - expliquer la distribution des caractéristiques en fonction des distributions dans un voisinage.
  - simuler des distributions spatiales du même type
- C’est un ensemble assez vaste de modèles incluant les processus sur réseau, les champs de Markov, mais aussi toutes les méthodes d’analyse d’image (cas particulier de pixels tous identiques selon une grille régulière).

Visualiser des données spatiales du premier type

En phase exploratoire, il est essentiel de

- ne pas faire d'hypothèse sur la régularité ou la continuité du phénomène
- visualiser les données brutes sans transformation.
- pouvoir se faire un première idée la variabilité spatiale avant de proposer une modélisation, d'utiliser la géostatistique ou toute autre méthode plus complexe.

Un exemple de données spatiales :

Le satellite SEAWIFS est en panne pour longtemps, j'ai besoin des données (concentration en chlorophyle a) pour la zone NW de la Méditerranée.

Un exemple de données spatiales :

Le satellite SEAWIFS est en panne pour longtemps, j'ai besoin des données (concentration en chlorophyle a) pour la zone NW de la Méditerranée.

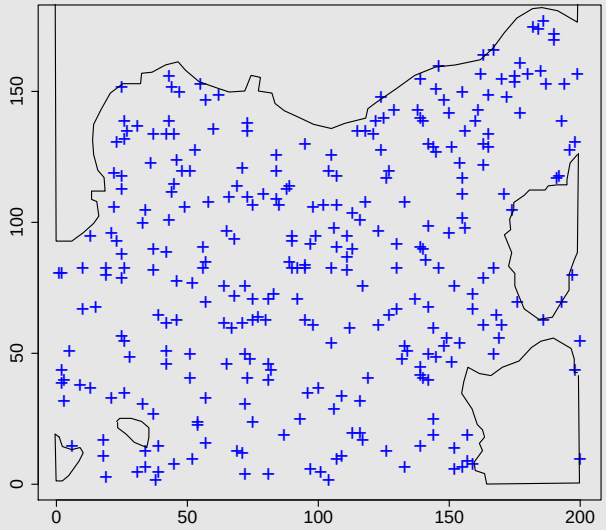
300 bateaux se proposent de rapporter chacun un échantillon

Un exemple de données spatiales :

Le satellite SEAWIFS est en panne pour longtemps, j'ai besoin des données (concentration en chlorophyle a) pour la zone NW de la Méditerranée.

300 bateaux se proposent de rapporter chacun un échantillon

Voici la localisation des données :

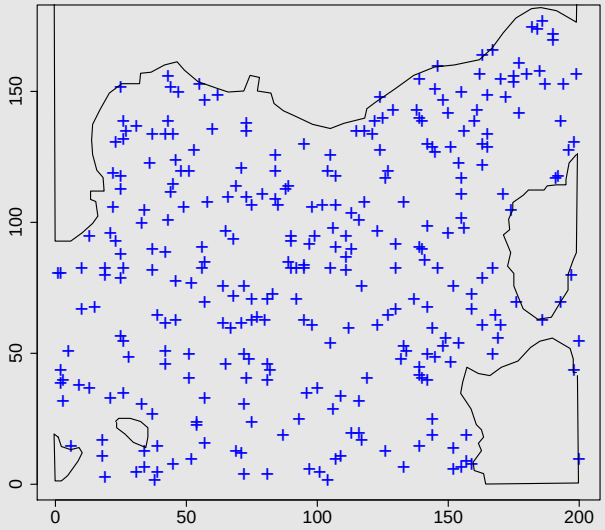


Un exemple de données spatiales :

Le satellite SEAWIFS est en panne pour longtemps, j'ai besoin des données (concentration en chlorophyle a) pour la zone NW de la Méditerranée.

300 bateaux se proposent de rapporter chacun un échantillon

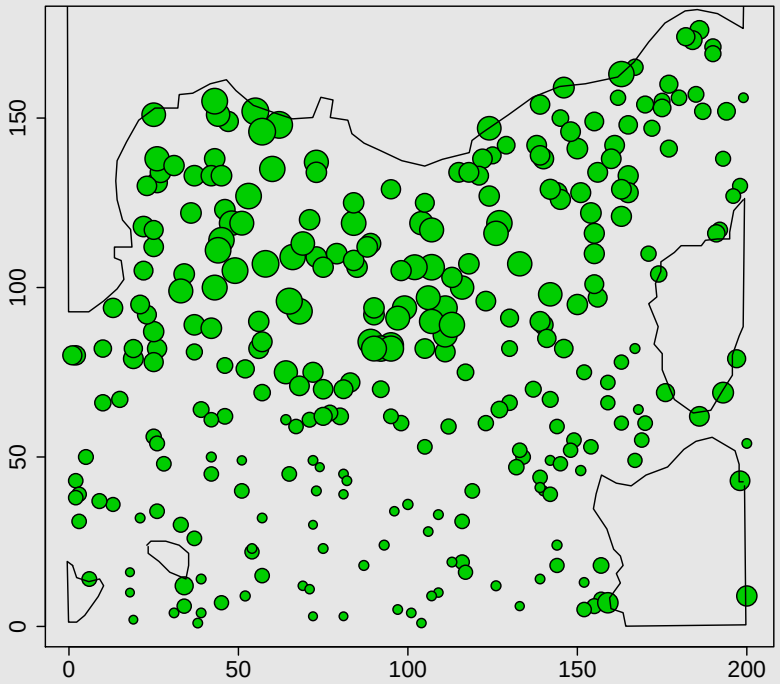
Voici la localisation des données :



En plus, ils sont spatialement "plutôt bien" répartis

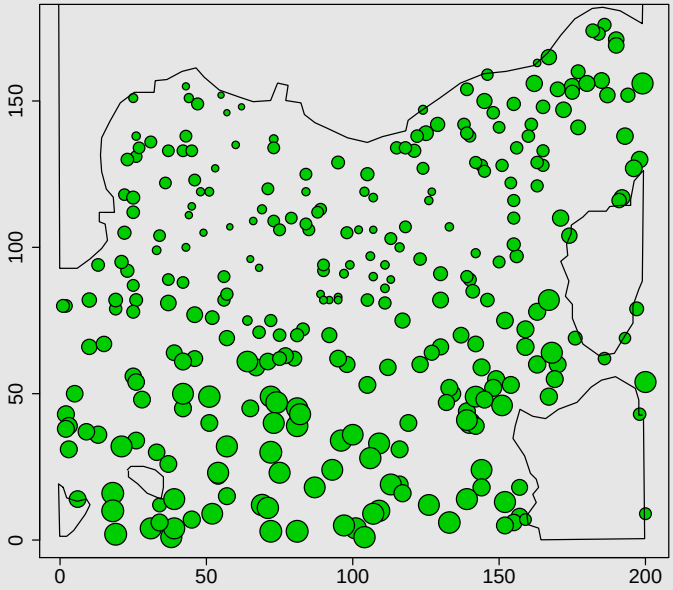
Visualiser les données :

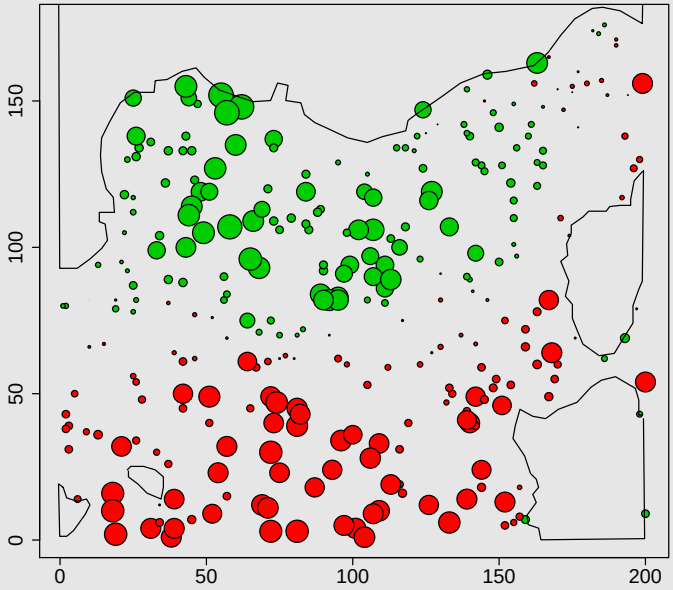
on peut représenter la variable mesurée avec des symboles proportionnels



Visualiser les données :

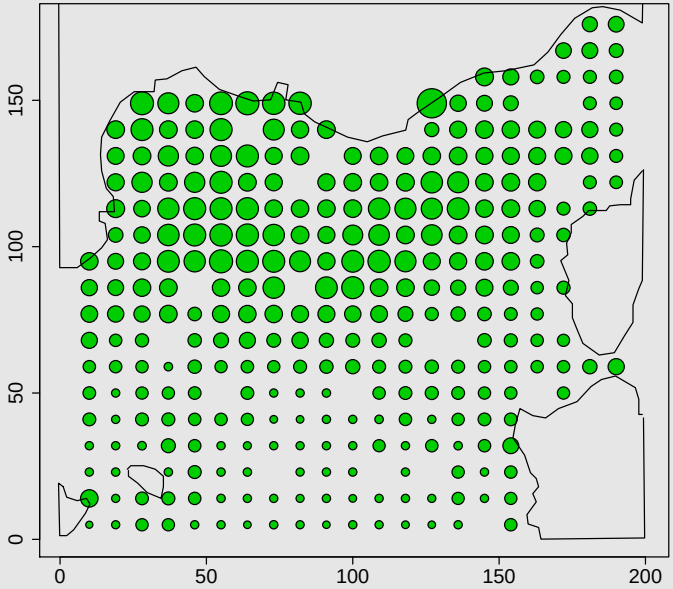
on peut aussi inverser les valeurs, ou centrer la variable

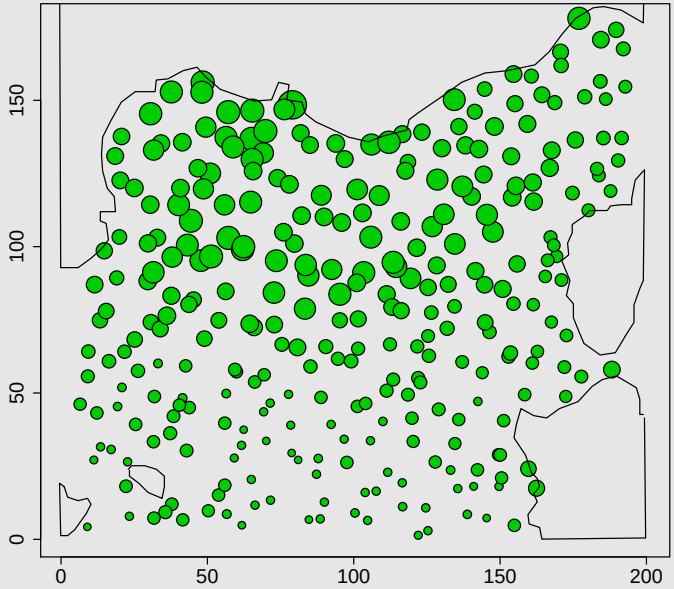




Visualiser les données :

Sans compter l'effet visuel de l'échantillonnage





Fin de la première partie

Partie 2 : Modèles et hypothèses

Pour atteindre les objectifs

- décrire la variabilité spatiale
- interpoler (cartographier) le mieux possible
- évaluer l'erreur d'interpolation

Partie 2 : Modèles et hypothèses

Pour atteindre les objectifs

- décrire la variabilité spatiale
- interpoler (cartographier) le mieux possible
- évaluer l'erreur d'interpolation

Il faudra un modèle :

- un champ aléatoire $Z(s)$ où $s \in \mathcal{D}$ est un point du domaine d'étude

et quelques hypothèses du fait de la réalisation unique:

- stationnarité d'ordre 2, ergodicité

Hypothèses de la Géostatistique

1. Stationnarité

”La loi est invariante par translation”

Mais dans la plupart des cas, il est suffisant de supposer la stationnarité d’ordre 2

Hypothèses de la Géostatistique

1. Stationnarité

”La loi est invariante par translation”

Mais dans la plupart des cas, il est suffisant de supposer
la stationnarité d’ordre 2

2. Ergodicité

”on peut inférer les paramètres (moments) de la loi spatiale à partir d’une réalisation unique si le domaine d’étude est suffisamment grand”

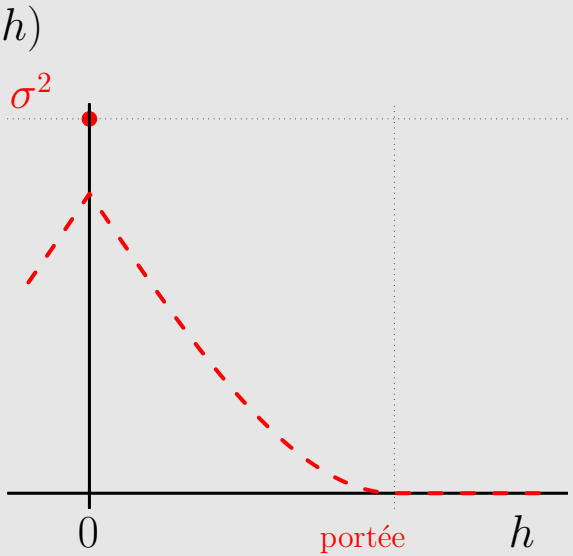
Stationnarité d'ordre 2

"Les deux premiers moments de $Z(x)$ existent et sont invariants par translation"

$$\begin{cases} \mathbf{E}\big[Z(x)\big] = m & \forall x \\ \mathrm{Cov}\big(Z(x), Z(x+h)\big) = C(h) & \forall x, \ \forall h \end{cases}$$

$C(h)$ est la
fonction de covariance

- $C(-h) = C(h)$
  fonction paire
- $C(0) = \mathrm{var}(Z(x)) = \sigma^2$
- peut être négatif
  $|C(h)| \leq C(0)$



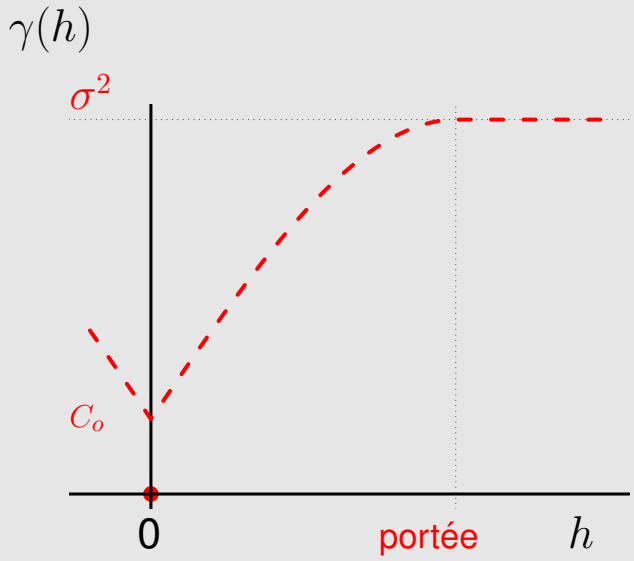
- Il faut que $\forall n, \ \forall (x_1, x_2, \ldots, x_n), \ \forall (\lambda_1, \lambda_2, \ldots, \lambda_n)$

$$\sum_{\alpha} \sum_{\beta} \lambda_\alpha \lambda_\beta \, C(h_{\alpha\beta}) \ \geq 0 \ \Rightarrow \ C(h) \text{ est semi-définie positive}$$

