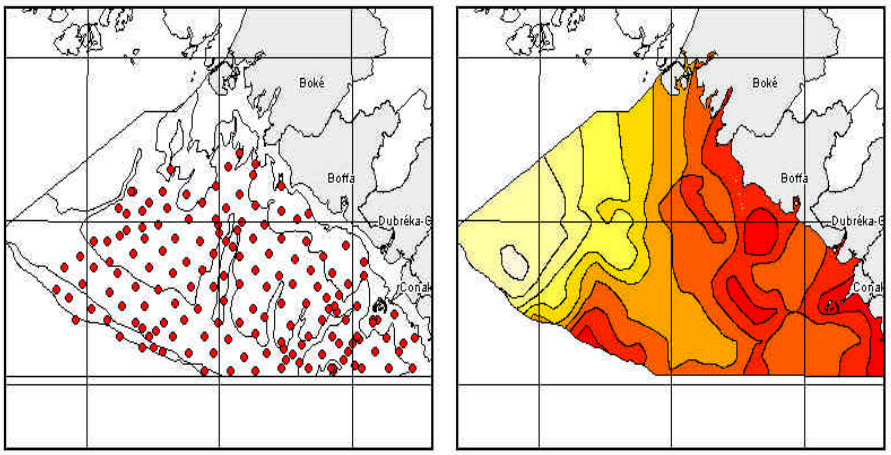
Statistiques et Interpolations dans les SIG

Laurent DRAPEAU
Centre I.RD Montpellier
Laboratoire HEA

Introduction

L'outil Système d'Information Géographique permet de gérer et d'analyser les données sur les critères des bases de données (valeurs des attributs), et sur les caractéristiques spatiales des éléments géographiques en jeu. Les fonctionnalités des SIG sont opérationnelles si l'on dispose d'une bonne «couverture spatiale » des phénomènes que l'on étudie. L'halieutique s'intéresse à la confrontation des données relatives à l'exploitation, à la ressource et à l'environnement. On ne dispose cependant pas dans chacune de ces composantes d'une même échelle spatiale d'observation. Il est impossible de collecter de façon exhaustive les données en tous les points de l'espace, ceci pour des raisons pratiques évidentes (coûts, inaccessibilité…). Le problème sous-jacent est celui de l'interpolation des données. L'interpolation est un moyen de générer l'information aux points de l'espace non enquêtés, cela pour la cartographie et l'analyse en 2D du phénomène. Il s'agit de fournir un maillage adapté. Il est important de noter que l'interpolation, qu'elle soit déterministe ou probabiliste, est le résultat d'un traitement des données, qui permet leur exploitation dans un SIG. Les SIG intègrent de plus en plus souvent des fonctionnalités avancées d'interpolation ou des liens souples avec les outils adaptés (ArcView et module spatial de S+). L'objectif est ici de rappeler quelques méthodes et principes élémentaires de l'interpolation. On peut classifier les méthodes en deux approches : l'approche déterministe, l'approche probabiliste. Les développements des réseaux neuronaux ont plus récemment abordé le problème de l'interpolation et fournissent une troisième approche.

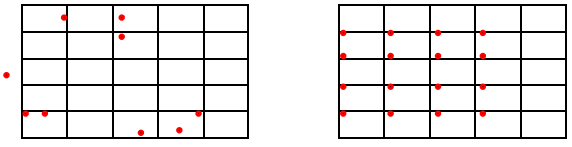


Température de surface sur la ZEE guinéenne (saison sèche) (ref: Atlas des Pêches Maritimes de Guinée)

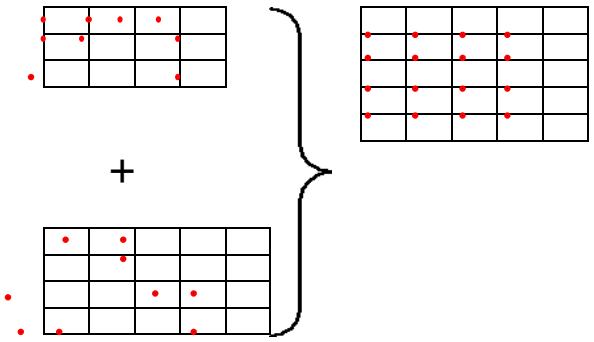
Si le principe de base est invariant (fournir des valeurs en des lieux non échantillonnés) les raisons peuvent être variées. Les données ont été échantillonnées selon un plan aléatoire et l'on désire avoir un maillage régulier, on désire confronter des variables échantillonnées dans des plans qui ne coïncident pas. On désire construire des isolignes ou iso-surfaces. Dans tous les cas, une analyse locale des données est nécessaire, et le choix de la méthode très important.

