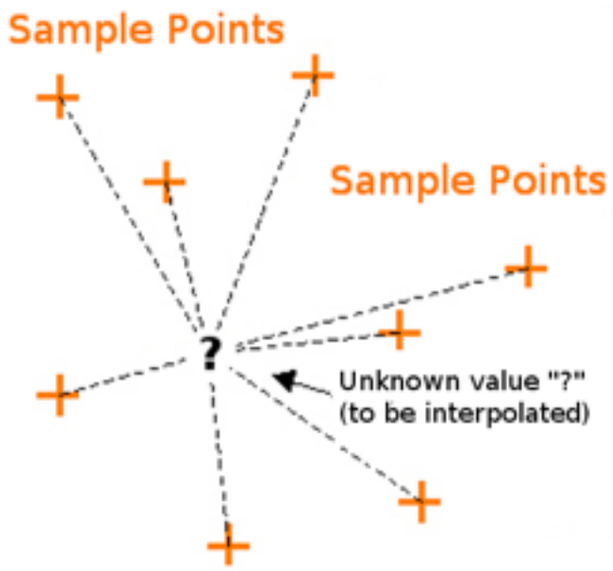
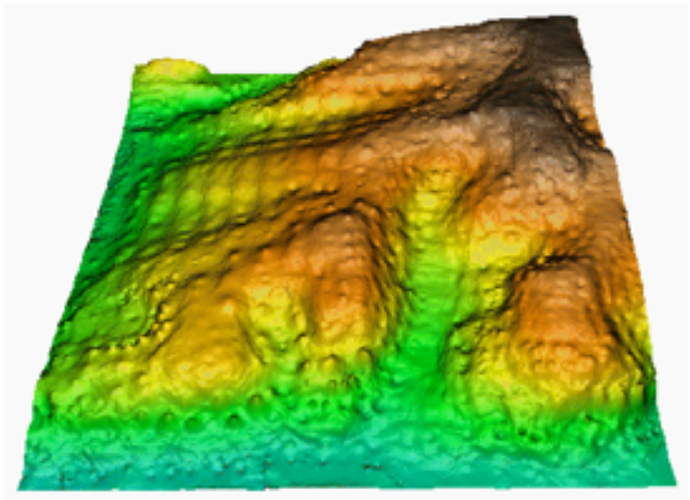
Introduction à l'interpolation statistique de données

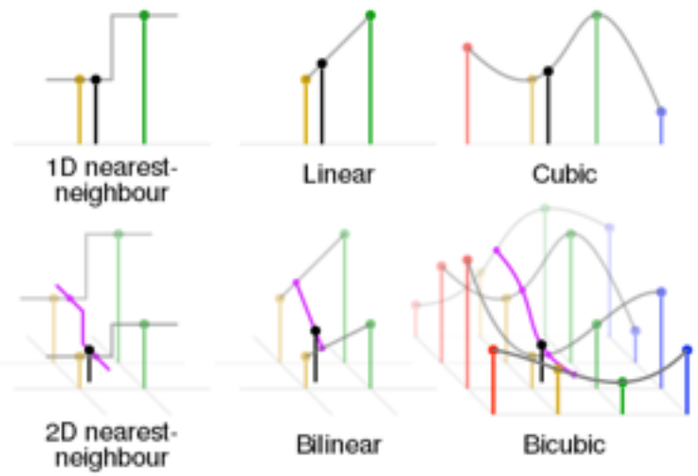
Présentation pour le Groupe transverse « Données » du LOPS, 29 Juin 2017, Brest

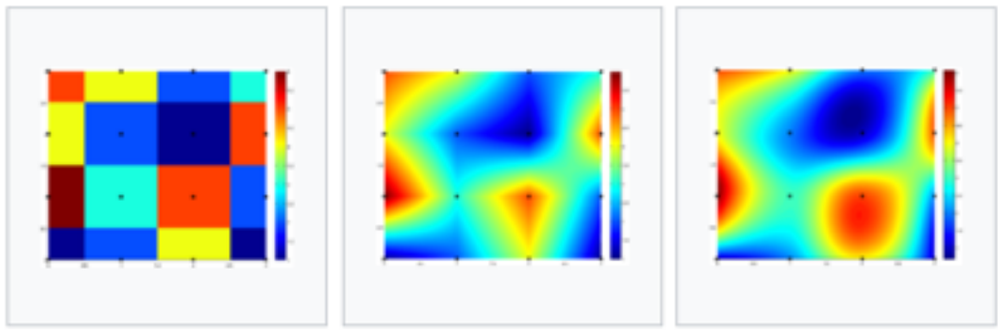
Le problème à résoudre

Estimer, prédire, expliquer une valeur d'un champ Z(**x**[,t]) en un point donné à partir d'un **échantillon** de valeurs connues en d'autres point « voisins » ?









Par proche voisin | Bilinéaire | Bicubique

La méthode subjective : à la main…

Des méthodes déterministes : interpolations « exactes » (pures maths, pas de stats, ni gestion des erreurs), pas de stats/incertitudes dessus, artefacts possibles (e.g. polynomial, splines cubiques)
exples : voisin (+ proche, naturel), linéaire, polynomial/spline, distance inverse, radial…

La méthode statistique : plus d'infos en input nécessaires (structure statistique du champ) et donc plus d'infos en sortie (e.g. incertitudes sur l'estimation, smooth ok, variances d'erreur etc), interpolation exacte OU smooth/fit possibles

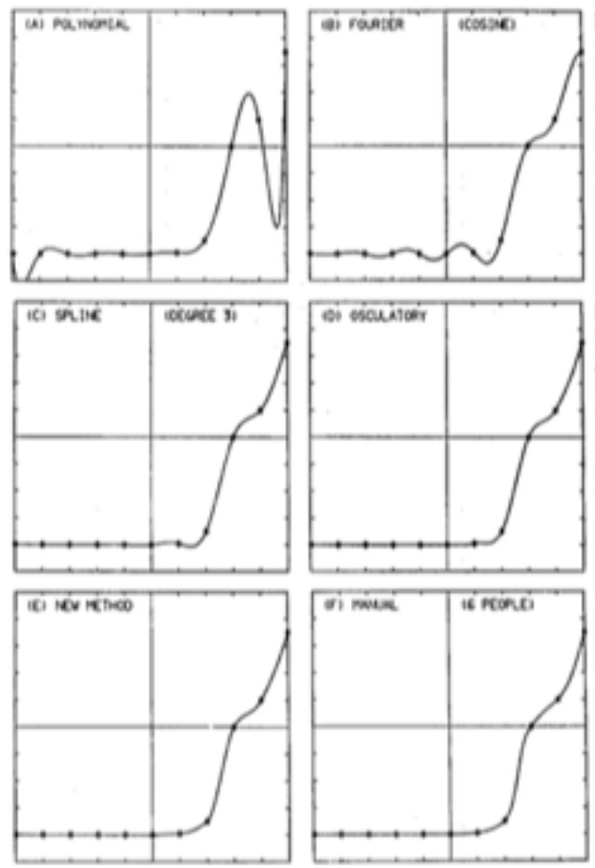


FIG. 1. Comparison of several methods of smooth curve fitting. (Encircled points are given data points.)

Historique rapide et contexte

Estimation OI (Atmosphère: Gandin 1965, Océan: Bretherton 1976), from Kolmogorov (1941), Wiener (1949)
=
Estimation par Kriging (Géologie, forages miniers : Krige 1951, Matheron 1963)
kriging : une regression multiple spatiale (Wackernagel 1995)

The use of the term kriging for this kind of method was proposed by Matheron after the original work by Krige [1951, 1966]

Krige, D. G., Two-dimensional weighted moving average trend surfaces for ore valuation, Symposium on mathematical statistics and computer applications in ore valuation, *J. S. Afr. Inst. Mining Met.*, *67*, 13–38, 1966.

Matheron, G., Random functions and their applications in geology, in *Geostatistics*, edited by D. F. Merriam, pp. 79–87, Plenum, New York, 1970.

Différents noms : interp stat, optimal interp, kriging, wiener-kolmogorov prediction, objective analysis, gauss-markov method…

MAIS tous basé sur le même principe de base : un modèle de régression linéaire estimé par les moindres carrés (car c'est le blue, cf th. Gauss markov, base de l'interp stat) —> **Méthode de régression linéaire spatio(-temporelle) d'estimation/prédiction basée sur l'autocorrelation (variables régionalisées)**

Le modèle de régression linéaire : rappels

Objectif : A partir d'un échantillon de n données/obs sur 2 variables (Xi, Yi, i=1,n), établir un modèle de relation linéaire entre ces 2 variables e.g. pour Galton (1886, inventeur de l'expression régression) entre la taille des fils et la taille des pères

—> **y = f(x,(a,b)) = ax + b** (y var expliquée, x var explicative, a,b parametres à déterminer)

Pour trouver cette droite on se base sur **l'estimateur par les moindres carrés** (on verra que c'est le meilleur estimateur linéaire en effet, cf th. Gauss-Markov) qui vise à **minimiser** la somme des carrés des résidus/erreurs (SCR, en vert sur la figure) = une fonction S(a,b) :

$$SCR = SSE = \hat{\epsilon}'\hat{\epsilon} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$ avec yi-chapeau = a.xi+b