On peut remplacer la Covariance par le variogramme

Par définition :

$$\gamma(h) = \frac{1}{2}\,\mathbf{E}\Big[\big(Z(s+h) - Z(s)\big)^2\Big]$$



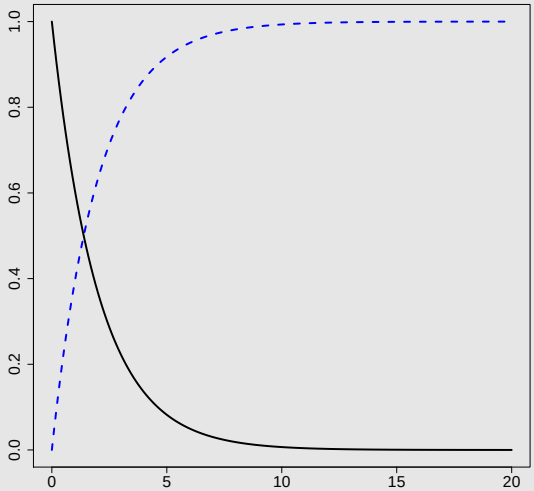
- $\gamma(h)$ La variance de la différence entre 2 observations distantes de h
- $\gamma(h)$ positif, $\gamma(0) = 0$
- $\gamma(-h) = \gamma(h)$ fonction paire
- $C_o = \lim_{h\to 0} \gamma(h)$ "effet de pépite"
- Si hypothèse stationnaire d'ordre 2

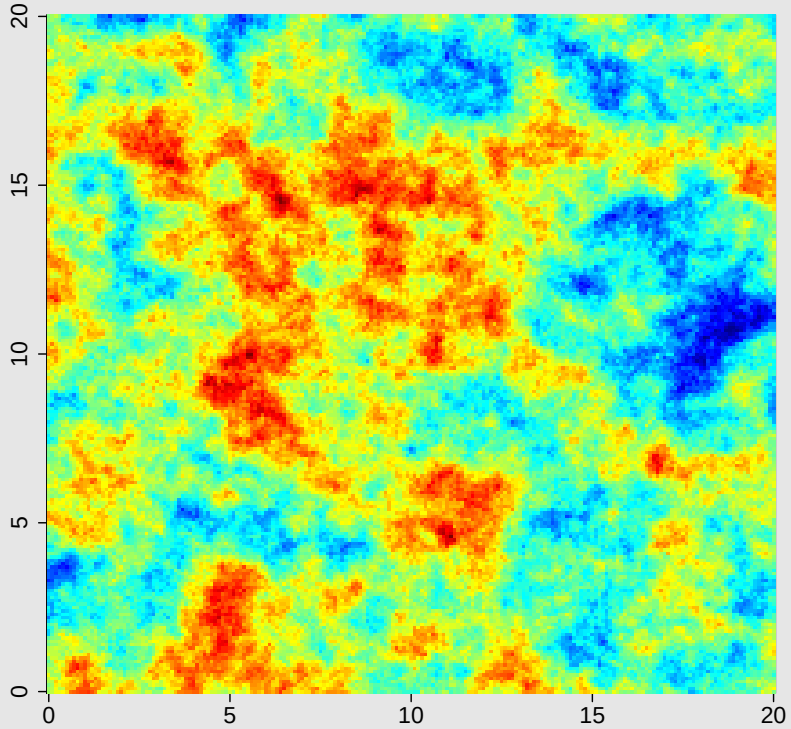
$$\gamma(h) = = C(0) - C(h) = \sigma^2 - C(h)$$

où $C(h)$ est la fonction de covariance spatiale

Exemples de champs aléatoires :

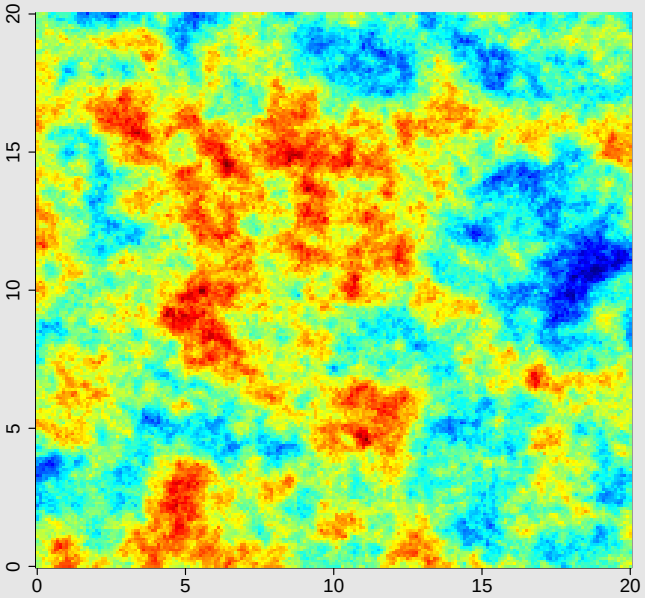
modèle de covariance exponentiel (multigaussien)

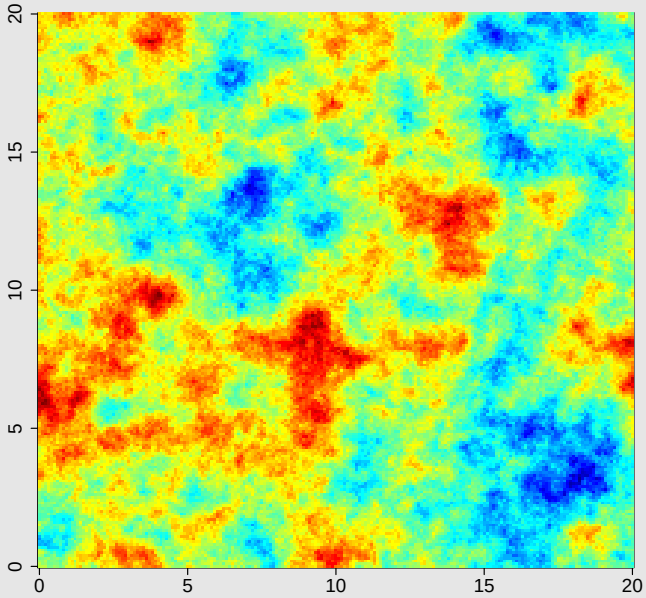




Exemples de champs aléatoires :

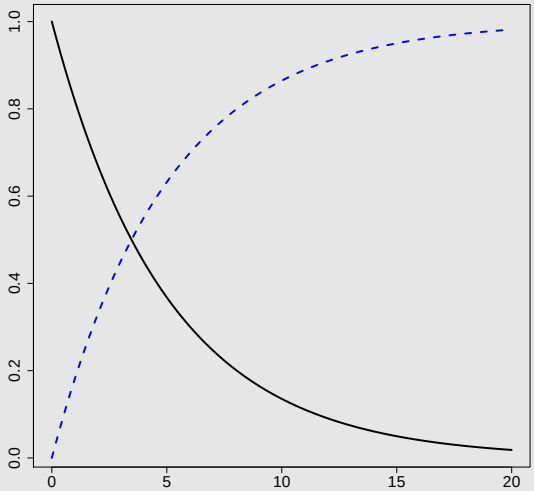
deux simulations avec les mêmes paramètres

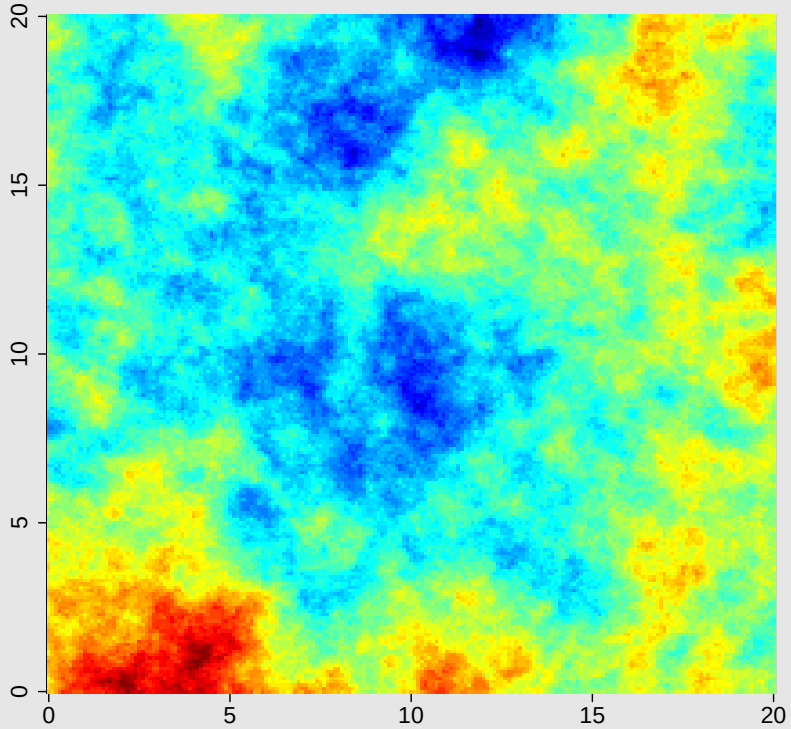




Exemples de champs aléatoires :

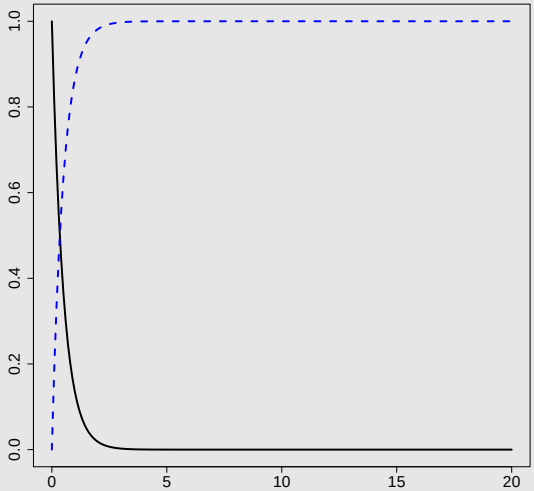
modèle de covariance exponentiel avec une portée plus grande

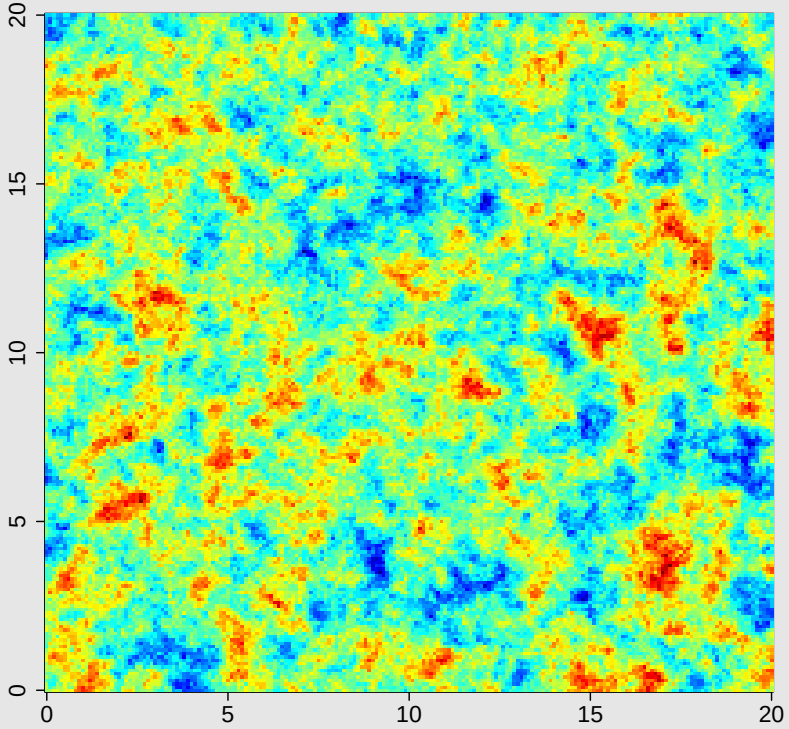




Exemples de champs aléatoires :

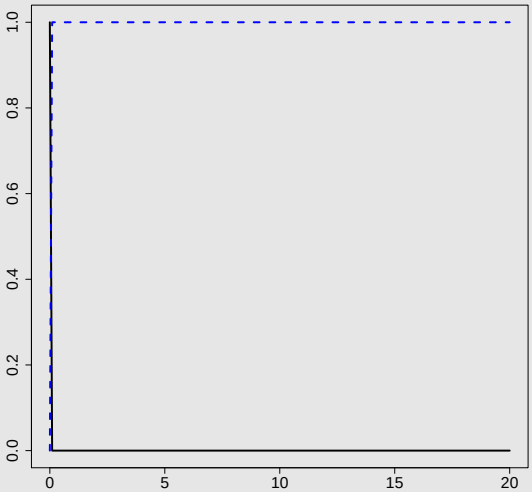
modèle de covariance exponentiel avec une portée plus petite

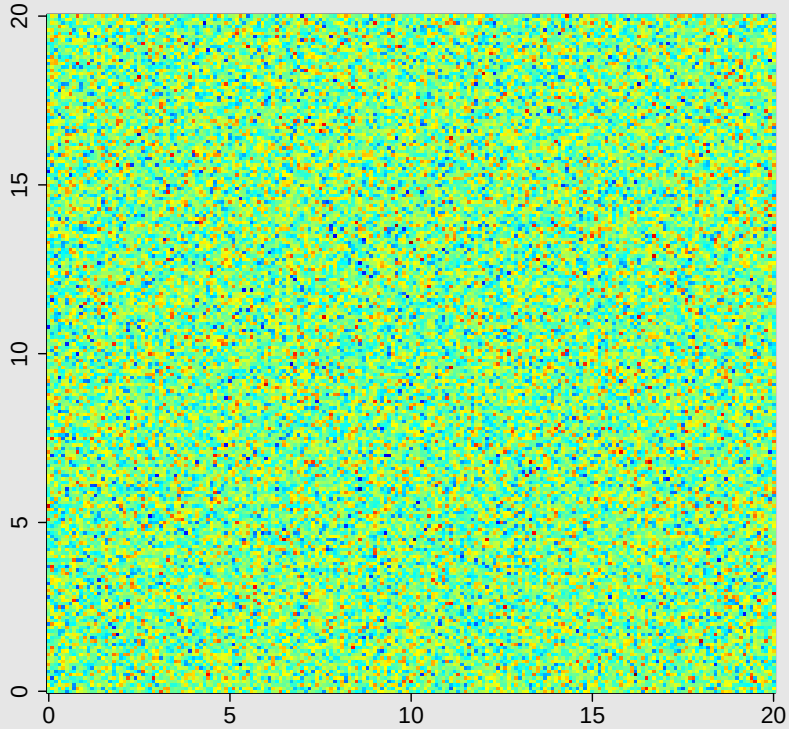




Exemples de champs aléatoires :

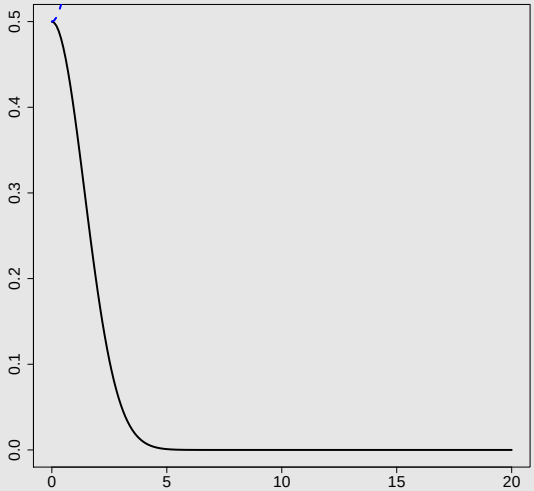
effet aléatoire pur (bruit blanc spatial)