Les besoins sous-jacents

- Exploitation dans un SIG, superposition de thématiques, agrégation spatiale
- Obtenir des informations sur un maillage régulier à partir d'un plan d'échantillonnage aléatoire



- **Redresser un plan d'échantillonnage systématique sur un maillage adapté à la comparaison de carte**

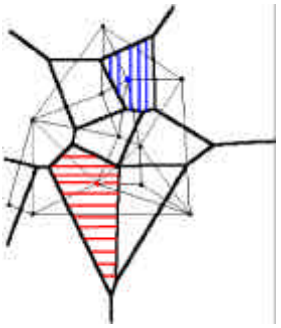


Méthodes d'interpolation

L'approche déterministe :

Elle regroupe les méthodes d'interpolation dont la fonction de structure (fonction de pondération) est choisie à priori. Elles ne fournissent pas d 'informations sur la variance d'estimation.

VORONOï :



Basée sur des critères de voisinages (partitionnment géométrique) simples elle construit une parcellisation du domaine d'étude. Chaque cellule contient un et un seul point de l'échantillon, l'ensemble des points de l'espace appartenant à la cellule a pour plus proche voisin le point d'échantillonnage associé à la cellule. La valeur du point échantillonné est associée à tous les points de la parcelle ou cellule. Cette approche est semblable à la triangulation. Les limites sont évidentes, car il y a de brusques saut de discontinuité.

Surface trend

Cette méthode utilise la combinaison linéaire de fonctions de bases. Il s'agit d'appréhender le phénomène par l'utilisation d'un polynôme dont le degré est laissé à l'estimation du thématicien. D'une manière générale on écrit : $Z=a+b_1x+b_2y+b_3xy+b_4x^2+b_5y^2+....$ Les paramètres du modèle sont estimés par la méthode des moindres carrés. Celle-ci vise à minimiser l'erreur entre la valeur prédite et la valeur observée

$$\sum_i (f(X_i) - Z_i)^2$$

Cette méthode s'intéresse peu aux changements locaux, aux irrégularités, mais seulement à la tendance. Le choix du degré fixe la complexité de l'interpolation dans la pratique on utilise des degrés de 3 à 5. Cette méthode d'interpolation à l'inconvénient de s'accroître ou décroître rapidement dans les zones ou l'échantillonnage est faible et particulièrement sur les bords de la zone.

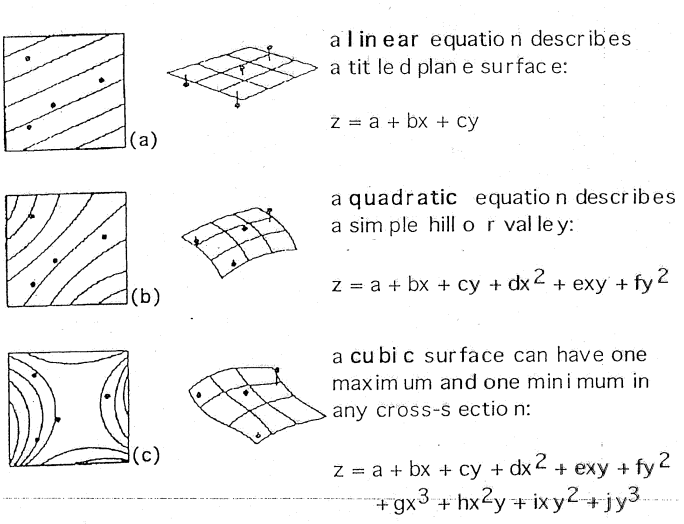


Figure 15 Types of trend surface analysis
Source: Based on Burrough (1986)

Les fonctions splines :

Les fonctions splines (Briggs 73) permettent d'interpoler à partir de données réparties aléatoirement. L'interpolation est rendue plus flexible que par le polynôme grâce à un paramètre de tension qui contrôle le comportement de la fonction d'interpolation et le paramètre de lissage.