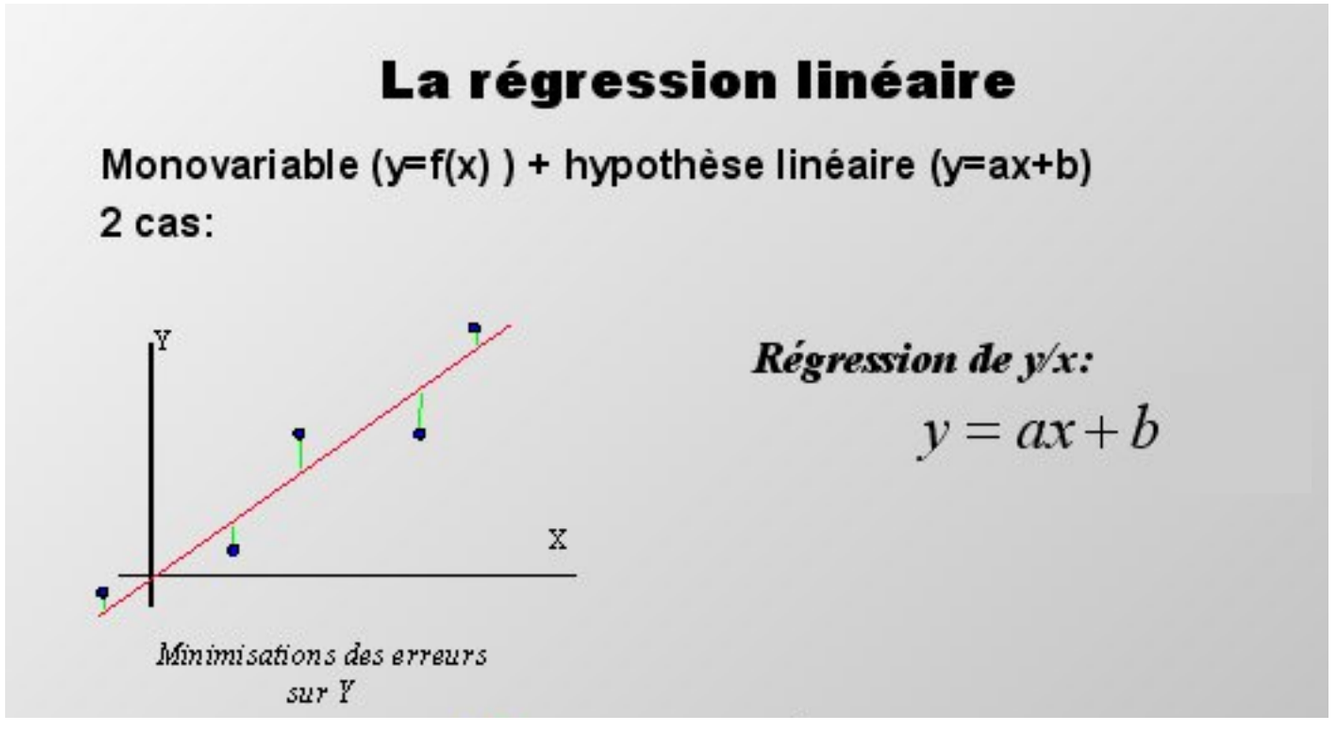
min(SCR) <—> annulation des dérivées de SCR(a,b), d'où le système d'équation linéaire à 2 inconnues a et b de solutions:

$$a = \frac{\mathrm{cov}(X,Y)}{Var(X)}$$

$$b = E(Y) - aE(X)$$

avec : $Cov(X,Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right] = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$

Var(*X*) = Cov(*X*, *X*).



Régression linéaire multiple et interpolation statistique

Même problème mais avec un échantillon de n données sur K variables (X1i, X2i,…,XKi, i=1,n), établir la relation entre ces K variables explicatives et la variable expliquée Y (Yi, i=1,n) : **Y = X.b en écriture matricielle** :
(a,b —> b1, b2, …bK, droite —> hyperplan dans espace de dimension K+1)

$$\Leftrightarrow \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{21} & . & . & x_{K1} \\ 1 & x_{22} & . & . & x_{K2} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{2N} & . & . & x_{KN} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ b_K \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_N \end{pmatrix}$$

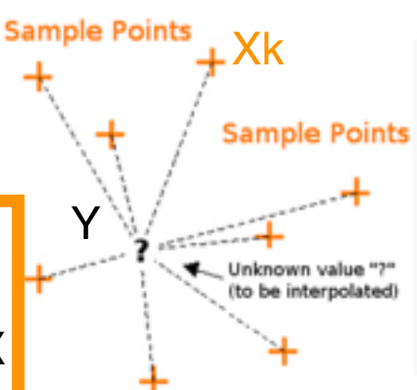
Minimisation de la SCR pour trouver **b** —> système linéaire à résoudre : **Coo.b = Coa <—> b = Coa . Coo^{-1}**

Avec :

Coo: matrice KxK de variance-covariance des X

$$Coo = \begin{pmatrix} Var(X_1) & Cov(X_1,X_2) & Cov(X_1,X_3) & \cdots & Cov(X_1,X_9) & Cov(X_1,X_{10}) \\ Cov(X_2,X_1) & Var(X_2) & Cov(X_2,X_3) & \cdots & Cov(X_2,X_9) & Cov(X_2,X_{10}) \\ Cov(X_3,X_1) & Cov(X_3,X_2) & Var(X_3) & \cdots & Cov(X_3,X_9) & Cov(X_3,X_{10}) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ Cov(X_9,X_1) & Cov(X_9,X_2) & Cov(X_9,X_3) & \cdots & Var(X_9) & Cov(X_9,X_{10}) \\ Cov(X_{10},X_1) & Cov(X_{10},X_2) & Cov(X_{10},X_3) & \cdots & Cov(X_{10},X_9) & Var(X_{10}) \end{pmatrix}$$

$$Coa = \begin{matrix} cov(Y,X1) \\ cov(Y,X2) \\ \dots \\ cov(Y,XK) \end{matrix}$$



Interpolation statistique: établir un lien entre la variable inconnue (dite expliquée) et les variables de l'échantillon connues (explicatives) —> si variables régionalisées, il existe une structure de covariance —> on a le système analogue à ci-dessus —> on peut prédire **b** et déduire **Y** à partir de **X**

Le théorème de Gauss-Markov

first to give the proof. [illegible]

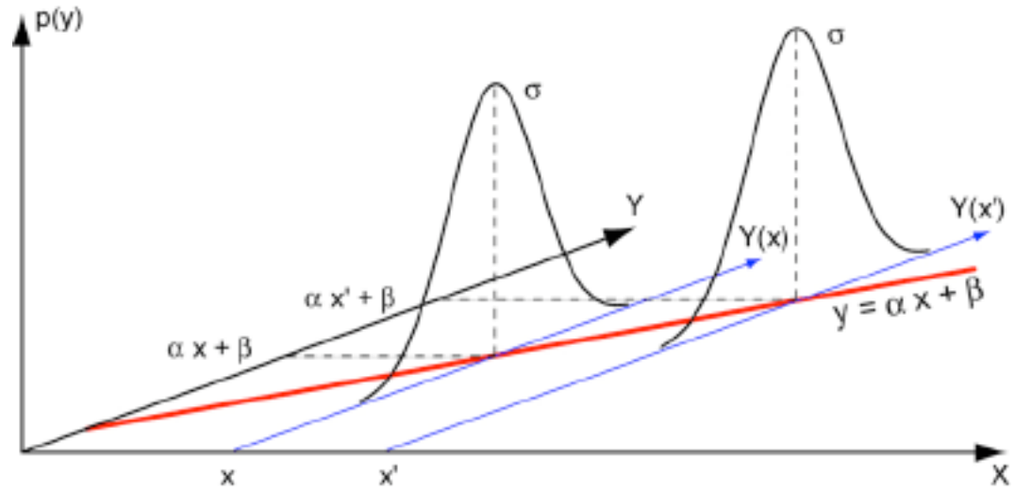
◆ **Theorem 6.2** If the general-linear-hypothesis model of full rank $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ is such that the following two conditions on the random vector $\mathbf{e}$ are met:

$$(1) \qquad E(\mathbf{e}) = \mathbf{0}$$

$$(2) \qquad E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}$$

the best (minimum-variance) linear (linear functions of the y_i) unbiased estimate of $\boldsymbol{\beta}$ is given by least squares; that is, $\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{X}'\mathbf{Y}$ is the best linear unbiased estimate of $\boldsymbol{\beta}$.