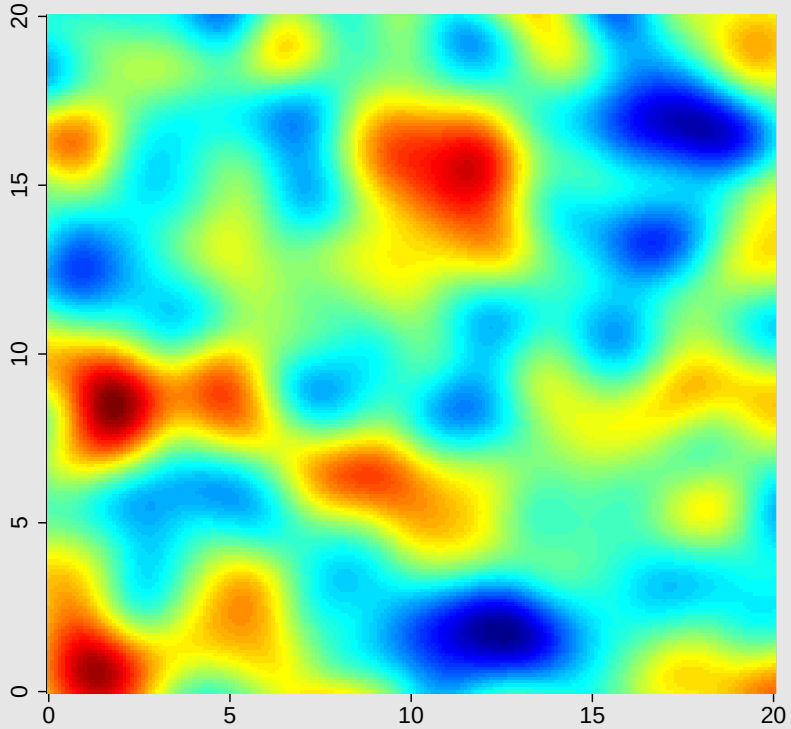




Exemples de champs aléatoires :

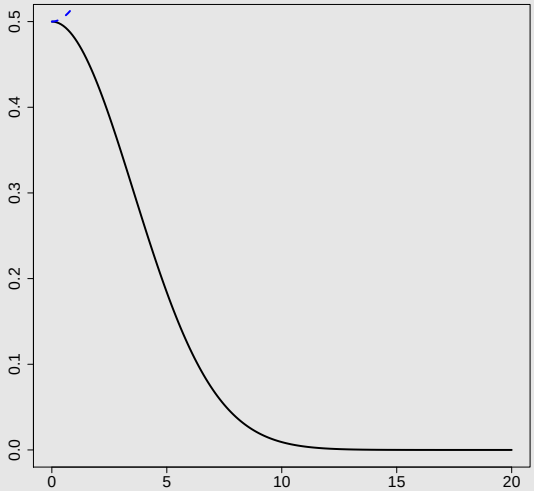
modèle de covariance gaussienne (plus lisse)

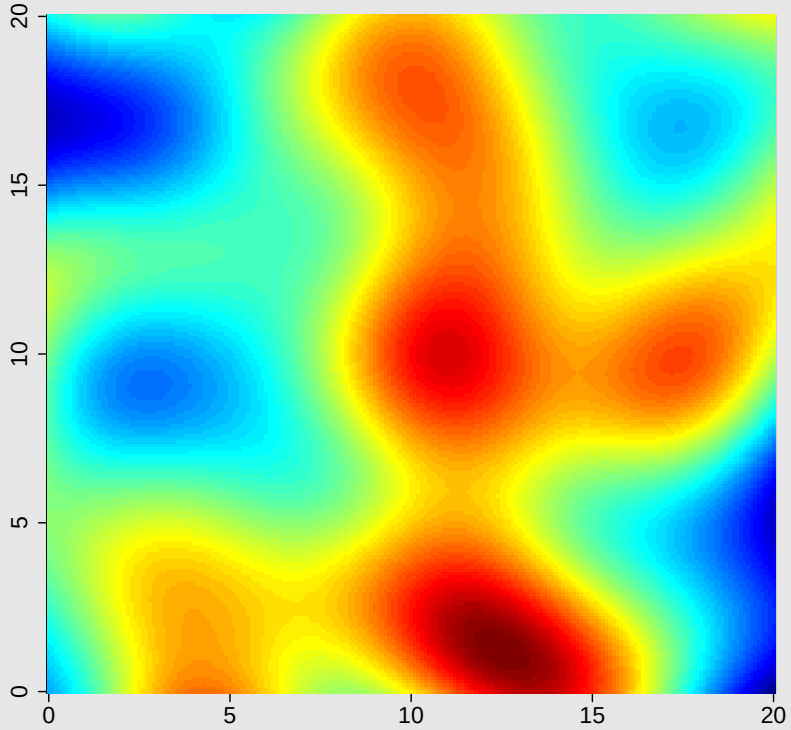




Exemples de champs aléatoires :

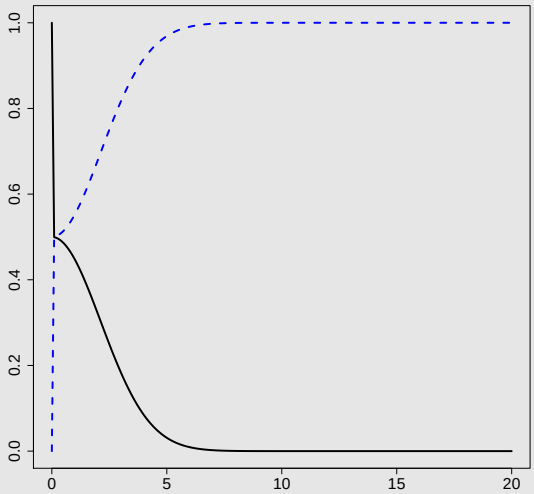
modèle de covariance gaussienne avec un portée plus longue

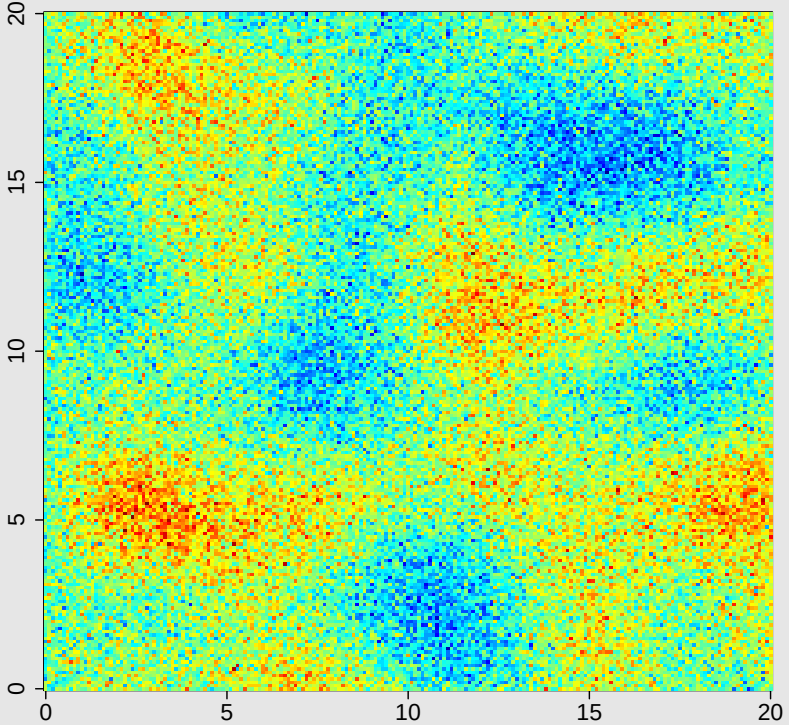




Exemples de champs aléatoires :

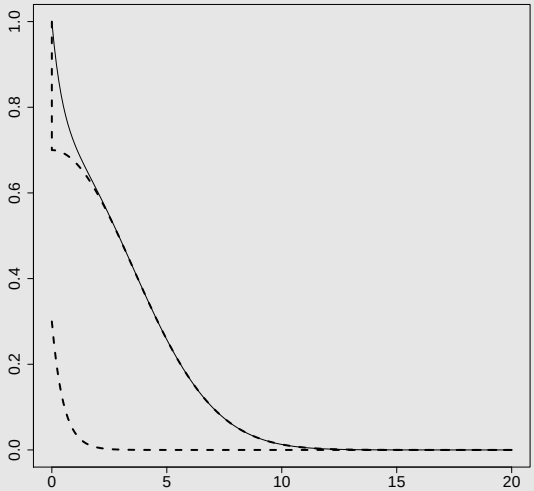
modèle de covariance gaussienne plus un effet aléaoire pur (50%, 50%)

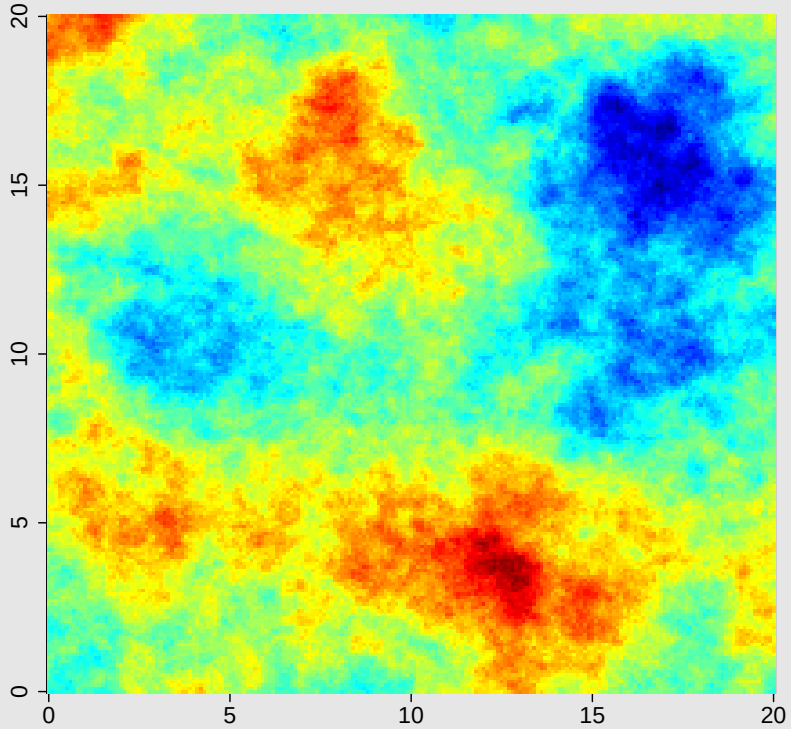




Exemples de champs aléatoires :

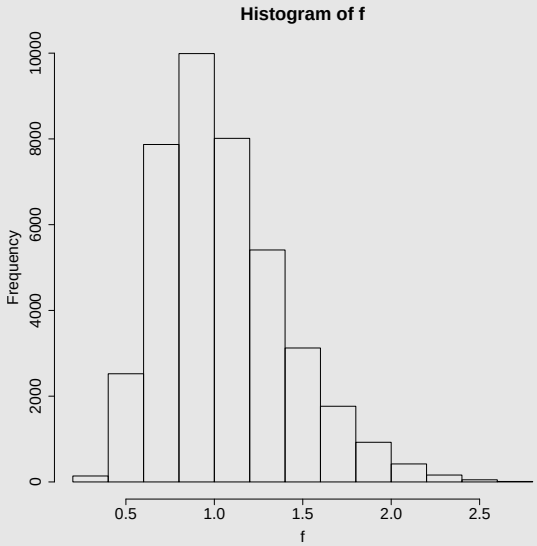
mélange de 2 modèles

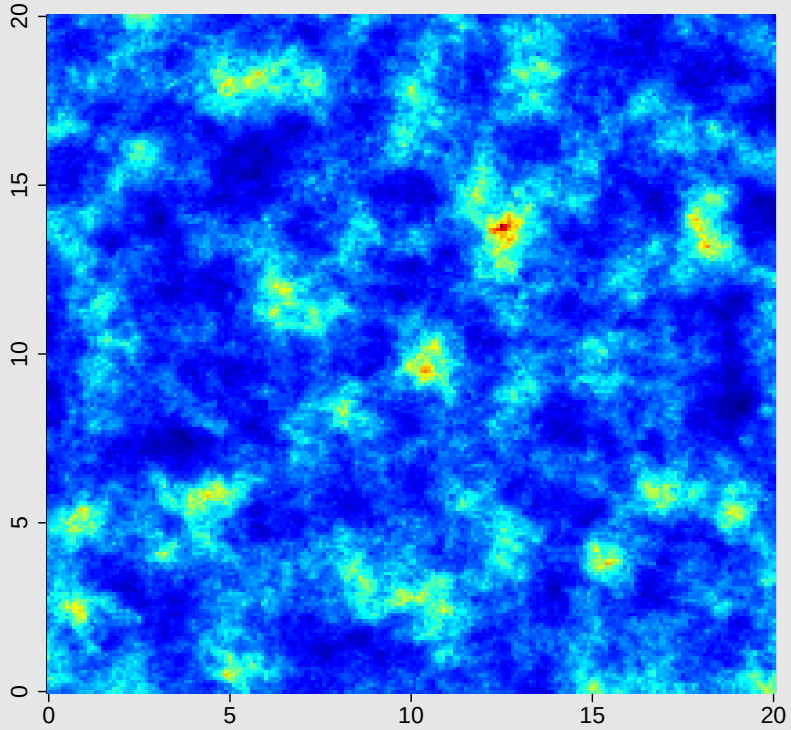




Exemples de champs aléatoires :

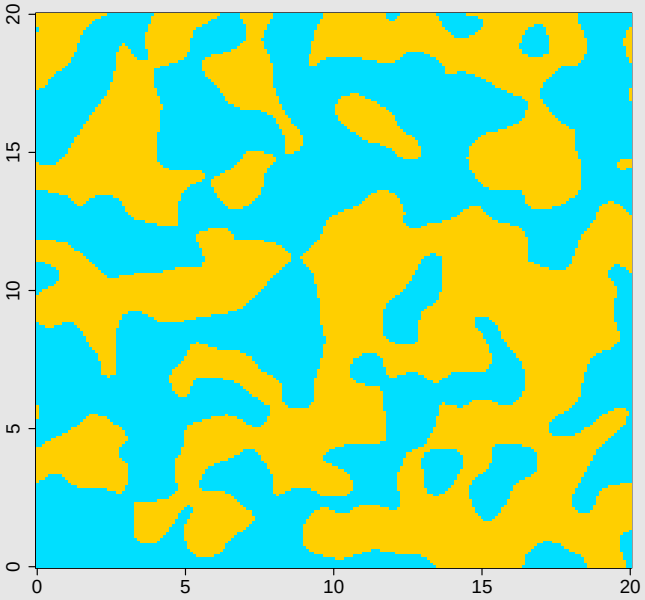
modèle de covariance sphérique et distribution lognormale

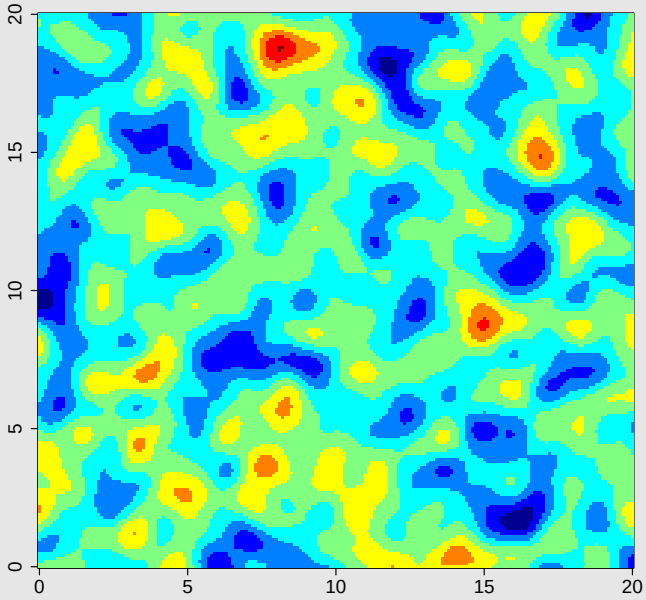




Exemples de champs aléatoires :

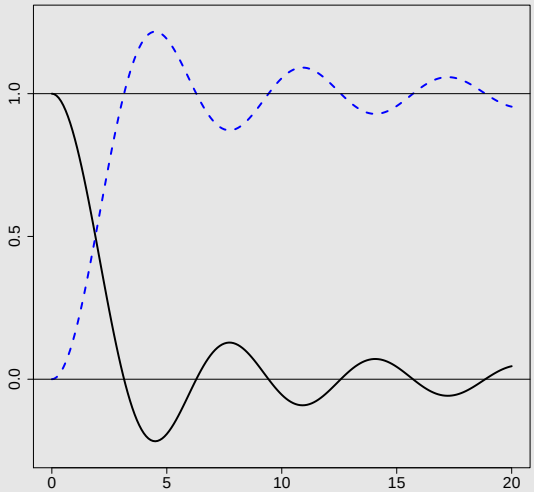
modèle de covariance gaussien et distribution seuillée