Inverse distance : Il s'agit d'une méthode de moyenne pondérée ou chaque valeur de la grille à interpoler est calculée comme une moyenne pondérée des observations. Les facteurs de pondérations sont calculés proportionnellement à l'inverse de la distance élevée à une puissance c'est à dire :

$$1/d^{a}_{ij}$$

Cette méthode permet d'obtenir des grilles très rapidement mais crée des zones circulaires autour des valeurs observées (bull'eyes). Cet aspect peut être lissé en jouant sur la puissance et le voisinage. C'est un interpolateur exact (il passe par les valeurs observées).

Courbature minimum : La surface est calculée pour être la plus lisse possible.
Méthode de Shepard : Elle fonctionne de la même manière que Inverse distance mais permet de limiter l'effet Bull'eyes en intégrant un critère de moindre carré.

L'approche probabiliste

Le Krigeage est une interpolation qui estime les valeurs aux points non échantillonnés par une combinaison des données. Les poids des échantillons sont pondérés par une fonction de structure qui est issue des données. On tient ainsi compte des distances, des valeurs et des corrélations. La fonction n'est pas fixée à priori mais suite à l'analyse du variogramme. On considère que la valeur estimée en un point est le produit d'un processus sous-jacent, il fournit une variance d'estimation contrairement aux autres approches. Elle permet d'appréhender la structure spatiale du phénomène étudié. Le Krigeage s'inscrit donc dans une démarche d'analyse des données géostatistique.

Objectif : Prise en compte la structure spatiale du phénomène et variance d'estimation
Intérêt : Peu de contrainte sur le plan d'échantillonnage ni sur l'indépendance des données
Méthode : Approche géostatistique = Analyse de la structure spatiale + Krigeage

La géostatistique fournit un ensemble de méthodes statistiques qui décrivent l'autocorrélation spatiale des données de l'échantillon et les modélisent dans différents types de modèles spatiaux. Elle change l'approche méthodologique de l'échantillonnage en ce sens que les méthodes traditionnelles ne travaillent pas avec l'autocorrélation et que l'objectif est d'éviter la corrélation spatiale. Avec la géostatistique il y a moins donc moins de contraintes et amélioration de l'estimation des moyennes, et de la cartographie des phénomènes.

L'analyse structurale des données :

Prérequis : bonne couverture spatiale de la zone à étudier, hypothèse de stationnarité
Variogramme : mesure le degré de dissimilarité entre les points en fonction de leur éloignement (variance)

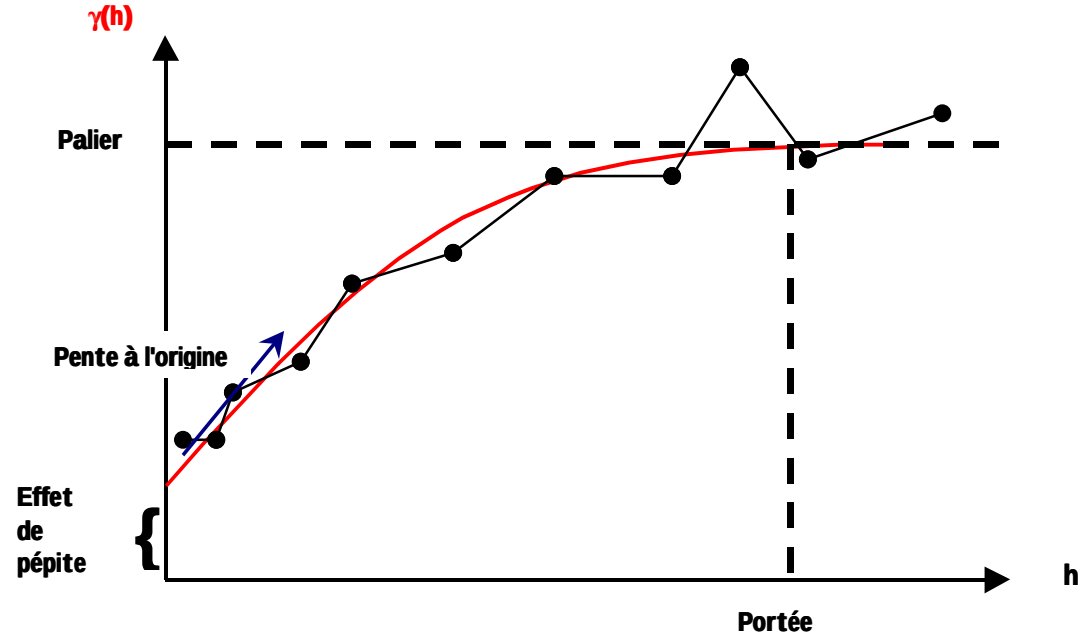
$$g^*(u,h) = \frac{1}{2n(u,h)} \sum_i (f(x_i) - f(x_i + h))^2$$