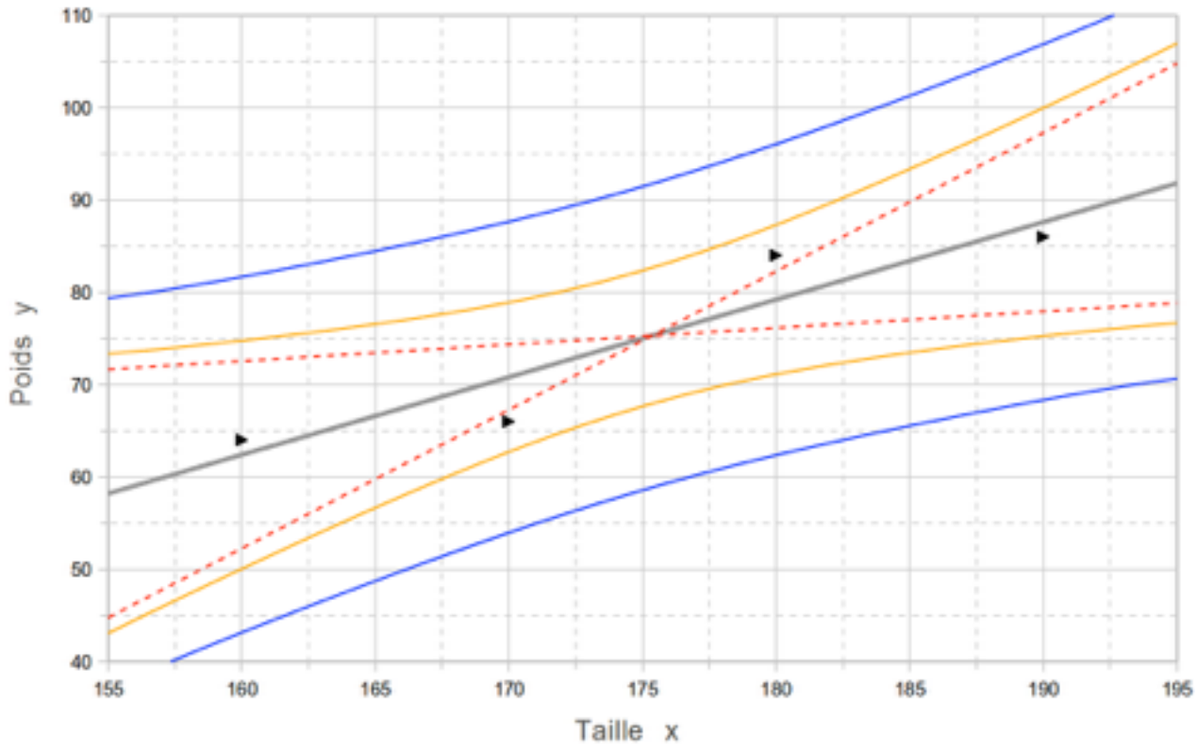
Proof: Let $\mathbf{A}$ be any $p \times n$ constant matrix and let $\boldsymbol{\beta}^* = \mathbf{A}\mathbf{Y}$; $\boldsymbol{\beta}^*$ is a general linear function of $\mathbf{Y}$, which we shall take as an estimate of $\boldsymbol{\beta}$. We must specify the elements of $\mathbf{A}$ so that $\boldsymbol{\beta}^*$ will be the best unbiased estimate of $\boldsymbol{\beta}$. Let $\mathbf{A} = \mathbf{S}^{-1}\mathbf{X}' + \mathbf{B}$. Since [illegible] we must find $\mathbf{B}$ in order to be able to specify $\mathbf{A}$.



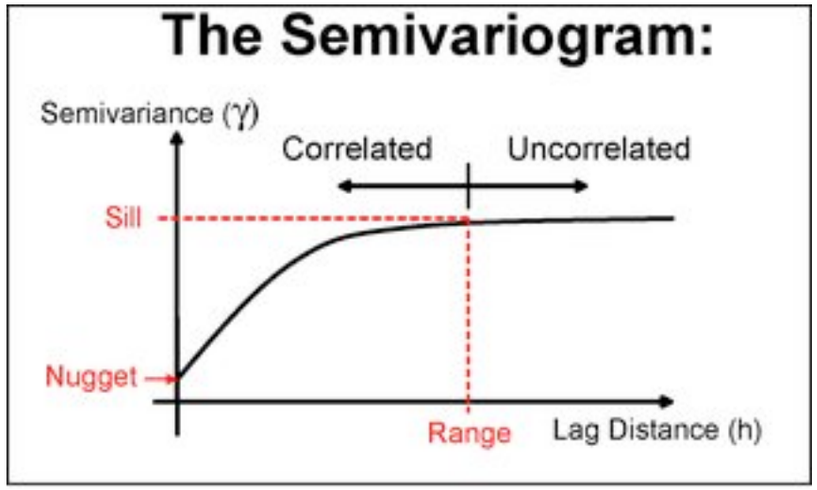
Les erreurs sont i.i.d (indépendantes et identiquement distribuées)

Least Square estimator = BLUE (Best Linear Unbiased Estimator)

Si en plus erreurs gaussiennes, alors estimation d'erreurs plus simples —>



L’analyse structurale : fonctions de covariance et variogrammes



Semivariance:

$$\gamma(h) = \sum_{i=1}^{N} \frac{[Z(x_i + h) - Z(x_i)]^2}{2N}$$

et on a :
gamma(h) = C(0) - C(h)
C(h) = gamma(inf) - gamma(h)

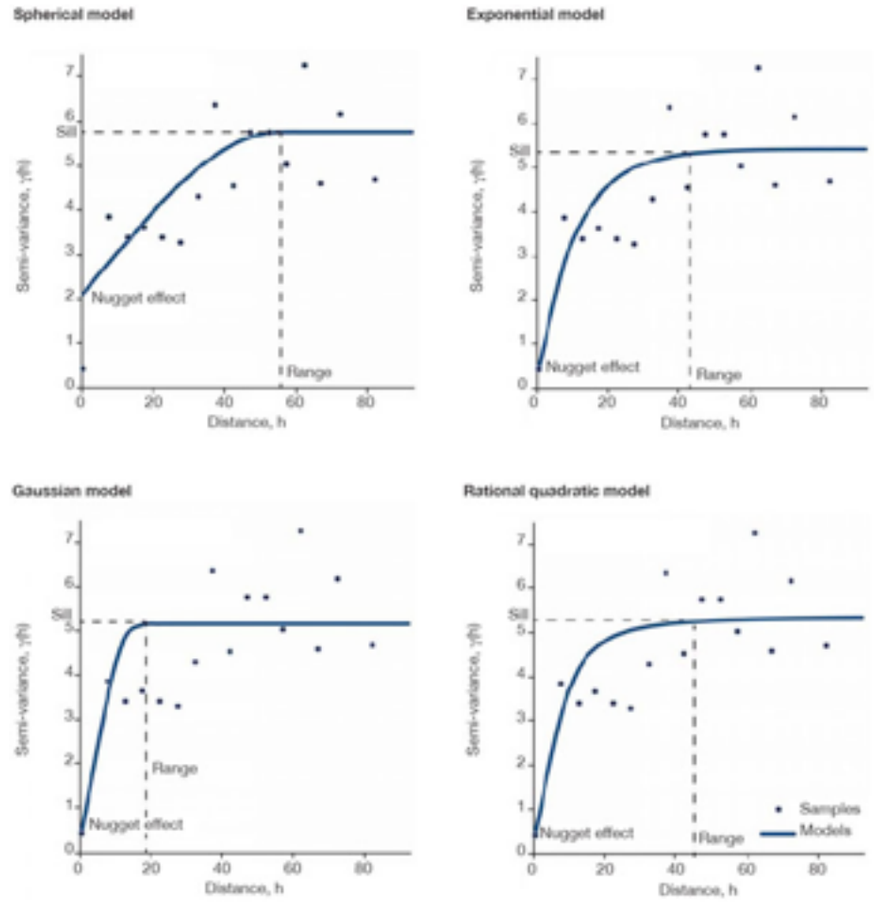
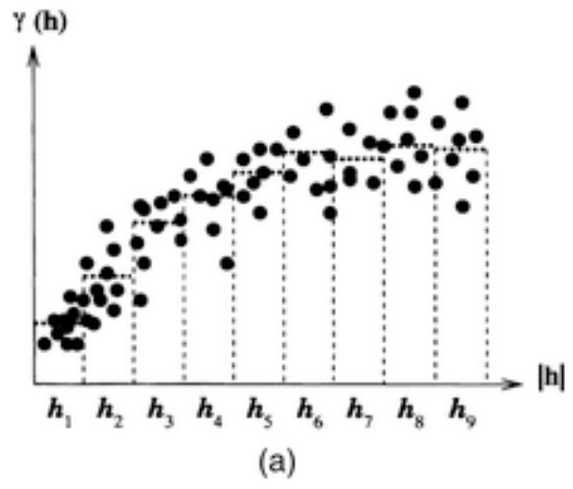
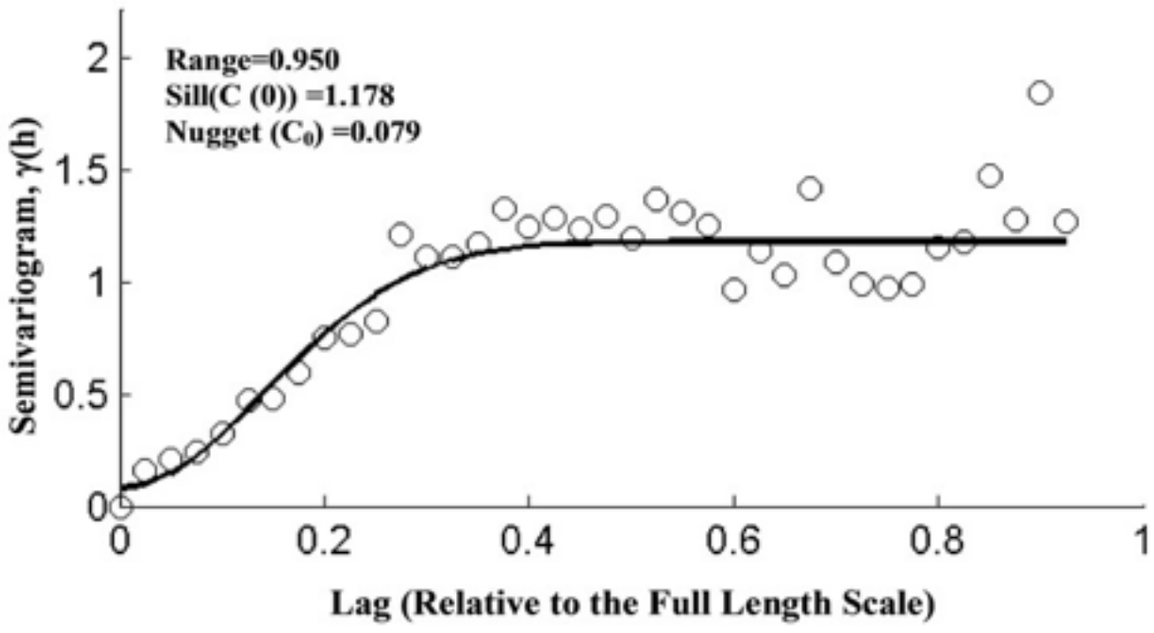
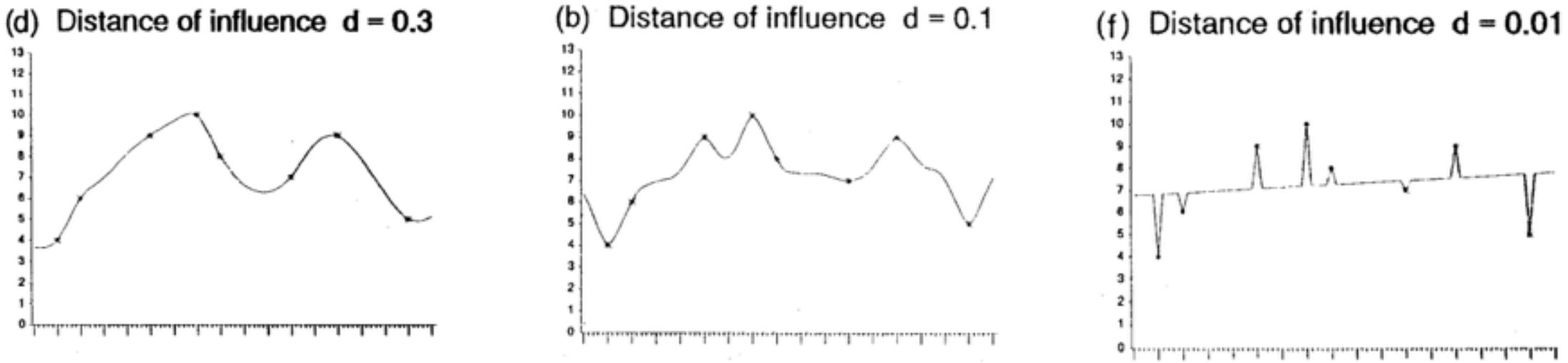