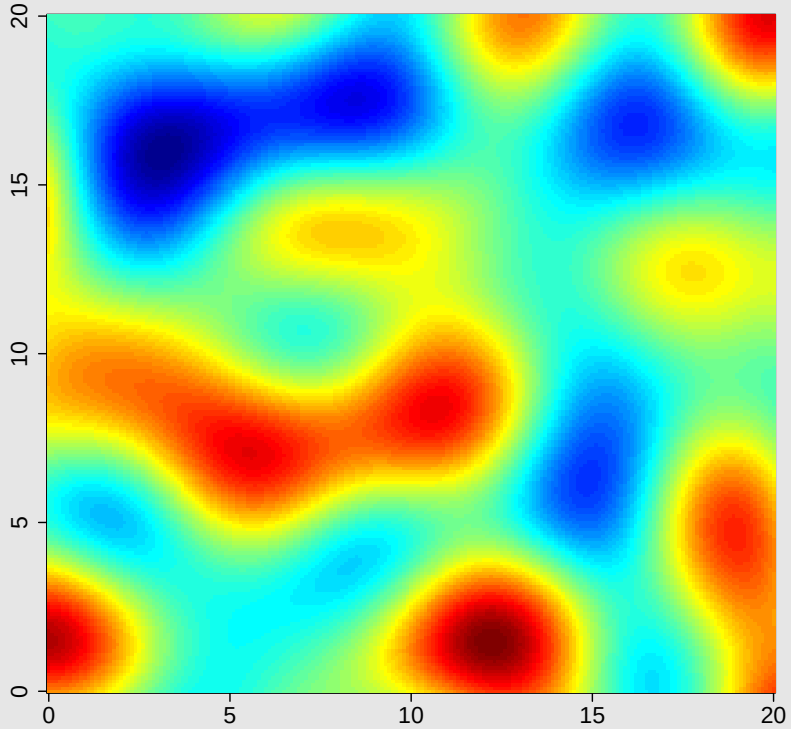




Exemples de champs aléatoires :

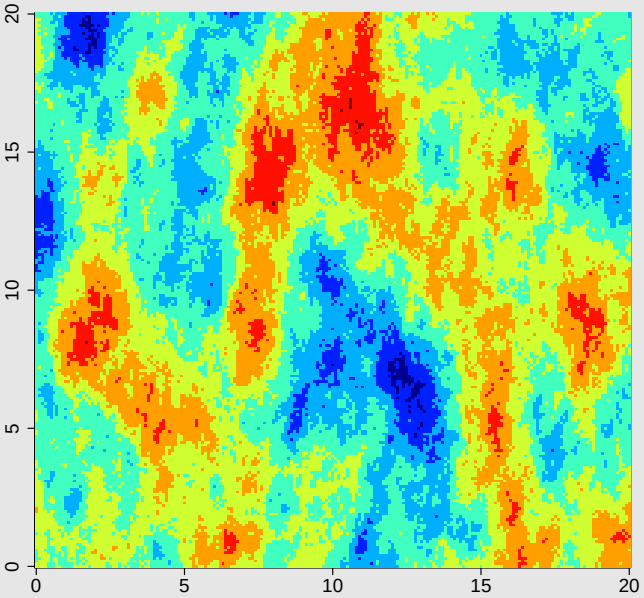
modèle de covariance plus complexes (Bessel avec effets de trous)

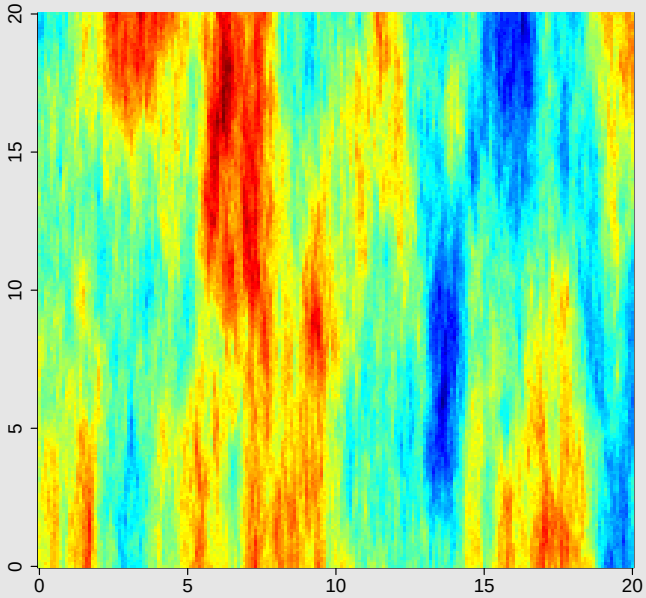




Exemples de champs aléatoires :

modèle anisotropes

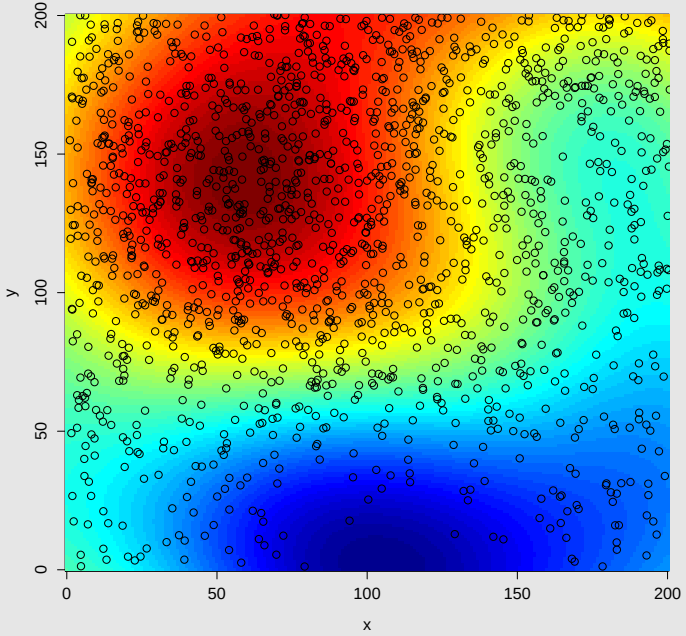


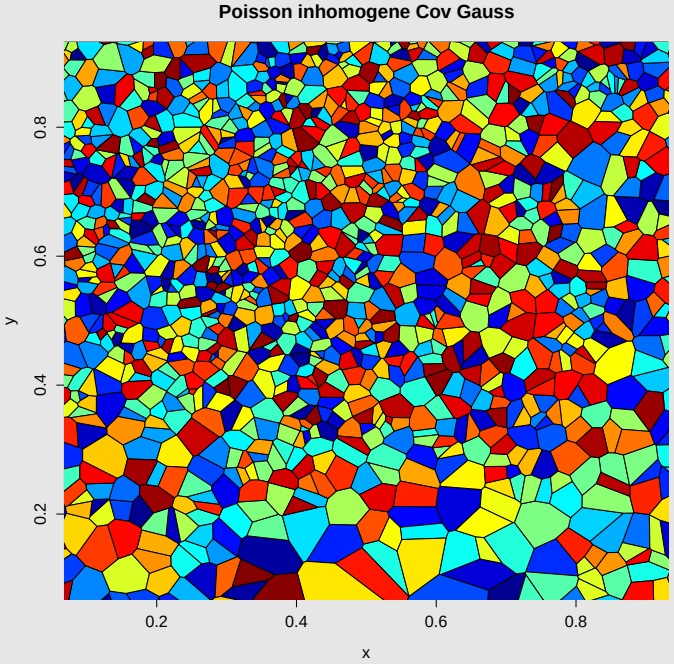


Plus compliqué avec des mosaïques

Exemples de champs aléatoires :

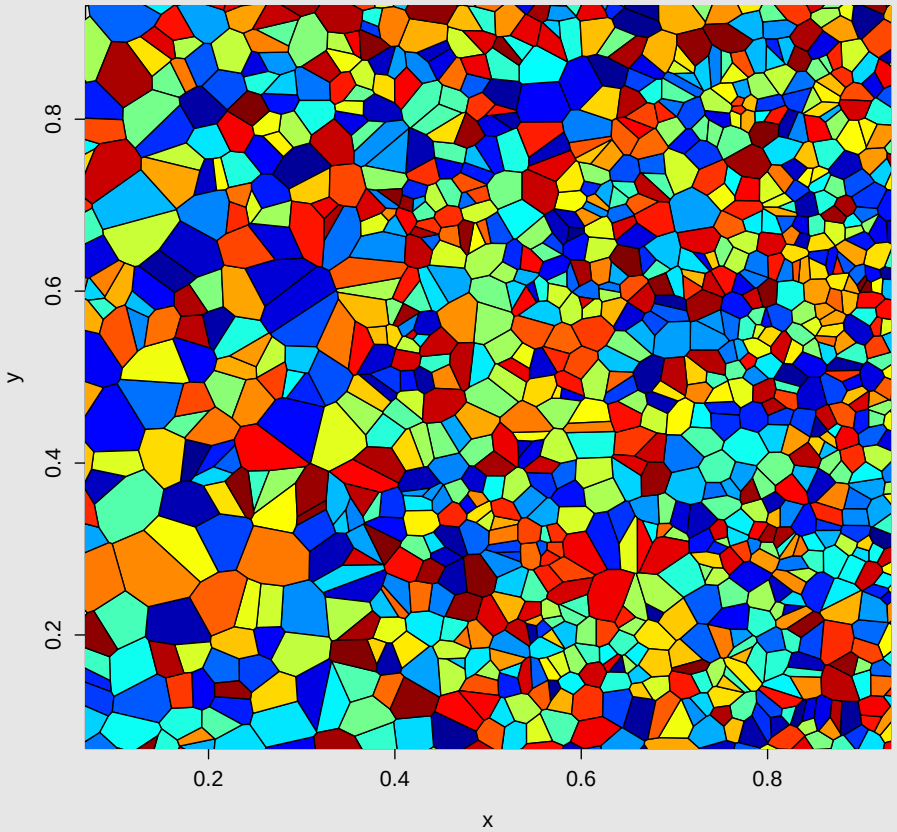
Voronoi sur Poisson inhomogène





Exemples de champs aléatoires :

Voronoi sur Poisson inhomogène

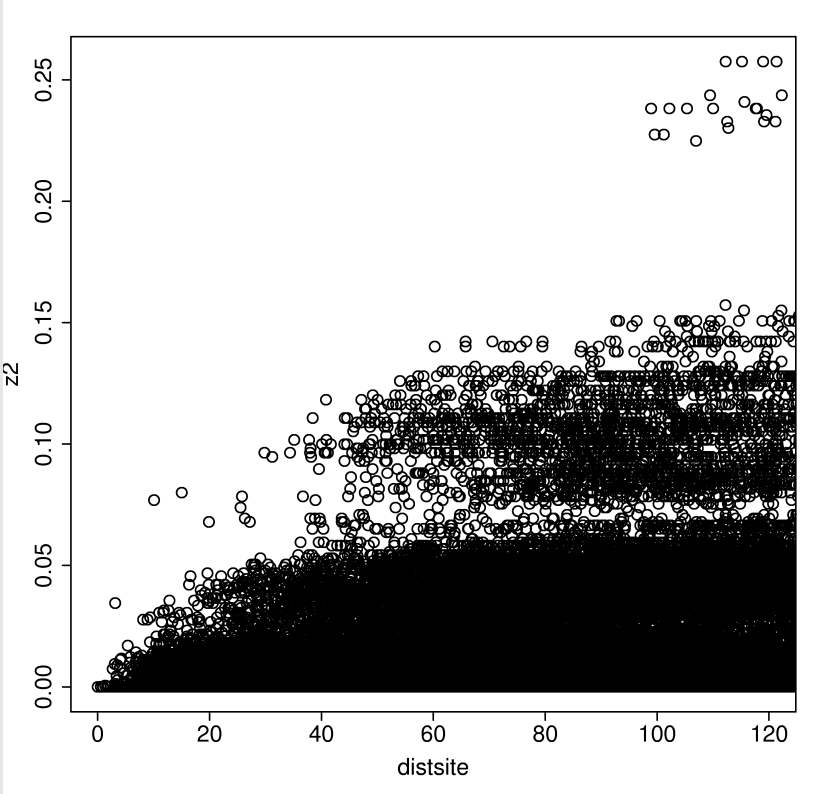


Fin de la seconde partie

Partie 3 : étude variographique

Une première approche est de calculer la nuée variographique :

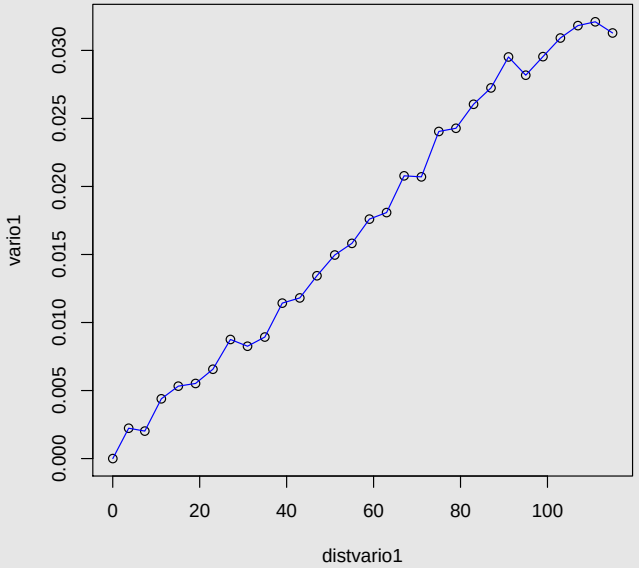
On trace les $\frac{1}{2}\big(Z(s_\alpha) - Z(s_\beta)\big)^2$ en fonction des distances $d(s_\alpha, s_\beta)$:



On calcule le variogramme expérimental

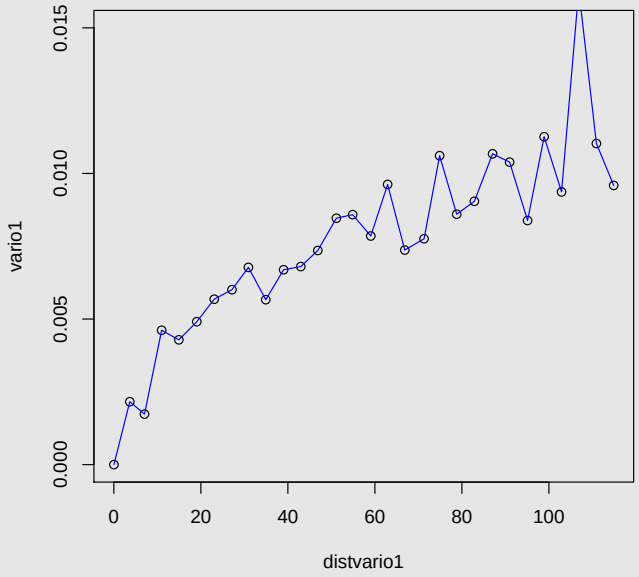
$$\gamma^*(h_k) = \frac{1}{2n_c} \sum_{\alpha,\beta} \big(Z(s_\alpha) - Z(s_\beta)\big)^2 \ \mathbf{1}_{d_{\alpha\beta} \simeq h_k}$$