$g^*(u,h)$ *valeur du variogramme expérimental pour la distance h*

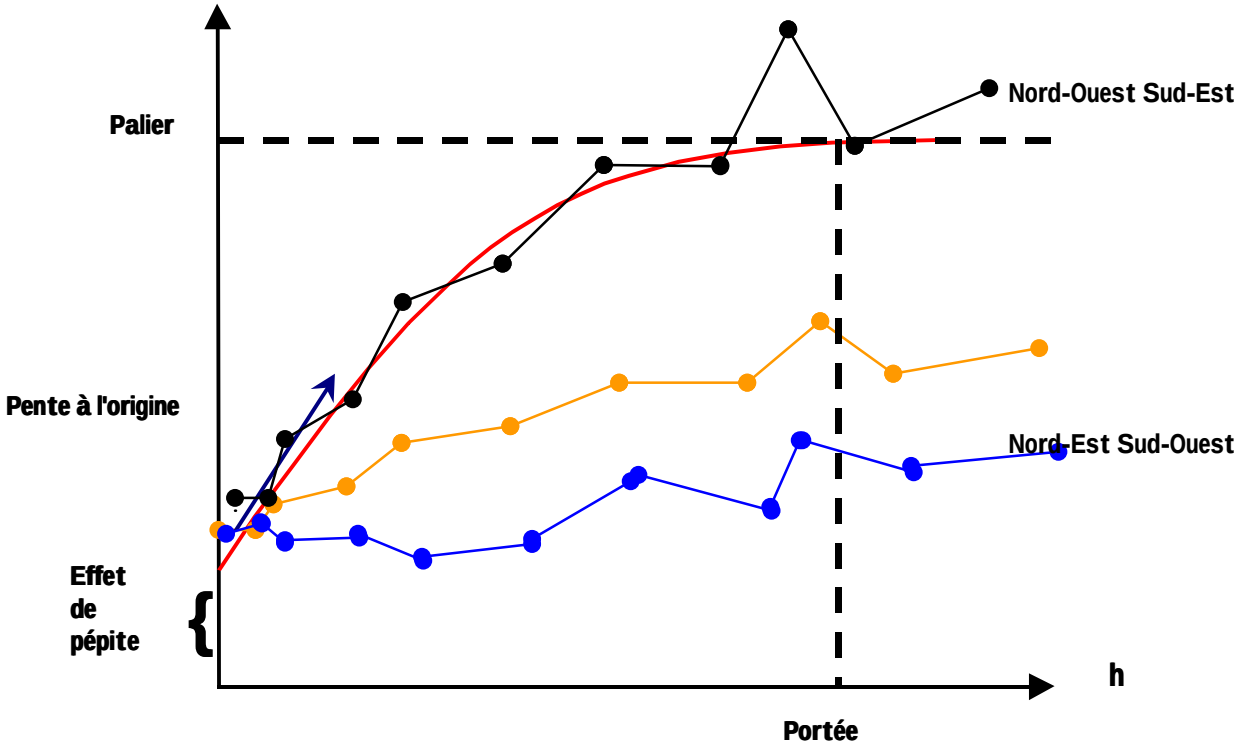
$f(x_i)$ *valeur du paramètre en* x_i

$n(u,h)$ *nombre de paires de direction u et de distance h*

Eléments d'un modèle de variogramme



Modèle de variogramme multidirectionnel



Précautions pour l'estimation du variogramme

☞**Sensibilité aux choix des classes de distances**

Le choix des classes (la tolérance) et leur nombre influence l'allure du variogramme
Une règle simple : le nombre de paires > 30 et h<D/2 (D=max(d_{ij})) (Journel et Huijbregts, 1978)

☞**Sensibilité au domaine d'étude**

La stationnarité du phénomène doit conduire à découper la zone d'étude si elle n'est pas observée sur le domaine global, des valeurs extrêmes (outliers) aussi.