Figure 1. Example of an experimental semi-variogram with different permissible models fitted — *Exemple d'un semi-variogramme expérimental sur lequel différents modèles possibles sont ajustés.*





Le « first guess » ou « rappel au background state »



Rayon de correlation / distance d'influence (Trochu 1993)

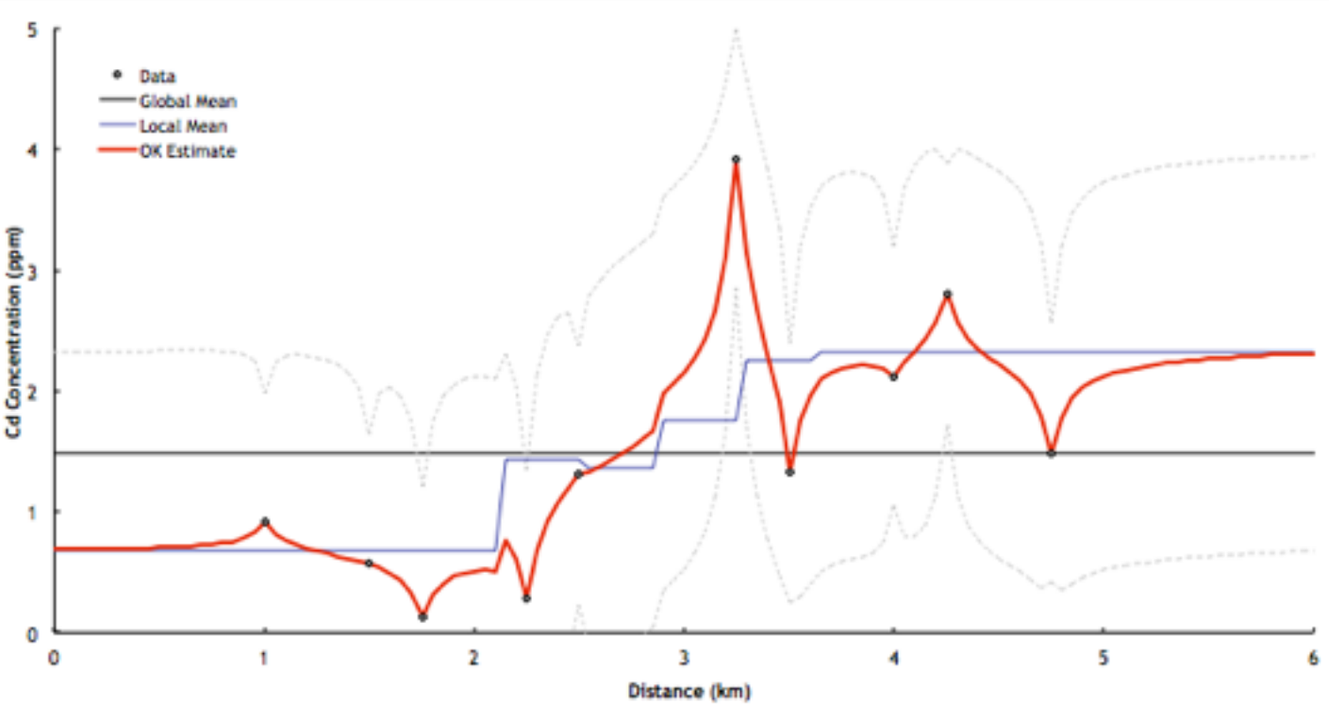
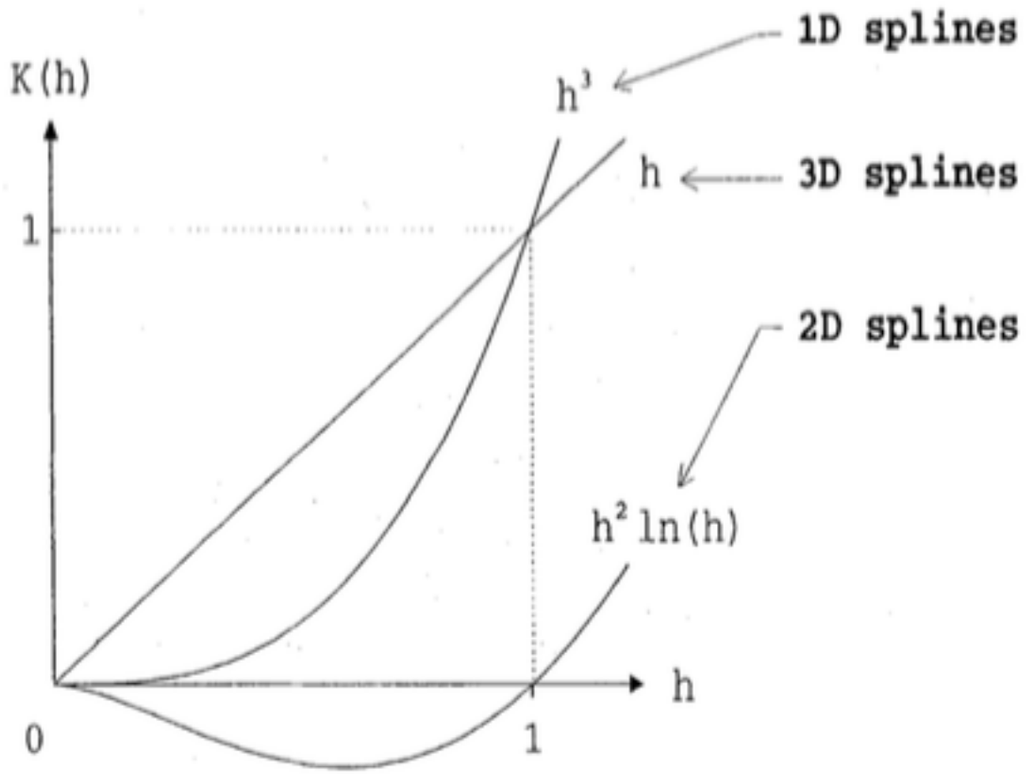
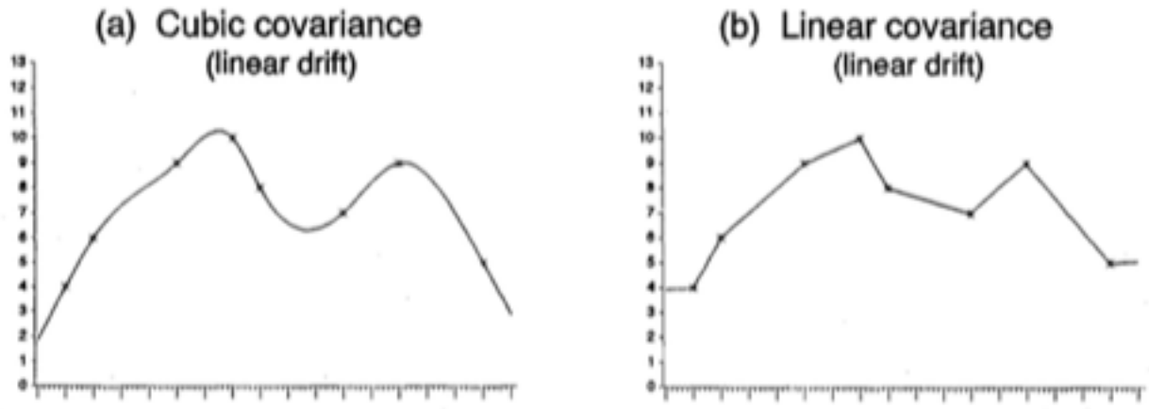


Fig 1: OK Estimates of the local trend and of Cd concentrations along a 6km transect. Estimates are based on a window of the n(u) = 5 ∀ u observations. The 95% CI of the estimate is indicated by the light gray broken lines which envelope the estimate and data points. The local mean is essentially a declustered mean of the 5 observations used to produce the estimate: Residual component is filtered by setting the data to unknown covariances to zero. Data sourced from Geostatistics for Natural Resources Evaluation (*Goovaerts, 1997*); see chapter 5 for details.