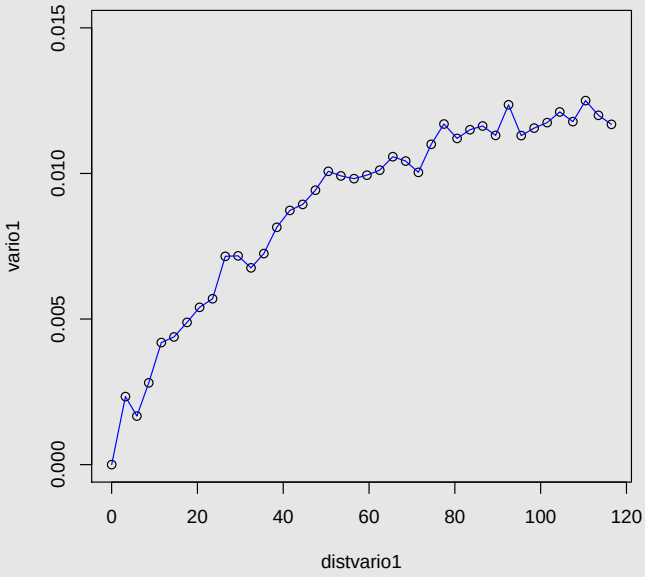
où n_c est le nombre de couples distants de $\simeq h_k$ ($n_c = \sum_{\alpha,\beta} \mathbf{1}_{d_{\alpha\beta} \simeq h_k}$)



Le variogramme expérimental

Du fait du gradient N-S, le variogramme expérimentale est fortement perturbé, on propose deux approches : (1) ne garder que les couples orientés E-W (2) travailler sur les residus

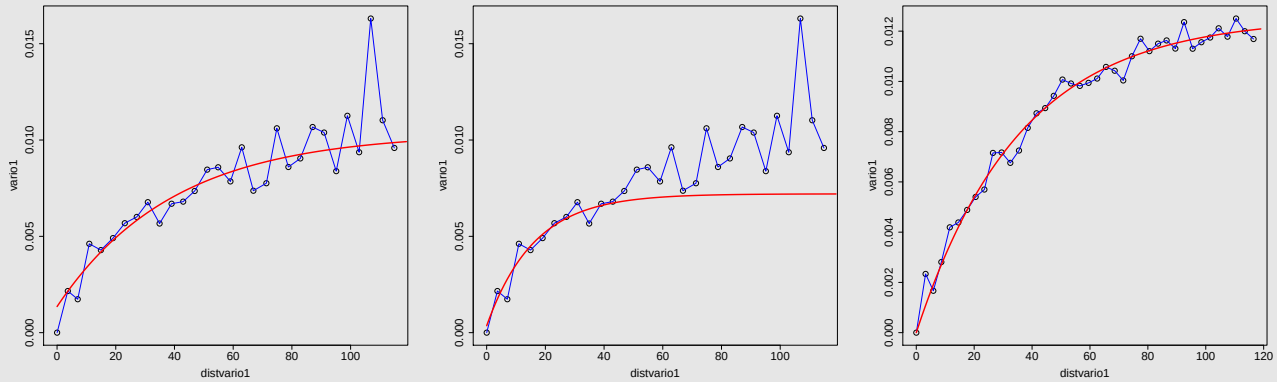




Ajustement des variogrammes expérimentaux

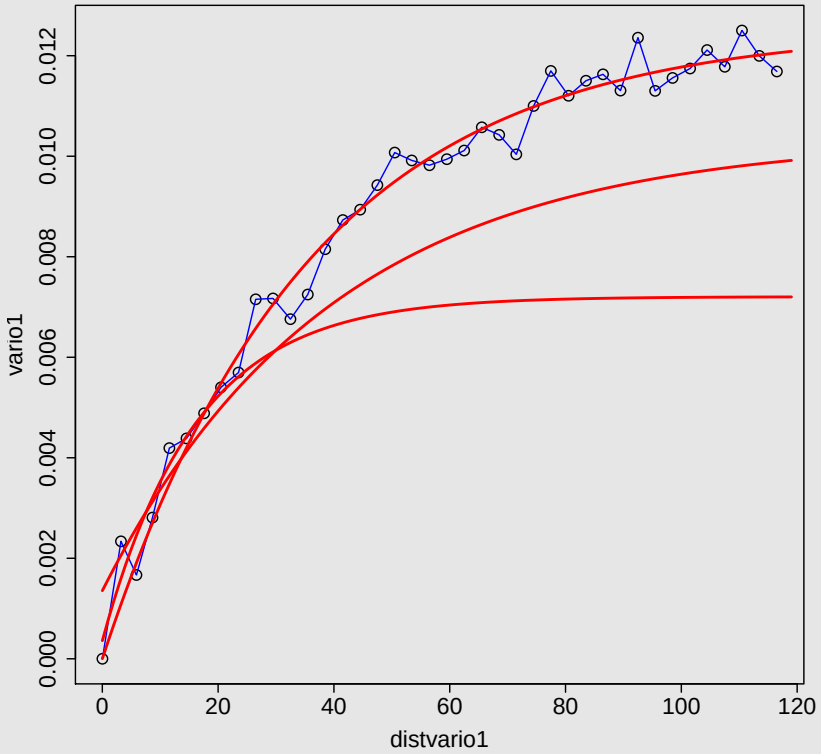
$\gamma_{\hat{\theta}}(h)$ modèle paramétrique ajusté (moindres carrés pondérés) au variogramme expérimental $h_k \longrightarrow \gamma^*(h_k)$

$\gamma_{\hat{\theta}}(h)$ assure la définie-positivité de la fonction de covariance associée et permettra de l'utiliser dans la méthode d'interpolation (Krigeage).



Ajustements pas si différents

Graphique des trois modèles paramétriques ajustés



Autres exemples des possibilités :

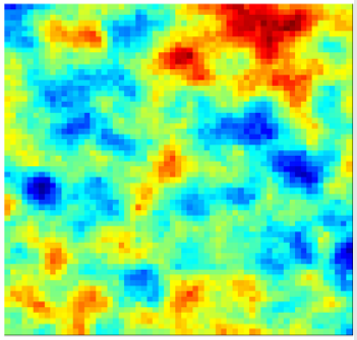
- Multi échelles (sens patchs hiérarchiques)
- Modèle hiérarchique (sens modèles Bayesiens)

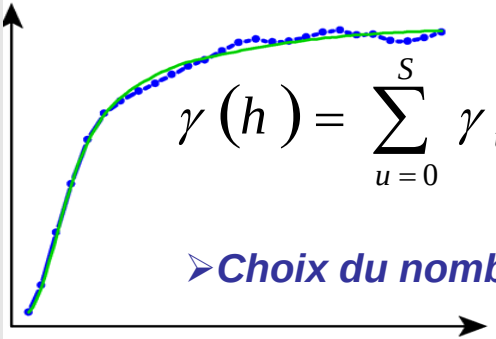
(E. Bellier 2007)

❑ Les modèles de régionalisation

Le phénomène observé est la somme de sous-phénomènes indépendants agissant à différentes échelles.

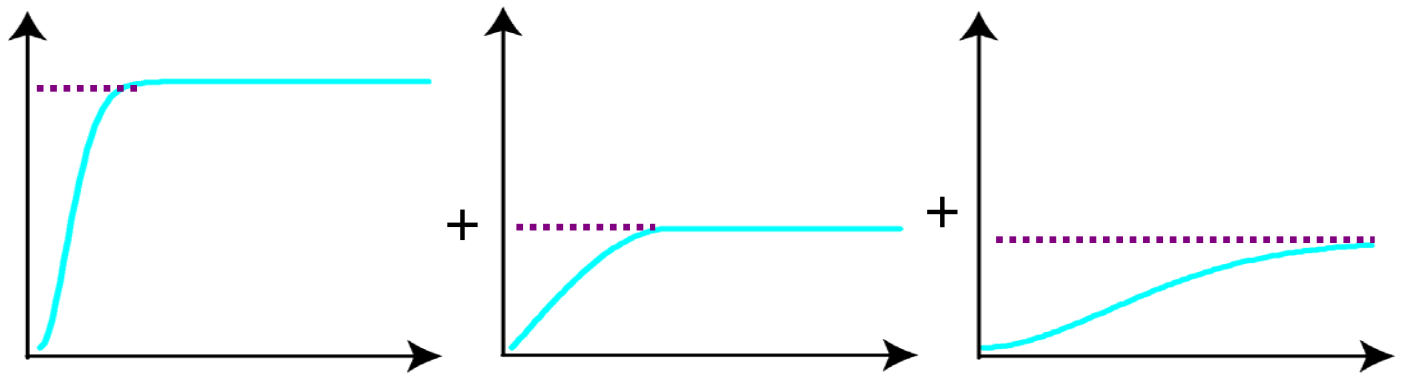
Modèle de variogrammes emboîtés :





$$\gamma(h) = \sum_{u=0}^{S} \gamma_u(h) = \sum_{u=0}^{S} b_u g_u(h)$$

➢ ***Choix du nombres d'échelles et des familles des modèles***



Portées différentes : échelles caractéristiques

Modèle hiérarchique multi-échelle non-stationnaire

Modèle :

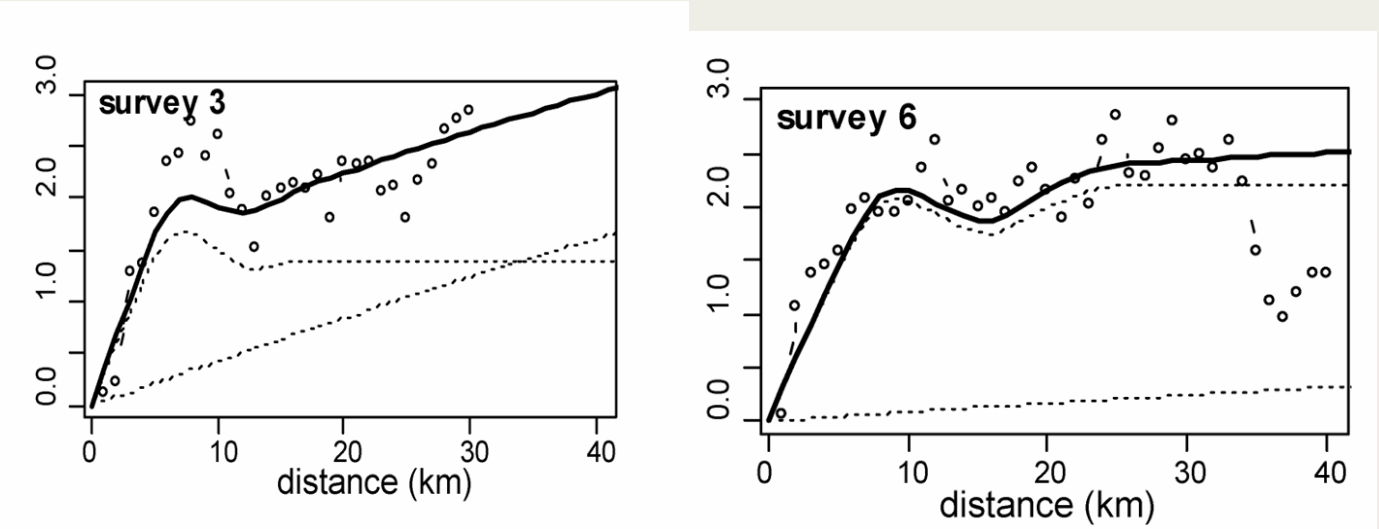
$$\begin{cases} Z_s | Y_s & \sim \mathcal{P}(Y_s) \\ Y_s & = m_s X_s \end{cases}$$

- Z_s : Données de comptage en s, Y_s densité spatiale
- m_s : tendance déterministe
- X_s : champ aléatoire positif, multi échelles de moyenne unitaire

Variogramme empirique de la variable latente X

$$\gamma_X^*(h) = \frac{1}{2\,N(h)} \sum_{\alpha,\beta \simeq d(h)} \left(\frac{m_\alpha\, m_\beta}{m_\alpha + m_\beta} \left(\frac{Z_\alpha}{m_\alpha} - \frac{Z_\beta}{m_\beta} \right)^2 - 1 \right)$$

Exemple: guillemots dans le Golfe de Gascogne (E. Bellier)



Fin de la troisième partie

Partie 4 : Les différents Krigeages :

Sous hypothèse de stationnarité d'ordre 2, la moyenne m est inconnue, on krige une valeur ponctuelle Z_o^* en x_o

L'estimateur de krigeage s'écrit sous la forme

$$Z_o^* = \sum_{\alpha=1}^{n} \lambda_\alpha Z_\alpha$$

où les λ_α sont des poids tels que :

- Z_o^* est sans biais
- l'erreur $Z_o^* - Z_o$ est de variance minimum

Pour le Krigeage Ordinaire, on obtient un système à $(n+1)$ équations

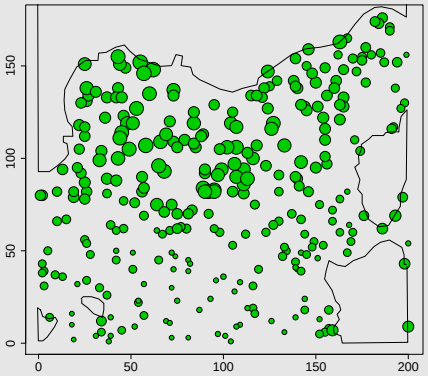
$$\begin{pmatrix} C_{11} & \dots & C_{1n} & 1 \\ \vdots & & \vdots & \vdots \\ C_{n1} & \dots & C_{nn} & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} C_{1o} \\ \vdots \\ C_{no} \\ 1 \end{pmatrix}$$

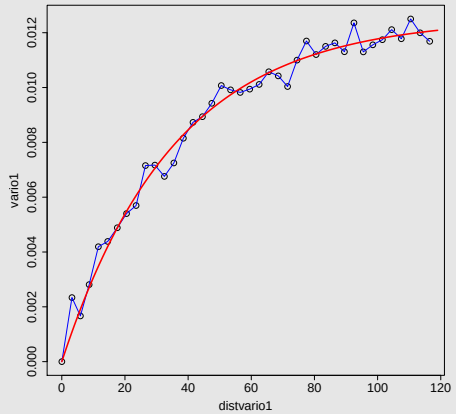
appellé système de krigeage.