☞**Sensibilité à l'anisotropie**

La présence d'anisotropie dans les données est très importante pour ajuster un modèle

Classes de distances	Classes de directions
- intervalles - nombre de classes - tolérance	- angle de référence - nombre de secteurs - tolérance angulaire

Analyse du variogramme

Caractéristiques : **Comportement à l'origine (h=0)**
Comportement à longue distance (palier et portée)
Comportement directionnel

A l'origine : **mesure de la continuité du phénomène.**
le degré d'irrégularité : tangente verticale = grande irrégularité, tangente horizontale = grande régularité

- discontinuité à l'origine purement aléatoire
- erreurs de mesures
- structures au rang inférieur à l'échelle de l'échantillon.

A longue distance : borné : les données ont un maximum d'hétérogénéité (variance = palier), la portée = distance à laquelle le palier est atteint. (théoriquement deux points séparés par d>portee sont non corrélés) Zone d'influence, autour des points. Ce paramètre est à relier au diamètre des « patches ». Non borné : pas de palier, la variance augmente indéfiniment.

Choix d'un modèle : Prise en compte des caractéristiques du variogramme
Ajustement purement statistique de type moindre carré possible

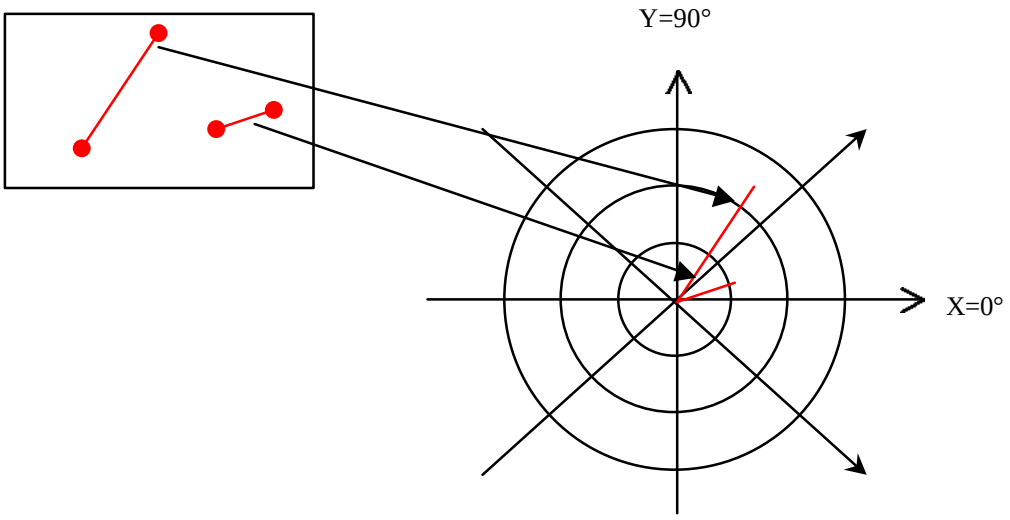
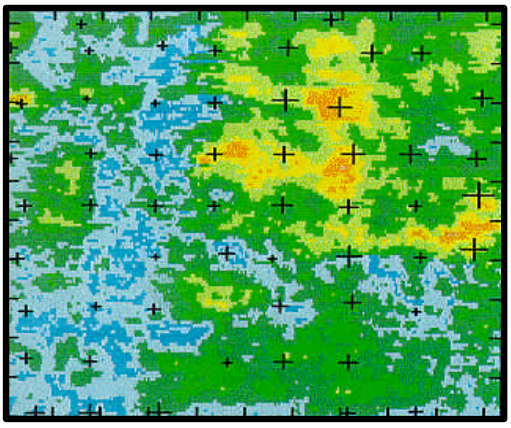