Lien avec l'interpolation « déterministe » et temps de calcul



kriging in the presence of a linear drift is equivalent to spline interpolation for the following generalized covariances



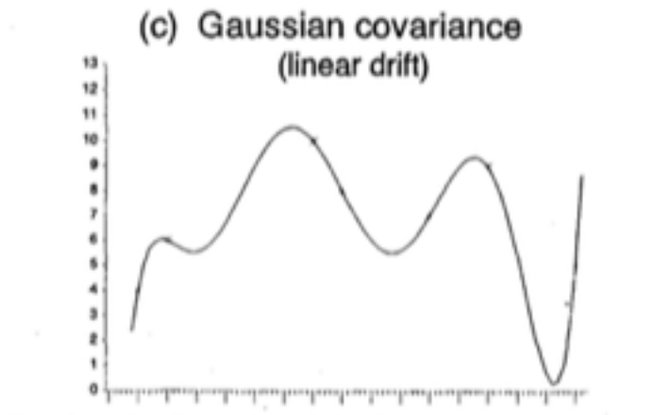
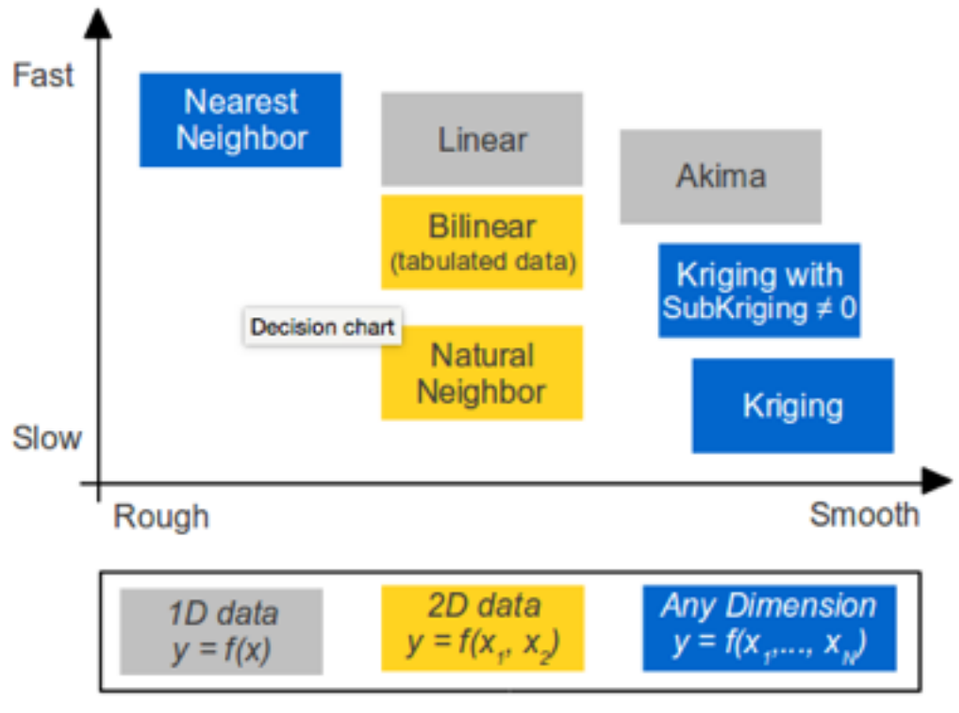
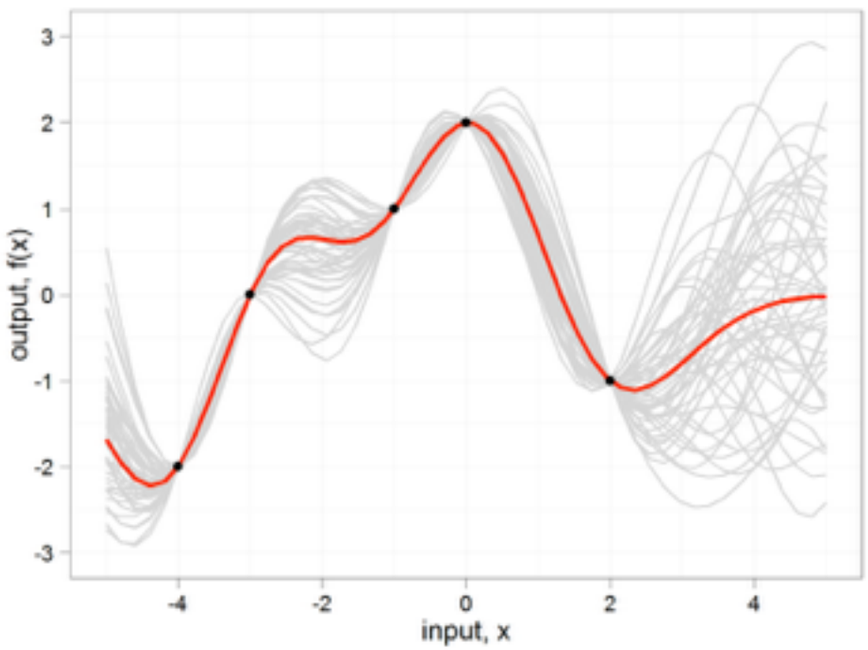
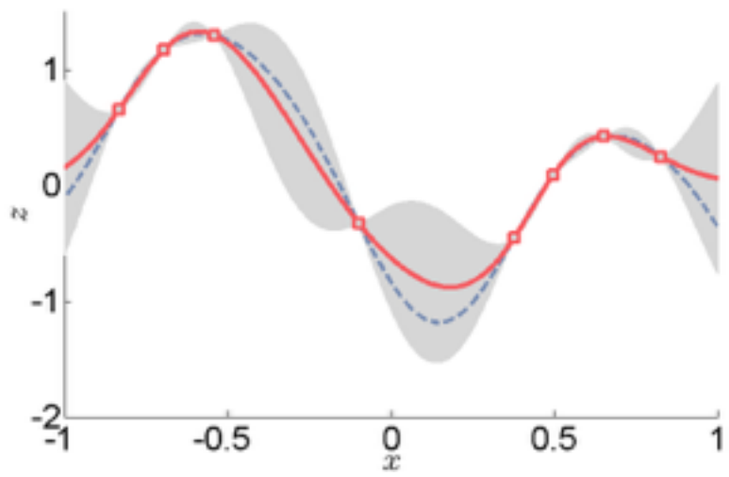


Fig. 3. Examples of kriging interpolations for various covariance functions.