La variance d'erreur de prédiction est appelée Variance de Krigeage σ^2_{KO}
Sa valeur est :

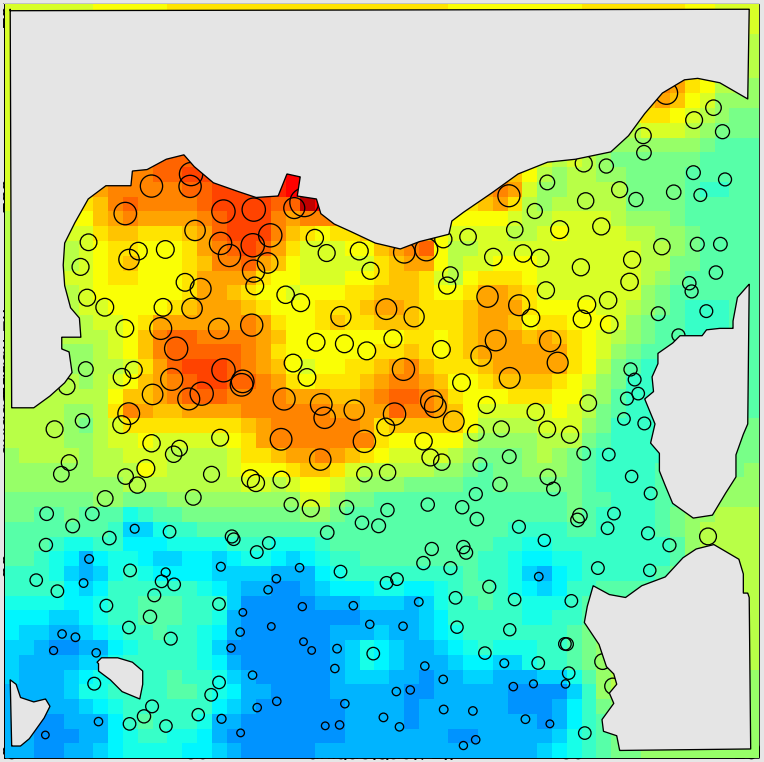
$$\sigma^2_{KO} = \sigma^2 - \sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha o} + \mu$$

Krigeage ordinaire:



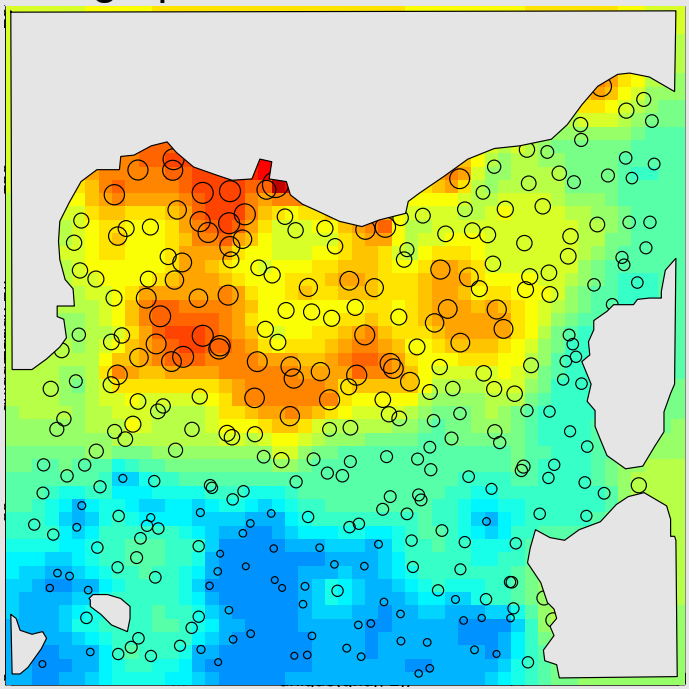


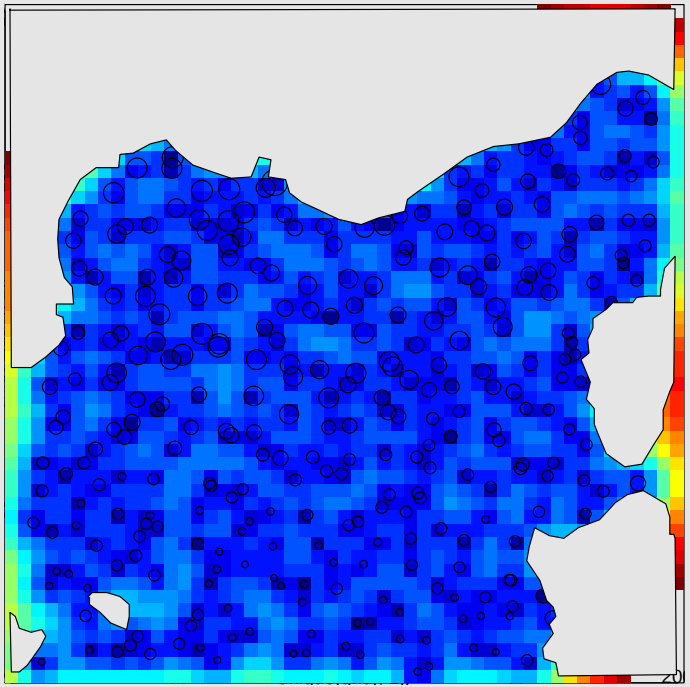
⇒



Krigeage ordinaire:

La carte des variances d'erreurs est aussi prédite avec la variable à cartographier





Krigeage Universel : modèle sous-jacent

- $Z(x) = m(x) + Y(x)$
  $m(x)$: tendance déterministe qui varie lentement
  $Y(x)$: résidu aléatoire, stationnaire d'ordre 2
  E $[Y(x)] = 0$
  Cov$(Y(x), Y(x+h)) = C(h)$

- On décompose $m(x)$ comme combinaison linéaire de fonctions de base.

$$m(x) = \sum a_l f_l(x) \qquad l = 0, \ldots, L$$

  $f_l(x)$: fonctions de base non quelconques avec $f_o(x) = 1$
  Ex: $1, x, y, x^2, y^2, xy$,...
  a_l : coefficients locaux, inconnus
  L : ordre de continuité de $m(x)$: constante, linéaire, quadratique ...

Krigeage Universel :

Prédiction de Z_o^* par $Z_o^* = \sum_{\alpha=1}^{n} \lambda_\alpha Z_\alpha$ où les λ_α vérifient :

- Z_0^* est sans biais et l'erreur est de variance minimum

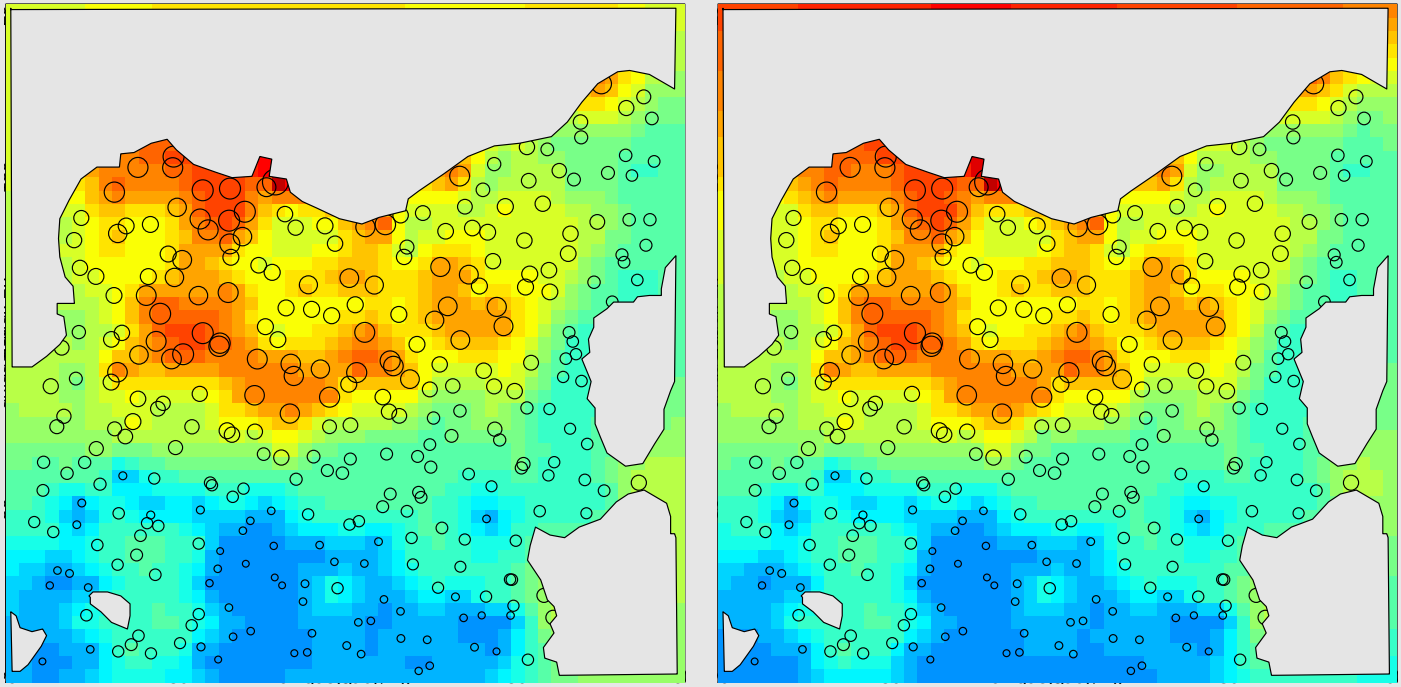
On procède de la même manière que pour le KO, et on obtient le système :

(n+L+1) équations

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_\beta C_{\alpha\beta} - \sum_{l} \mu_l f_\alpha^l = C_{\alpha o} & \text{pour } \alpha = 1, \ldots, n \\ \sum_{\alpha=1}^{n} \lambda_\alpha f_\alpha^l = f_o^l & \text{pour } l = 0, \ldots, L \end{cases}$$

Krigeage ordinaire et universel :

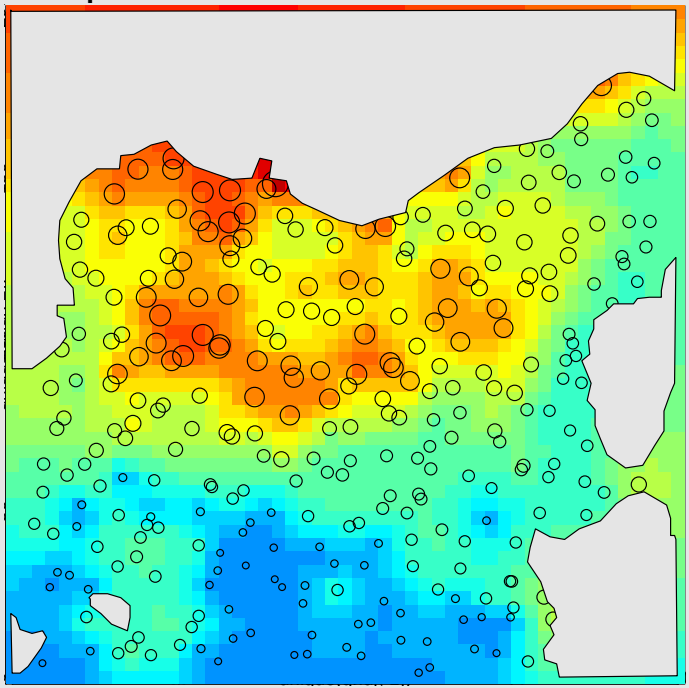
Comparaison

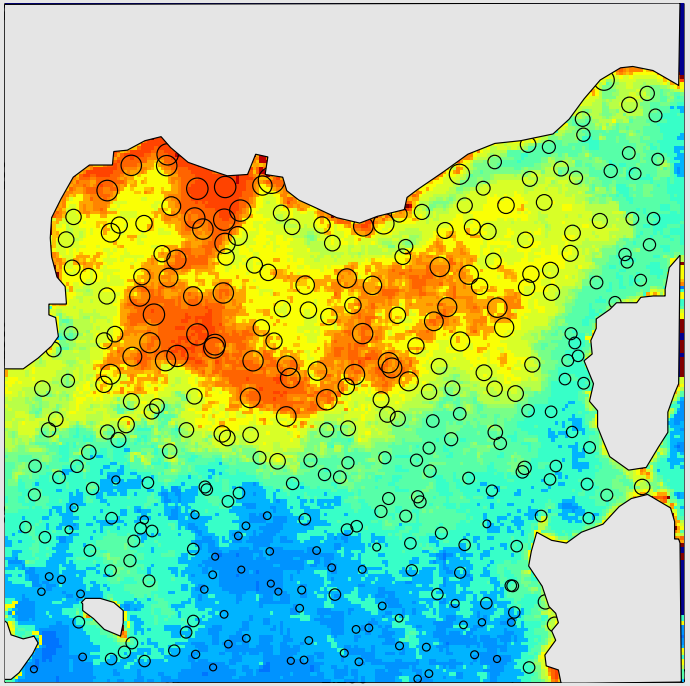


Un autre possibilité est d'estimer le gradient et de kriger les résidus

Krigeage universel:

Comparaison avec la ”vérité terrain”

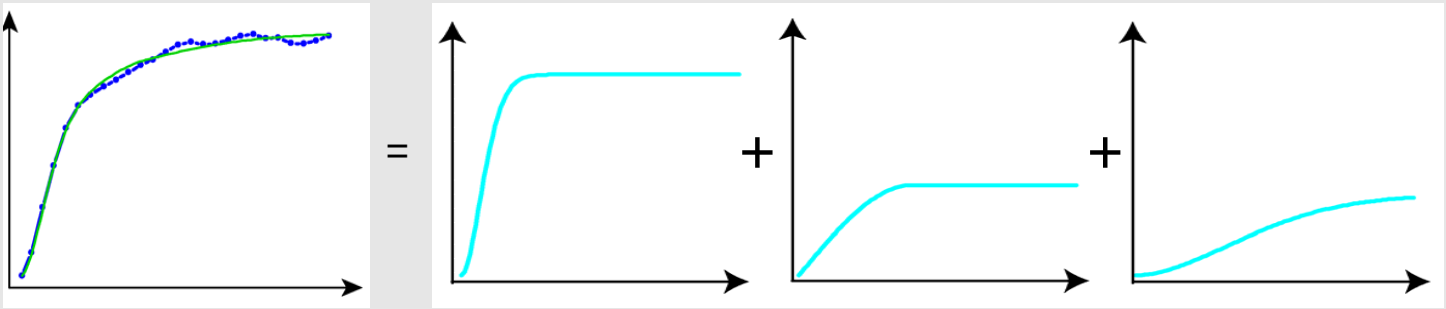


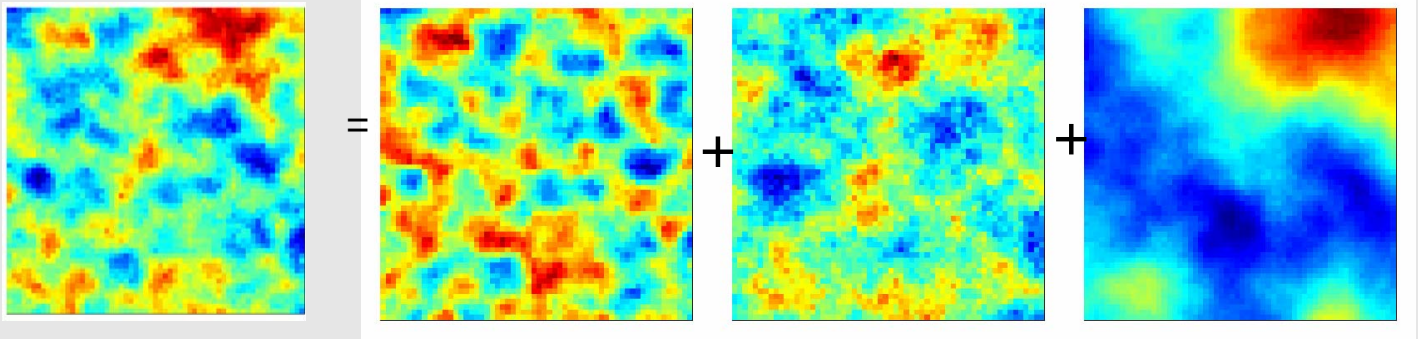


Autres exemples : krigeages

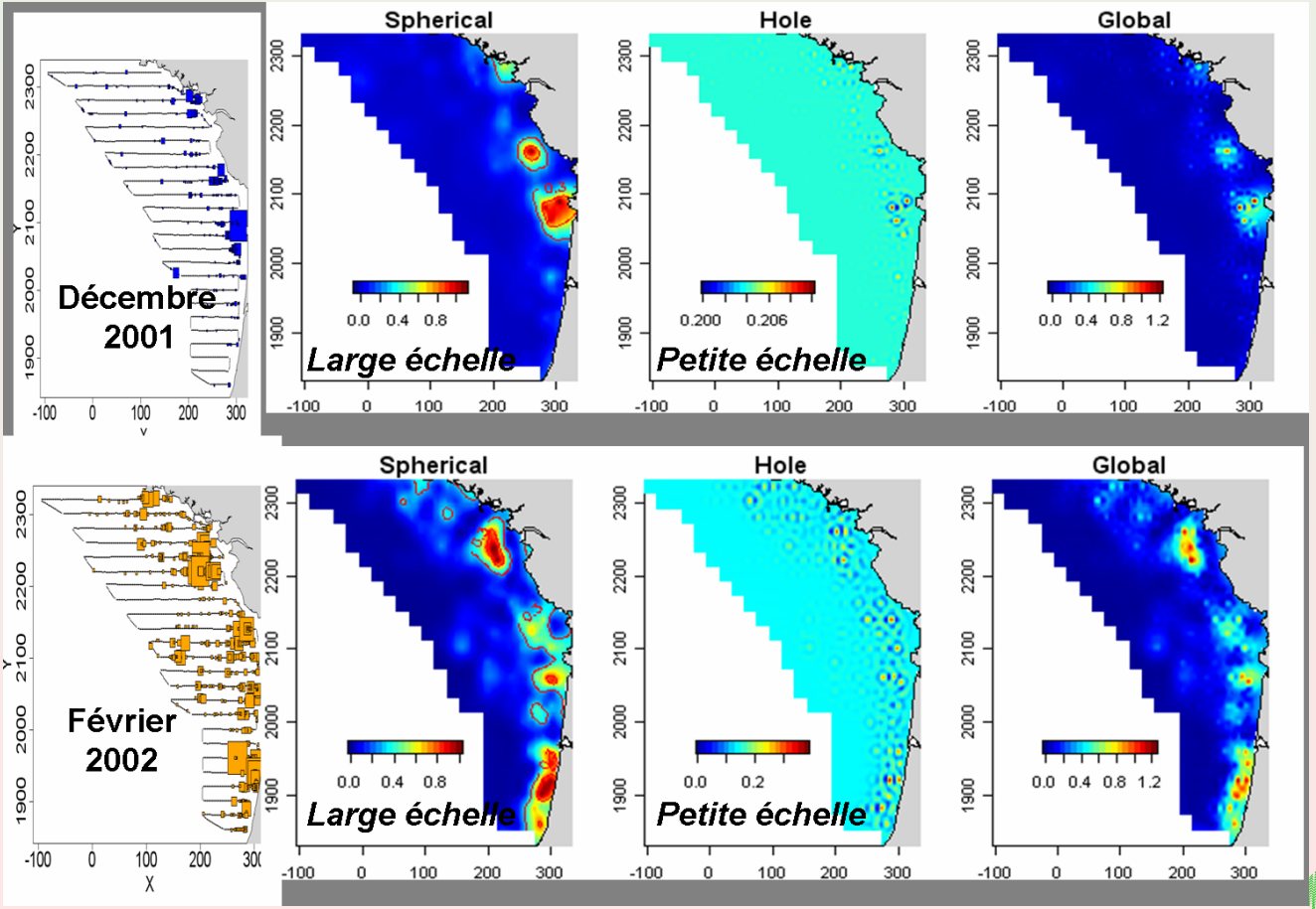
Les composants d'un modèle de régionalisation peuvent être extraits par krigeage filtrant :