Illustration d'une anisotropie

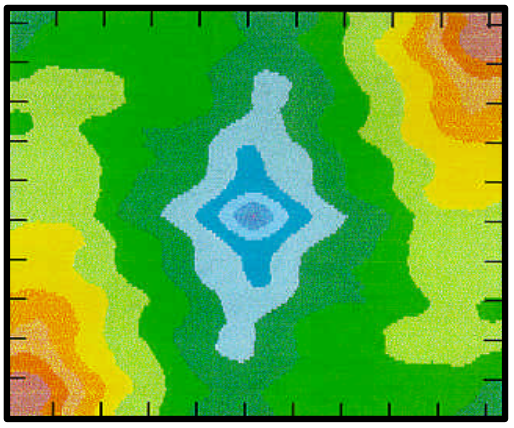
Cartographie du phénomène

0 10km



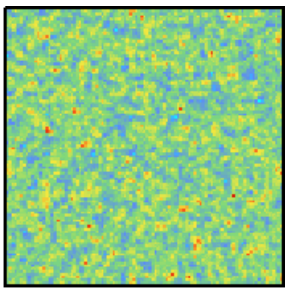
Carte du variogramme

-5 0 5

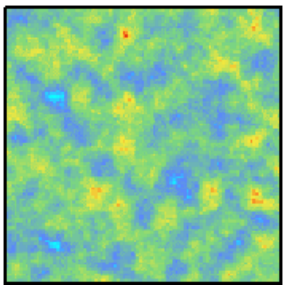


Influence de la portée

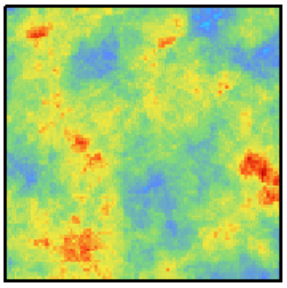
3 unités



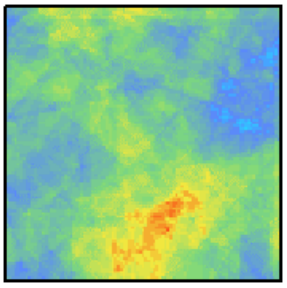
10 unités



25 unités

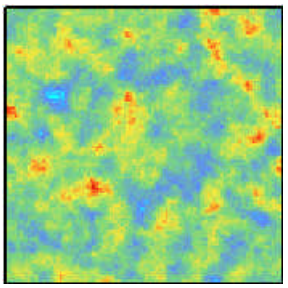


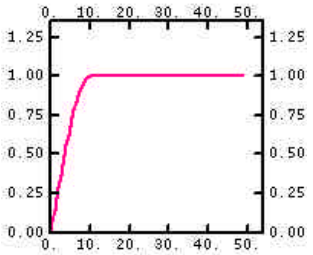
50 unités



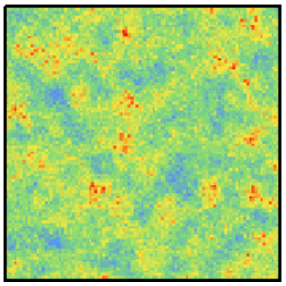
Influence de l'effet de pépite

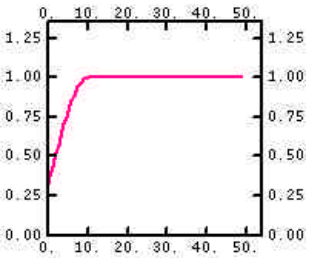
pas de pépite



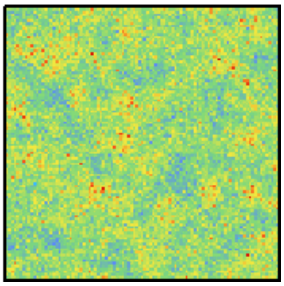


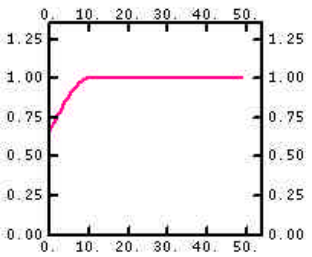
33% pépite



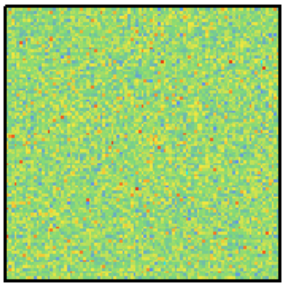


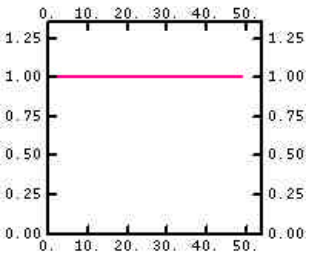
66% pépite





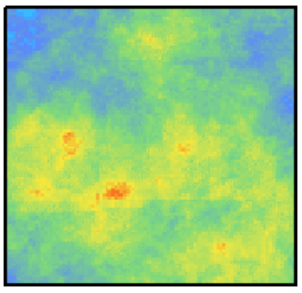
100% pépite

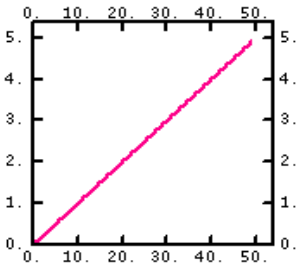




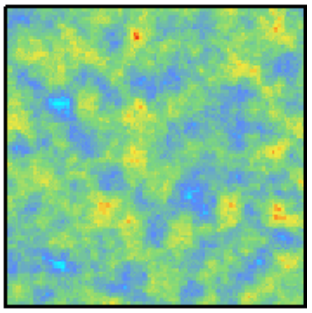