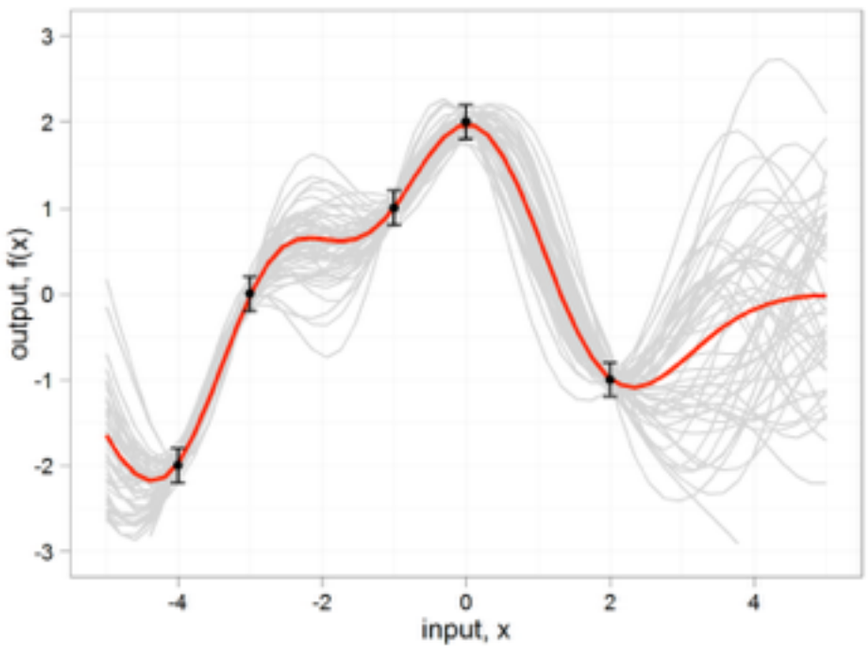


Exemples simples

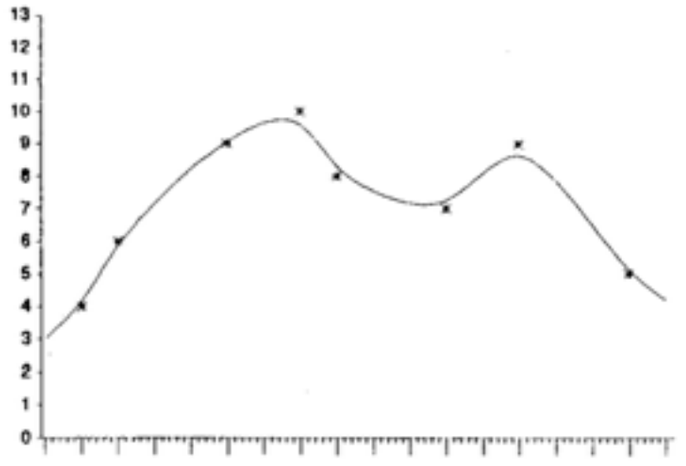




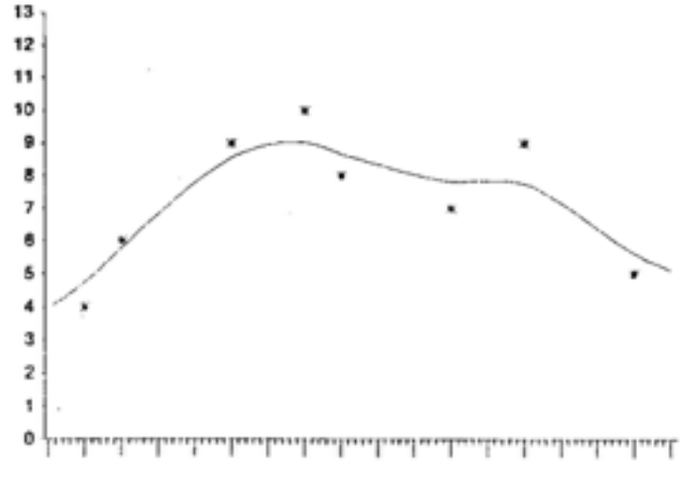
noise free vs noisy data with different gaussian variograms



Erreur P = 0.01



Erreur P = 0.1



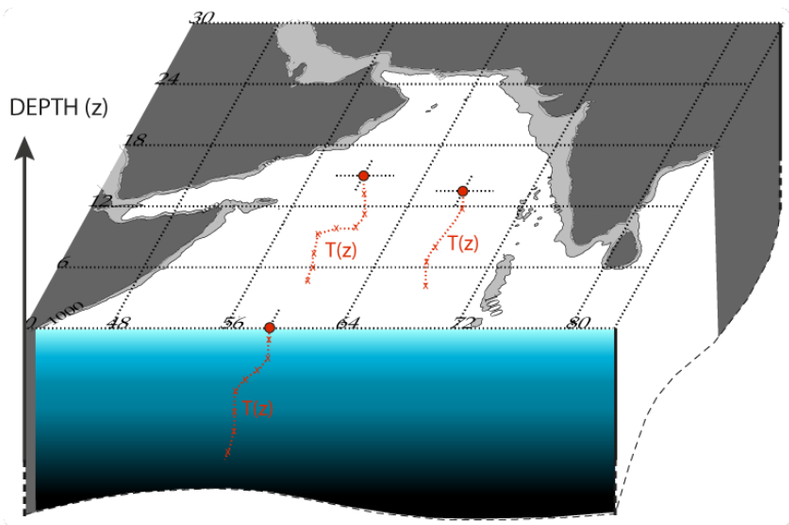
Exemple pratique : l'outil de génération des climatologies

TOOL : **G**enerator of **A**tlas from **I**n situ **O**bservations (GAIO, code IDL, ~ 60 routines, 25,000 lines)

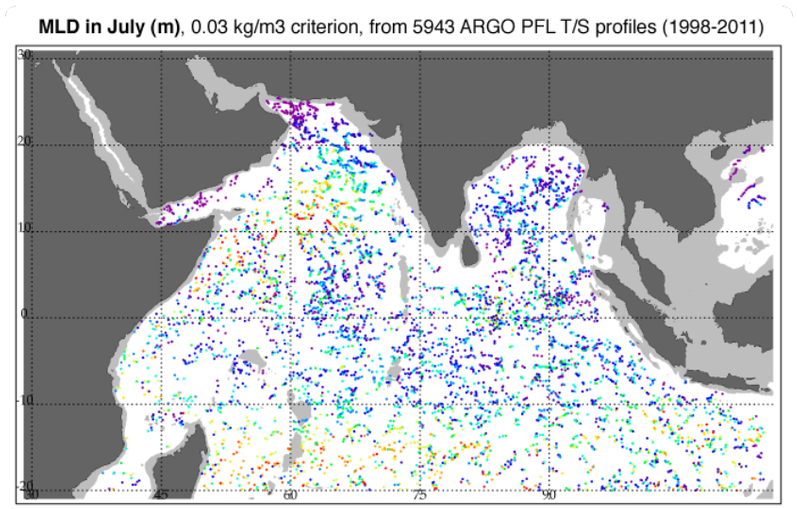
Working from **T/S/O2 profiles**, what I call **Level 1 datasets** (e.g. Argo) to get :

- **Level 2 datasets** : "advanced" ocean variables computed from each stations/profiles
  . Surface-measured Variables ..... T/S/O2 (10m depth)
  . MLD variables ..... DT02, DR003, DReqDT02, optim, BLT, TTD
  . Below-MLD measured vars ..... T/S/O2 10/25 m below MLDs
  . Below MLD Temperature-inversions ... Dinvmax, Tinvmax, Dt10
  . Isothermal depths D20, D26
  . CI-index ..... Cooling Inhibition index (Vincent et al. jgr 2012)
  . N2-average ..... Average of N2 over 400 m [TO INTEGRATE]
  . Heat Content ..... Heat content over 300 m
  . Sea Steric Height ..... SSH 1000/1500 m

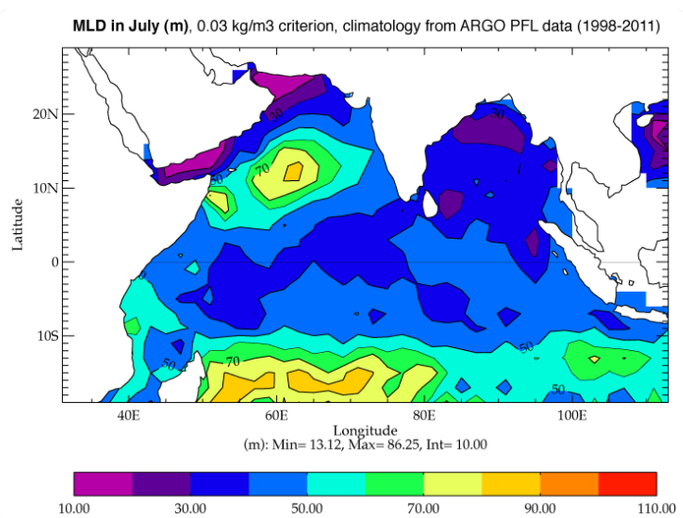
- **Level 3 datasets** : gridded fields(x,y,t) of those variables (Ordinary Kriging, cov. expon, Rcorr≈800km)



L1



L2

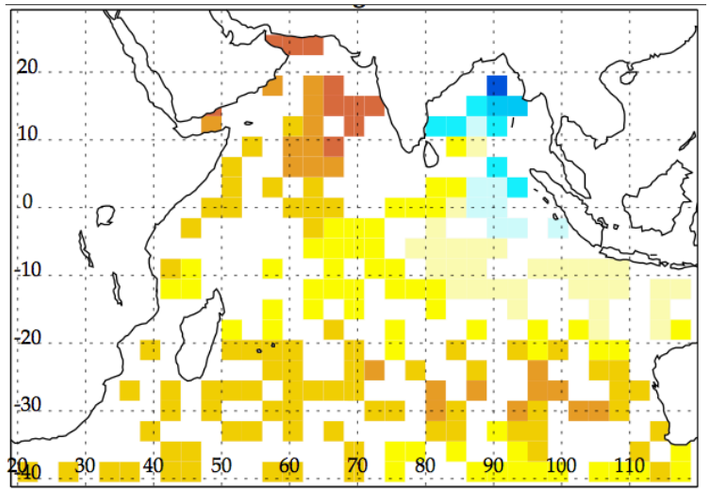


L3

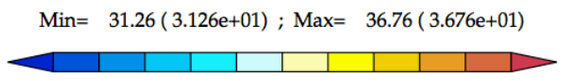
www.ifremer.fr/cerweb/deboyer/mld (Mise à jour site et données prévue pour 2017)