$$Z_S^*(x_0) = \sum_{\alpha=1}^{n} w_\alpha^S Z(x_\alpha)$$

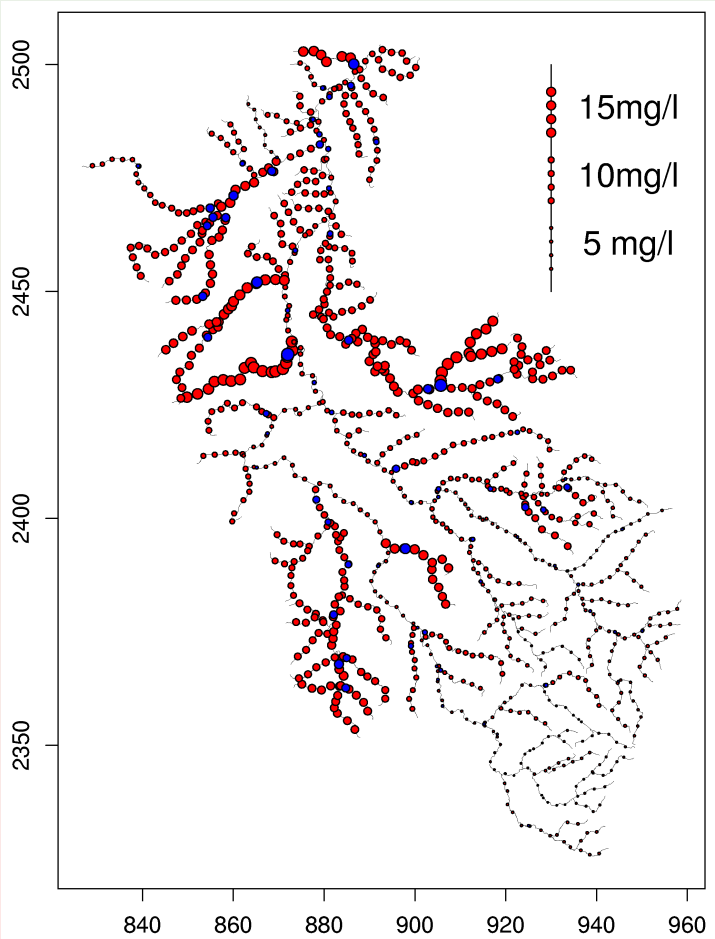




Exemple: guillemots dans le Golfe de Gascogne (E. Bellier)

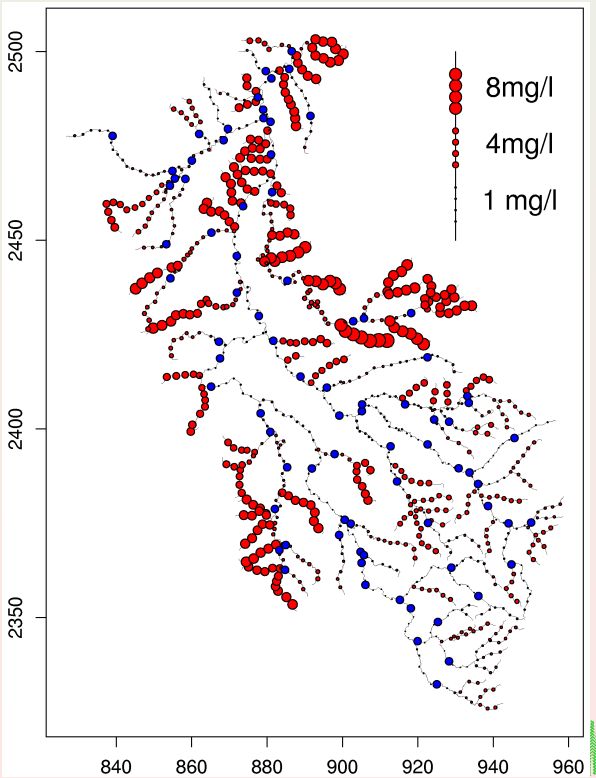


Kriging : Nitrate Rates, *Hybrid* type model



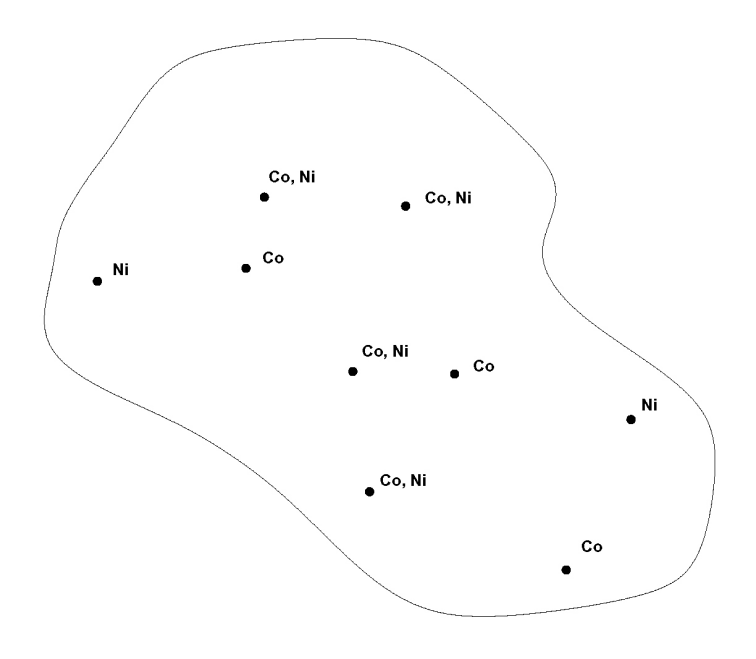
⇐ Kriged map

Confidence interval width ⇓



Fin de la quatrième partie

Partie 5 : Géostatistique Multivariable



1. Notations
2. Modélisation multivariable
   - Covariance croisée
   - Variogramme croisé
3. Cokrigeage

Cela se complique, même à 2 variables

- On considère 2 fonctions aléatoires simultanées $Z_1(x), Z_2(x)$.
- Chaque $Z_i(x)$ est échantillonnée sur un ensemble S_i de n_i points.

  Ce ne sont pas forcément les mêmes points
- On utilise les lettres grecques pour les échantillons : $\alpha, \beta, \ldots$

  $Z_1(x_\alpha) \;\; \alpha = 1, n_1 \quad Z_2(x_\beta) \;\; \beta = 1, n_2$
- Peut être généralisé à p fonctions aléatoires.

Covariance croisée

- On est en hypothèse stationnaire d'ordre 2 :
  moyennes et covariances existent et sont stationnaires.

$$E[Z_i(x)] = m_i$$

$$E\left[(Z_i(x) - m_i).(Z_i(x+h) - m_i)\right] = C_{ii}(h),$$

  où $i = 1, 2 \quad x, x+h \in D$.

- On définit la covariance croisée :

$$C_{ij}(h) = E\left[(Z_i(x) - m_i).(Z_j(x+h) - m_j)\right],$$

  où $1 \leq i \neq j \leq 2 \quad x, x+h \in D$. $C_{ij}(h)$ n'est pas symétrique autour de 0, et est sensible aux "effets de décalage".

Variogramme croisé

- On est en hypothèse intrinsèque,

$$E[Z_i(x+h) - Z_i(x)] = 0$$

$$E[(Z_i(x+h) - Z_i(x)).(Z_i(x+h) - Z_i(x))] = 2\gamma_{ii}(h)$$

où $i = 1, 2 \quad x, x+h \in D.$

- On définit le variogramme croisé :

$$\gamma_{ij}(h) = \frac{1}{2} E[(Z_i(x+h) - Z_i(x)).(Z_j(x+h) - Z_j(x))]$$

où $1 \leq i \neq j \leq 2, \quad x, x+h \in D.$

Estimation

Comme pour les covariances et les variogrammes simples, on utilise les estimateurs empiriques :

$$\hat{C}_{ij}(h) = \frac{1}{2N(h)} \sum_{x_\beta - x_\alpha \simeq h} (Z_i(x_\alpha) - m_i)(Z_j(x_\beta) - m_j)$$

$$\hat{\gamma}_{ij}(h) = \frac{1}{2N(h)} \sum_{x_\beta - x_\alpha \simeq h} \Big(\big(Z_i(x_\beta) - Z_i(x_\alpha)\big).\big(Z_j(x_\beta) - Z_j(x_\alpha)\big) \Big)$$

où

$$N(h) = \text{nombre de couples tels que } x_\beta - x_\alpha \simeq h$$

Pour estimer le variogramme croisé il faut connaître les 2 variables aux mêmes points d'échantillonnage (x_α, x_β).

Cela n'est pas nécessaire pour la covariance croisée.

Lien entre covariance et variogramme croisés

On se place en hypothèse stationnaire d'ordre 2. Il existe une covariance croisée. Alors:

$$\begin{aligned}\gamma_{ij}(h) &= \cdots \\ &= C_{ij}(0) - \frac{1}{2}(C_{ij}(h) + C_{ij}(-h))\end{aligned}$$

Or,

$$C_{ij}(h) = \frac{1}{2}(C_{ij}(h) + C_{ij}(-h)) + \frac{1}{2}(C_{ij}(h) - C_{ij}(-h))$$

$\Rightarrow \gamma_{ij}(h)$ correspond à la partie paire de $C_{ij}(h)$.

$C_{ij}(h)$ est un outil structural plus puissant que $\gamma_{ij}(h)$ mais nécessitant des hypothèses plus fortes.

Modèle de corégionalisation

Problème difficile :

Trouver un ensemble de $p \times p$ variogrammes (ou covariances) tel que la variance de toute combinaison linéaire de covariable soit positive, i.e. :

$$Var\left(\sum_{i=1}^{p}\sum_{\alpha_i}\lambda_{\alpha_i}Z_i(x_\alpha)\right) \geq 0$$

Il existe très peu de modèles généraux assurant cette condition.

Modèle de corégionalisation linéaire

Chaque variable $Z_i(x)$ est décomposée en une somme de composantes non corrélées:

$$Z_i(x) = \sum_{u=0}^{S} Z_i^u(x) + m_i$$

avec

$$\begin{aligned} E[Z_i^u(x)] &= 0 \\ Cov(Z_i^u(x), Z_j^u(x+h)) &= C_{ij}^u(h) = b_{ij}^u \rho_u(h) \\ Cov(Z_i^u(x), Z_j^v(x+h)) &= 0 \quad \text{si } u \neq v \end{aligned}$$

Alors, chaque covariance $C_{ij}(h)$ est décomposée en somme de fonctions de corrélation emboîtées :

$$C_{ij}(h) = \sum_{u=0}^{S} C_{ij}^u(h) = \sum_{u=0}^{S} b_{ij}^u \rho_u(h)$$

où S est le nombre de structures de base. $u = 0$ correspond à l'effet de pépite.

Propriétés :

1. Les fonctions $\rho_u(h)$ sont les mêmes pour toutes les fonctions $C_{ij}(h)$. Elles correspondent par exemple à des échelles différentes (ex : une pépite, une courte portée, une longue portée.)
2. On a
$$\sum_{u=0}^{S} b_{ii}^u = Var(Z_i).$$
Les b_{ii}^u sont donc une décomposition de la variance pour les différentes échelles.
3. Ce qui caractérise un couple de variable (Z_i, Z_j), ce sont les valeurs $b_{ij}^u,\ u = 1, \ldots, S$.
4. Ce qui caractérise une échelle de travail (donc, une structure $\rho_u(h)$), c'est la matrice $(p \times p)$ $\mathbf{B}^u$ des valeurs $b_{ij}^u,\ 1 \leq i, j \leq p$.

La condition de définie positivité devient :
il faut que chaque matrice $\mathbf{B}^u,\ u = 1, \ldots, S$ soit définie positive.

Pour deux variables :

$$|b_{12}^u| \leq \sqrt{b_{11}^u b_{22}^u}$$

Si on travaille en variogramme, $g_u(h) = 1 - \rho_u(h)$ est la fonction structurale de base, et on a

$$\gamma_{12}(h) \leq \sum_{u=0}^{S} \sqrt{b_{11}^u b_{22}^u}\, g_u(h)$$

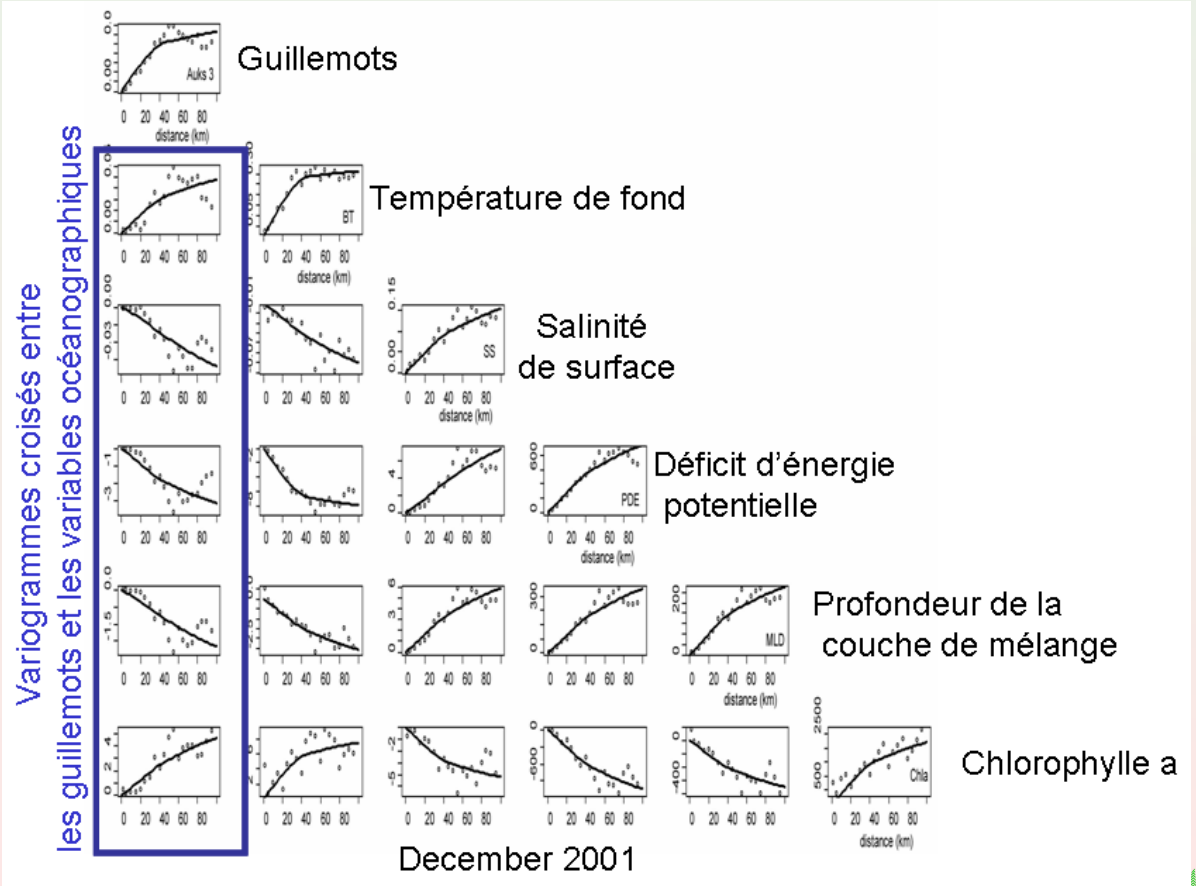
En écriture matricielle:

$$\mathbf{\Gamma}(h) = \sum_{u=0}^{S} \mathbf{B}^u g_u(h)$$

Ajustement : on propose les $g_u(h)$, et on recherche les $\mathbf{B}^u$ qui minimisent un critère de moindre carré.