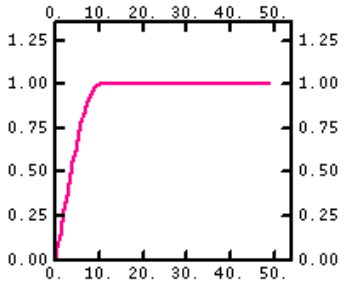
Exemple de variogrammes stationnaire

Linéaire

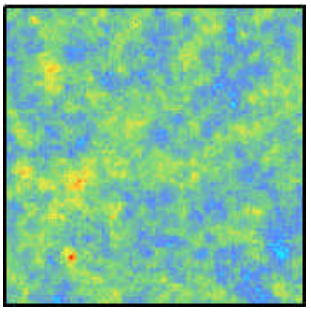
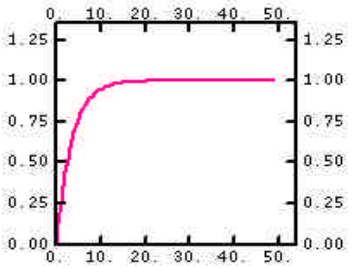




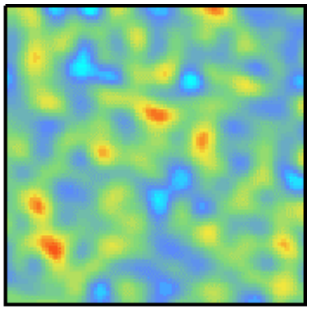
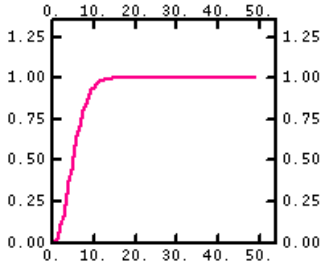
Sphérique
linéaire à l'origine ->bonne continuité
g(h) = C [(3h/2a) - ((h^3)/(2a^3))] + Co pour (h <a
g(h) = C + Co pour h>a
Co effet de pépite (palier C),
G(h) est la semi-variance, a est la portée
La tangente à l'origine (ligne passant par 2 à 3 points) coupe le palier à une distance de 2a/3. Si ce n'est pas le cas un modèle emboîté est peut être nécessaire

Exponentiel
Linéaire à l'origine, mais as*ymptotique*,
g(h) = C [1 - exp(-h/a)] + Co
L'approche graduelle du palier implique une portée réelle 1/3 de la portée.

Gaussien
Parabolique à l'origine (très bonne continuité) et attend le palier asymptotiquement
g(h) = C [1 - exp(-h2/a2)] + Co
La portée vaut 1/sqrt(3) la portée réelle

Cubique