Reconstruction de champs de MLD (m) à partir d'Argo —> INTRASAISONNIER

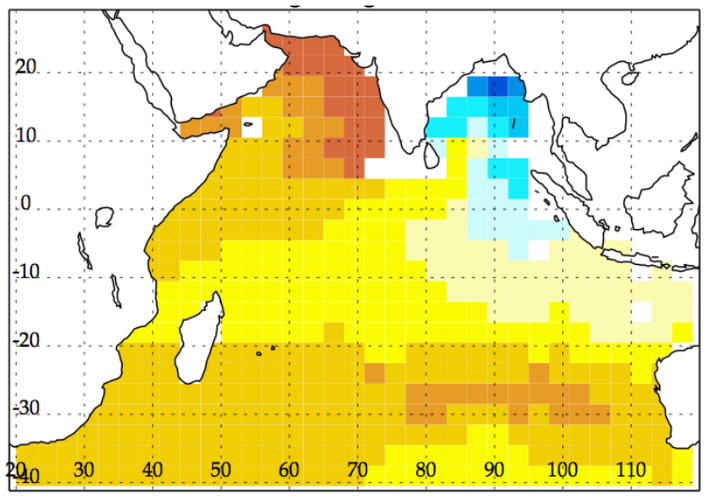
SSS 1-10 JUL 2007



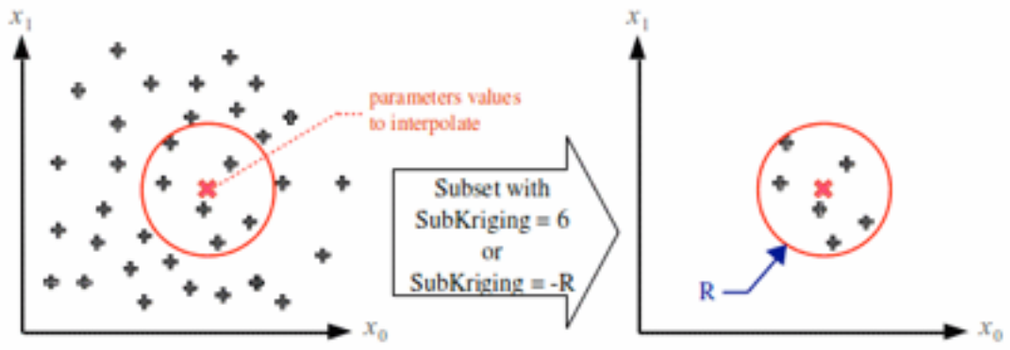
Min= 31.26 (3.126e+01) ; Max= 36.76 (3.676e+01)

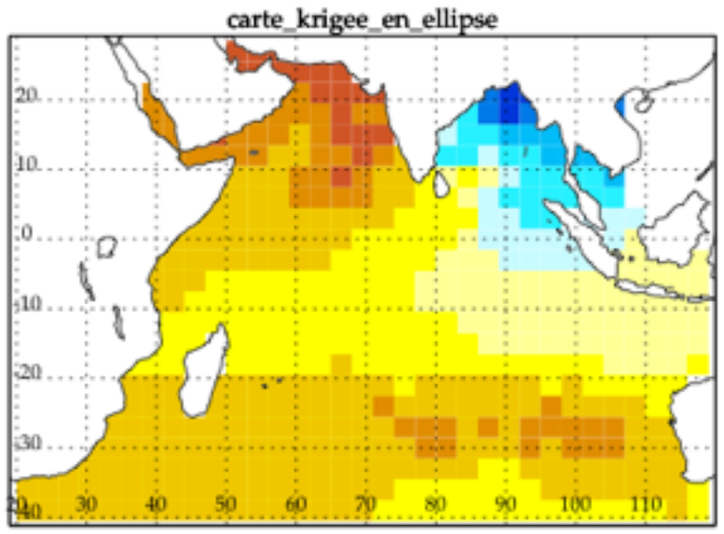


31 37

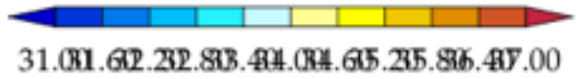


Min= 31.26 (3.126e+01) ; Max= 36.76 (3.676e+01)





Min= 31.26 (3.126e+01) ; Max= 36.76 (3.676e+01)



31.00

Keerthi et al., 2016

Statistical interpolation is a powerful and widely used technique for the objective analysis of atmospheric data. At the time of this writing, most numerical weather prediction centers use variants of this method in their data assimilation cycles. Statistical interpolation has been widely studied, and many of its properties are now well understood. For these reasons, we devote two chapters to this subject. This chapter considers the univariate scalar problem, and Chapter 5 will examine multivariate analysis with implicit physical constraints.

The technique of statistical interpolation can be traced back to Kolmogorov (1941) and Wiener (1949), who applied it to problems in various branches of science and engineering. It has often been referred to in the literature as *optimal interpolation*, a term apparently coined by Wiener. In practice, the technique is rarely optimal so the more appropriate term *statistical interpolation* is used in this book. In the oceanographic literature, the method is called the Gauss–Markov method (Section 13.8).

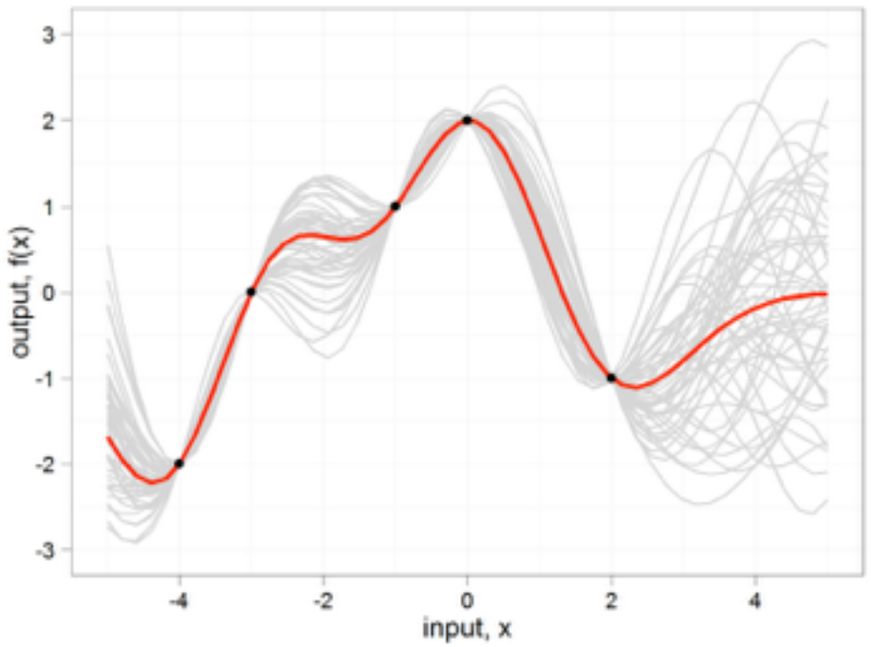
In the atmospheric sciences, attempts to use statistical interpolation for objective analysis date back to the 1950s. Variants of this technique were studied in the West by Eliassen (1954), Kruger (1964), Eddy (1964, 1967), and Petersen and Middleton (1964). In the Soviet Union, the development of statistical interpolation techniques for the objective analysis of atmospheric data was pursued more vigorously. In 1963, L. S. Gandin published a textbook that appeared in English translation in 1965 as *Objective Analysis of Meteorological Fields*. This book had an enormous influence on the subsequent development of objective analysis techniques in both the Soviet Union and the West. The practical use of this technique awaited the development of adequate computer power, but by the mid-1970s it was being used operationally in Canada, and most of the major western meteorological services followed suit shortly thereafter.

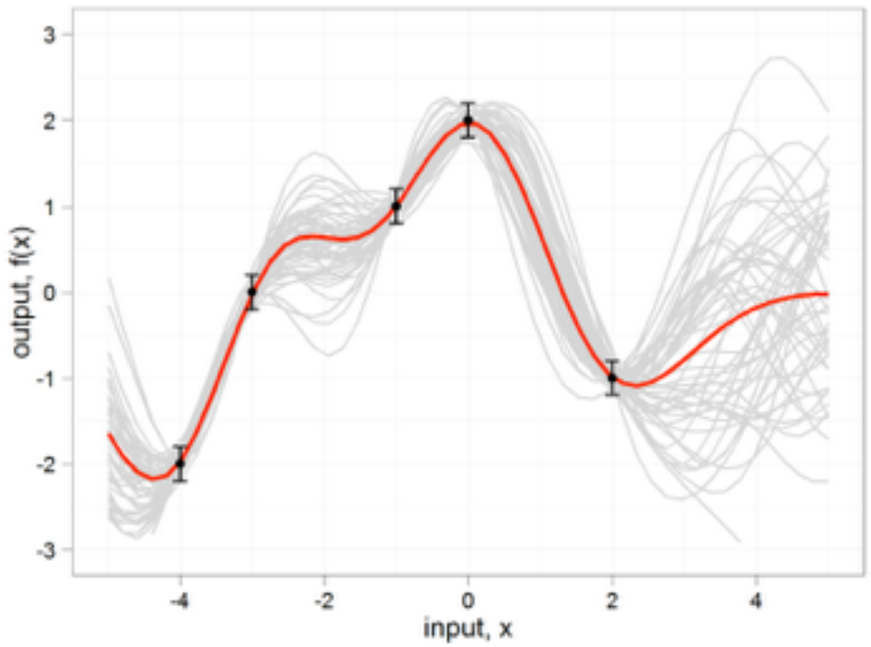
Statistical interpolation is a minimum variance method that is closely related to

the Kriging technique widely used in seismology. Before we derive the statistical interpolation algorithm, a brief discussion of minimum variance estimation is in order.