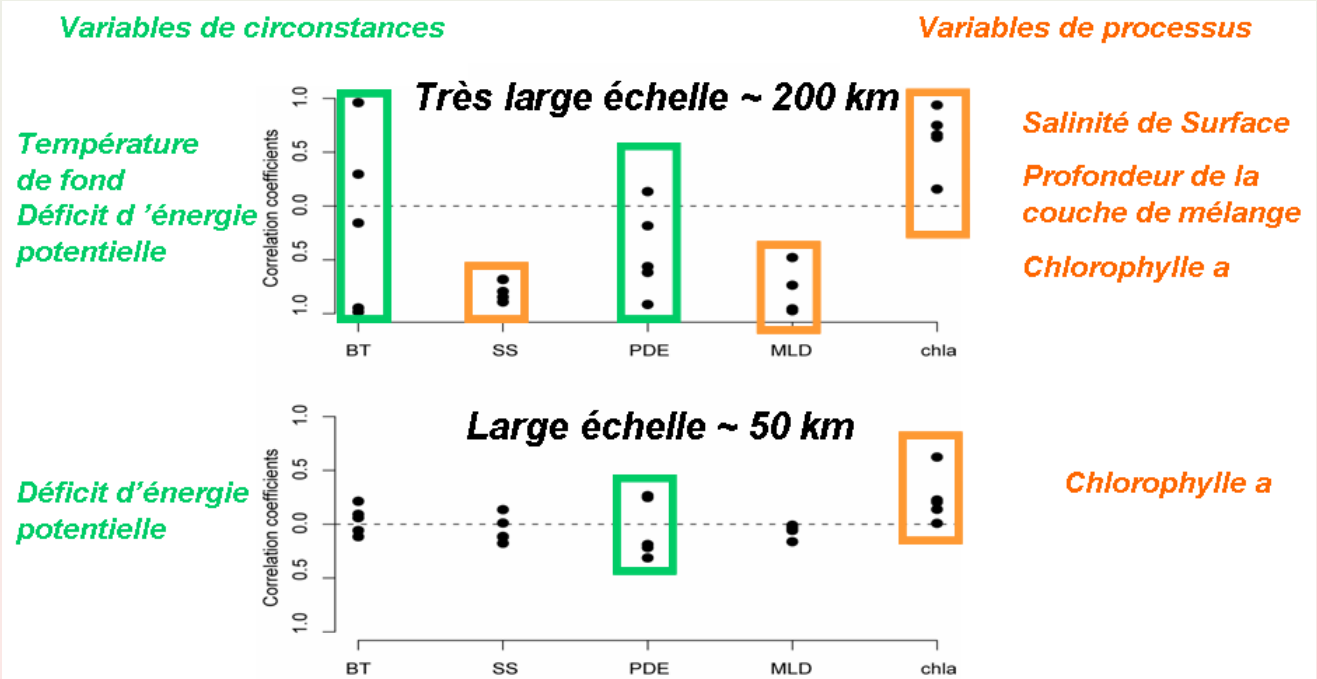
Un exemple avec les guillemots

Exemple: distribution de guillemots et variables océaniques (E.Bellier)



Exemple: distribution de guillemots et variables océaniques (E.Bellier)

Liaisons croisées avec l'environnement :



Fin de la cinquième partie

Partie 6 : Le(s) cokrigeage(s)

Le co-krigeage est l'extension naturelle du krigeage au cas multivariable. On estime (au point x_0) une variable d'intérêts (Z_1, par convention) à partir des autres valeurs $Z_1(x_\alpha)$ et des variables auxiliaires $Z_i(x_\beta)$, $i \neq 1$.

On suppose qu'on connaît les variogrammes $\gamma_{ij}(h)$ simples et croisés.

On construit

$$Z_1^*(x_0) = \sum_{i=1}^{p} \sum_{\alpha=1}^{n_i} \lambda_\alpha^i Z_i(x_\alpha)$$

note: les n_i peuvent être différents.

Système de Cokrigeage

$$\sum_{j=1}^{p}\sum_{\beta=1}^{n_j}\lambda_\beta^j\,\gamma_{ij}(x_\alpha - x_\beta) + \mu_i \;=\; \gamma_{1i}(x_\alpha - x_0) \quad \text{pour} \quad \begin{matrix} i = 1,\ldots,p \\ \alpha = 1,\ldots,n_i \end{matrix}$$

$$\sum_{\alpha=1}^{n_i}\lambda_\alpha^i \;=\; \delta_{1i} \qquad \text{pour } i = 1,\ldots,p$$

C'est un système dont la taille devient vite énorme en voisinage unique.

Variance de cokrigeage :

$$\sigma^2_{CK} = \sum_{j=1}^{p}\sum_{\alpha=1}^{n_i}\lambda_\alpha^i\gamma_{1i}(x_\alpha - x_0) + \mu_1.$$

Remarques :

- Des variables non ou très peu corrélées ne servent à rien dans le cokrigeage.
- Pour reduire la taille du système $\Rightarrow$ utiliser le voisinage glissant $\Rightarrow$ besoin de points informant de Z_1 dans le voisinage.

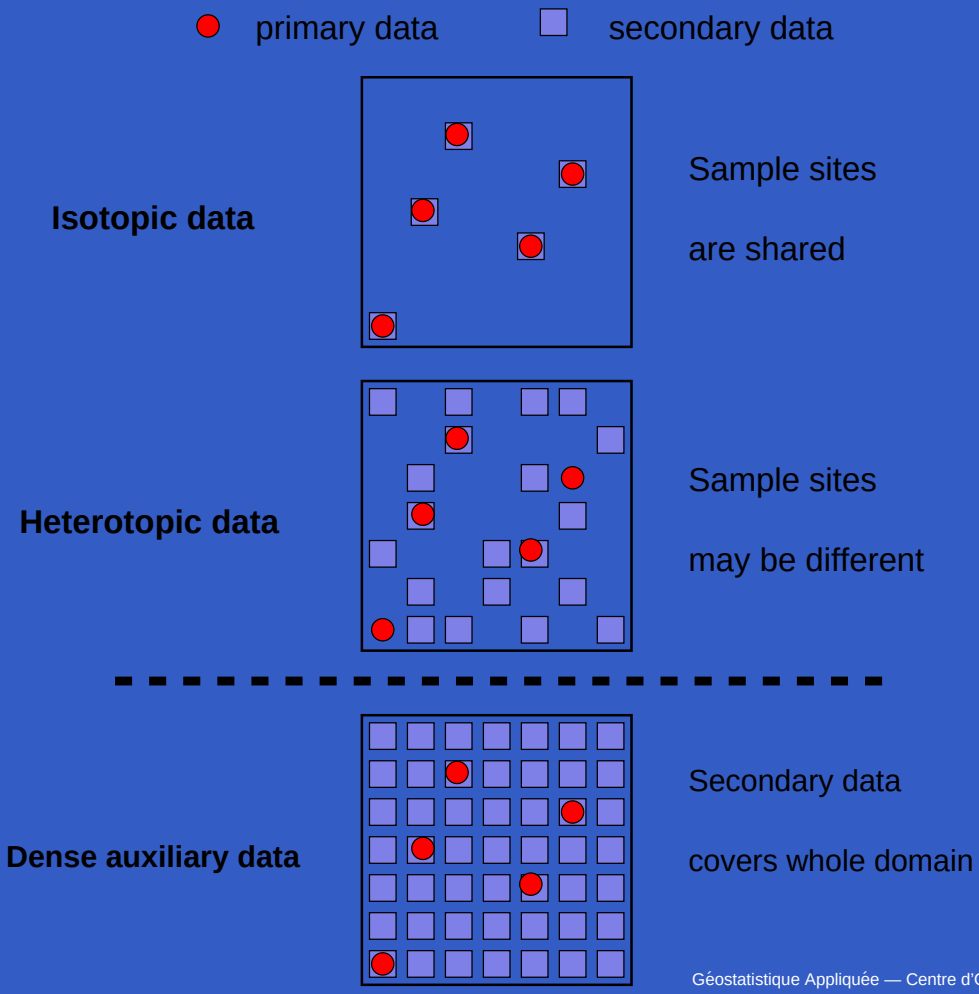
Remarques :

- Des variables non ou très peu corrélées ne servent à rien dans le cokrigeage.
- Pour reduire la taille du système $\Rightarrow$ utiliser le voisinage glissant $\Rightarrow$ besoin de points informant de Z_1 dans le voisinage.
- $\sigma^2_{CK}(Z_1(x_0) \mid (Z_1(x_\alpha)), \ldots, (Z_p(x_\alpha))) \leq \sigma^2_K(Z_1(x_0) \mid (Z_1(x_\alpha)))$.
  En théorie, il est toujours avantageux de prendre en compte les autres variables.
- Si
$$C_{ij}(h) = b_{ij}.\rho(h), \quad \forall i, j$$
on parle de *modèle de corrélation intrinsèque*. Il n'y a qu'une seule structure spatiale, et les variances / covariances entre variables sont données par les b_{ij}, indépendamment de l'échelle.
Dans ce cas, $Z^{CK}(x) = Z^{KS}(x)$.

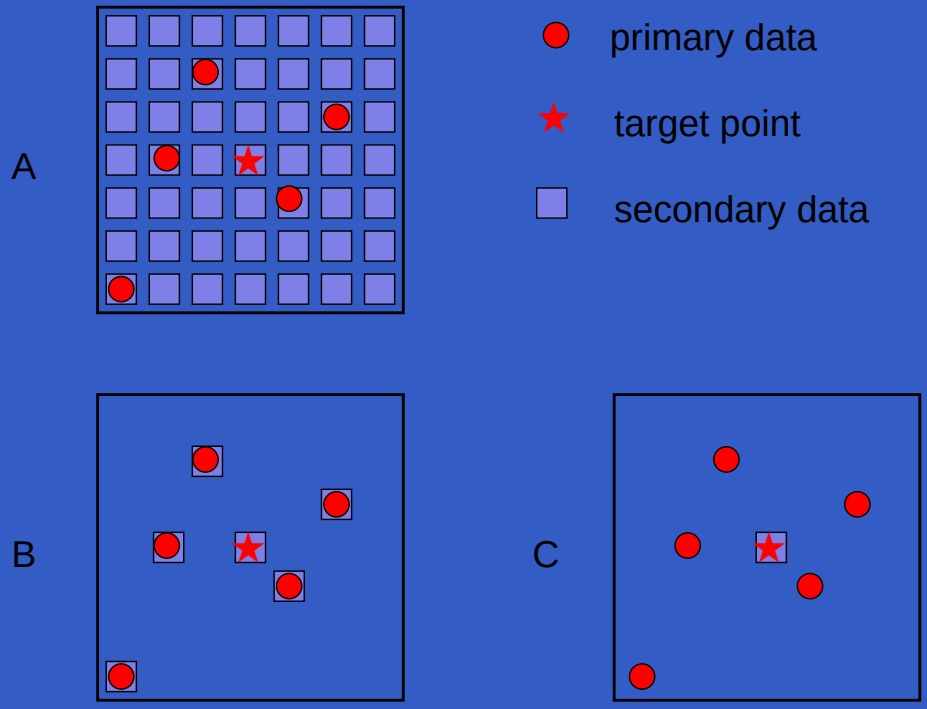
Remarques :

- Des variables non ou très peu corrélées ne servent à rien dans le cokrigeage.
- Pour reduire la taille du système $\Rightarrow$ utiliser le voisinage glissant $\Rightarrow$ besoin de points informant de Z_1 dans le voisinage.
- $\sigma^2_{CK}(Z_1(x_0) \mid (Z_1(x_\alpha)), \ldots, (Z_p(x_\alpha))) \leq \sigma^2_K(Z_1(x_0) \mid (Z_1(x_\alpha)))$.
  En théorie, il est toujours avantageux de prendre en compte les autres variables.
- Si
$$C_{ij}(h) = b_{ij}.\rho(h), \quad \forall i, j$$
  on parle de *modèle de corrélation intrinsèque*. Il n'y a qu'une seule structure spatiale, et les variances / covariances entre variables sont données par les b_{ij}, indépendamment de l'échelle.
  Dans ce cas, $Z^{CK}(x) = Z^{KS}(x)$.
- Lorsque toutes les variables sont informées en tout point de données, on parle d'*isotopie*. Sinon on parle d'hétérotopie

Configurations: Iso- and Heterotopic Data



Neighbors: dense auxiliary data



A: neighborhood using all data
B: multi-collocated
C: collocated

- Le *"Collocated cokriging"* propose de ne retenir que $Z_2(x_0)$ pour estimer $Z_1^*(x_0)$. $\Rightarrow$ réduit la taille du système à résoudre.
- Dans ce cas, on peut aussi utiliser le krigeage avec dérive externe. Il s'agit d'un modèle différent où la covariable modélise une non-stationnarité en moyenne.

Krigeage avec dérive externe

On veut interpoler $Z(x)$ faiblement échantillonnée, en utilisant une variable auxiliaire $u(x)$, connue en tous points d'interpolation.

Ex : Z = température ; u=altitude ;

Le modèle :
On considère $Z(x)$ aléatoire et $u(x)$ déterministe.

$$\begin{aligned} E[Z(x)] &= a + bu(x) \\ Cov(Z(x), Z(x+h)) &= C(h) \end{aligned}$$

Estimateur :

$$Z^*(x_0) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha)$$

Système de Krigeage avec dérive externe :

Variance $\sigma_E^2 = Var(Z^*(x_0) - Z(x_0))$ minimisé sous les deux contraintes de non-biais mène au système

$$\sum_{\beta=1}^{n} \lambda_\beta C(x_\alpha - x_\beta) - \mu_1 - \mu_2 u(x_\alpha) = C(x_\alpha - x_0) \text{ pour } \alpha = 1, \ldots, n$$

$$\sum_{\beta=1}^{n} \lambda_\beta = 1$$

$$\sum_{\beta=1}^{n} \lambda_\beta u(x_\beta) = u(x_0)$$

Krigeage avec dérive externe

- Difficulté d'estimer $C(h)$, car l'estimateur empirique $\hat{C}(h)$ dépend fortement de $u(x)$.
- Dérive externe vs. co-krigeage:
  - La DE n'est pas adaptée si il faut introduire de l'aléatoire dans $u(x)$.
  - Le CoK utilise $u(x)$ de façon linéaire, alors que la DE utilise cette information au travers des conditions de non biais (de façon non-linéaire).
  - Le lien induit est plus fort en DE qu'en CoK.

Merci