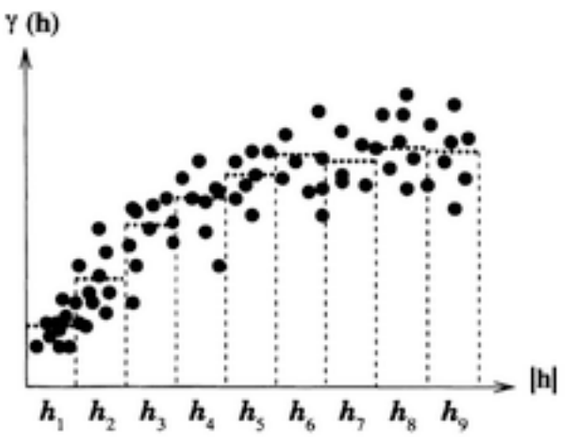
of the form (4.2.11). Interpolation of the form (4.2.9) using weights (4.2.10) is called *optimum interpolation*. However, these weights are optimal only if the observation and background error variances $\underline{\underline{B}}$, $\underline{\underline{Q}}$, and $\underline{\underline{B}}_i$ are *correct*.

If (4.2.10) is used to derive the weights, but the assumed observation and background error variances and correlations are *not* correct, then (4.2.11) is not strictly minimized. In this case, interpolation of the form (4.2.8) using these weights is not optimal and is called statistical interpolation. This distinction between optimal and statistical interpolation is important. The correct values of $\underline{\underline{B}}$, $\underline{\underline{Q}}$, and $\underline{\underline{B}}_i$ cannot be known precisely because they involve differences between observations and truth and background and truth; and since the truth is not known, $\underline{\underline{B}}$, $\underline{\underline{Q}}$, and $\underline{\underline{B}}_i$ cannot be known precisely and must be estimated. How this is done will be discussed in Section 4.3. Section 4.9 will consider the sensitivity of the statistical interpolation procedure to errors in the estimates of the statistical quantities $\underline{\underline{B}}$, $\underline{\underline{Q}}$, and $\underline{\underline{B}}_i$.

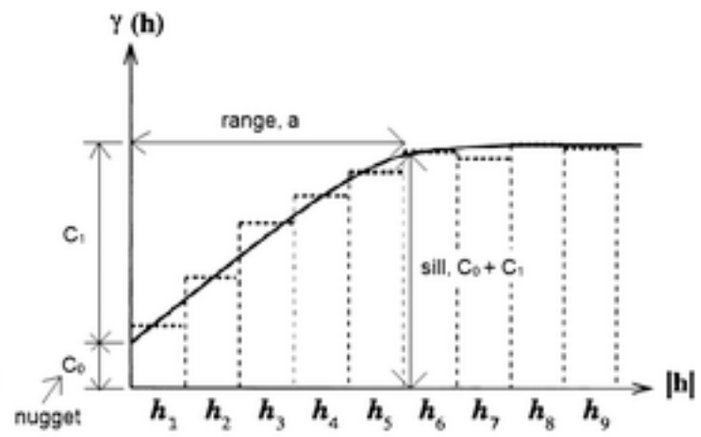




noise free vs noisy data

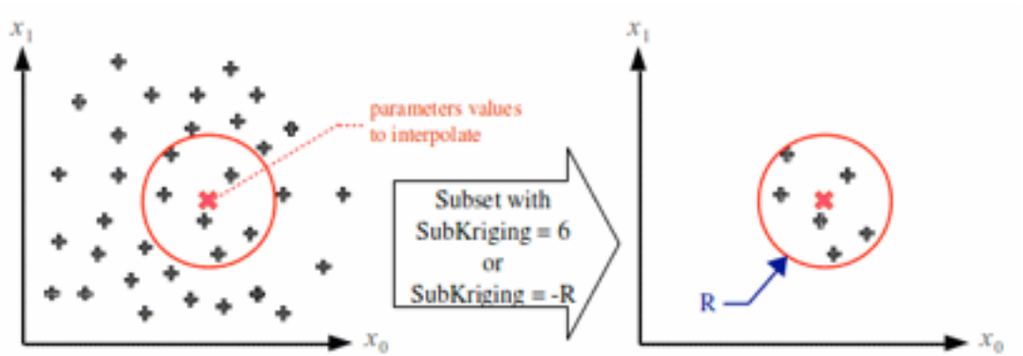


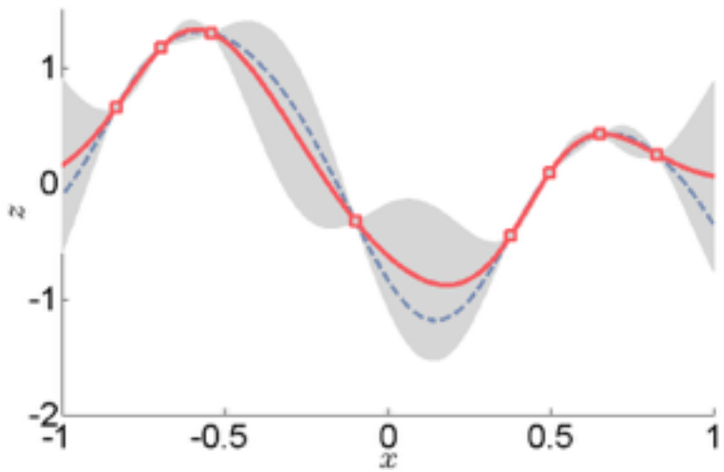
(a)



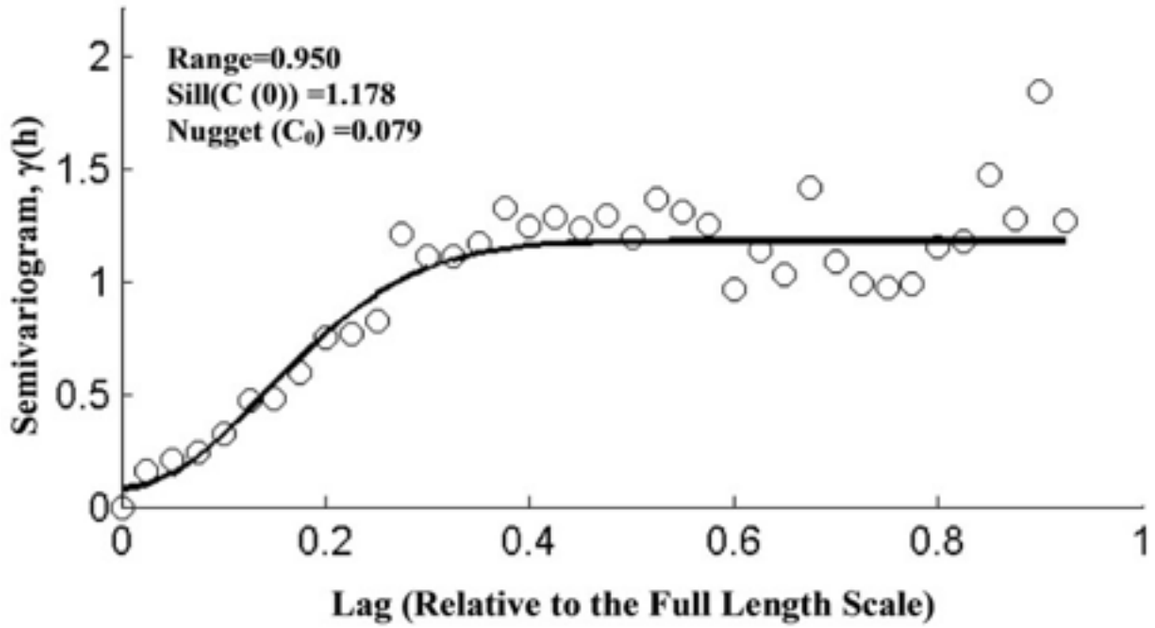
(b)

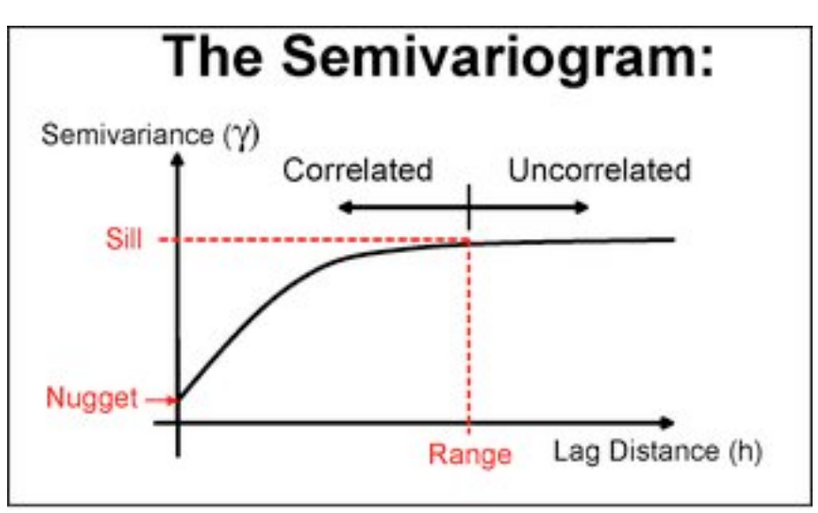
$$\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y)$$





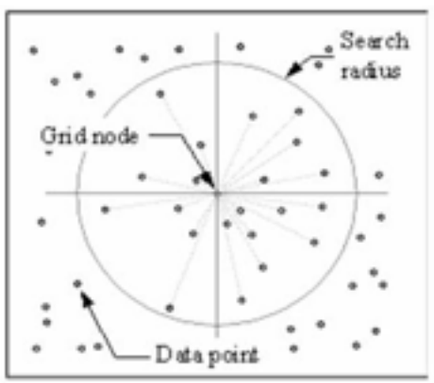
krig en rouge, avec spline cub en bleu





Semivariance:

$$\gamma(h) = \sum_{i=1}^{N} \frac{[Z(x_i + h) - Z(x_i)]^2}{2N}$$



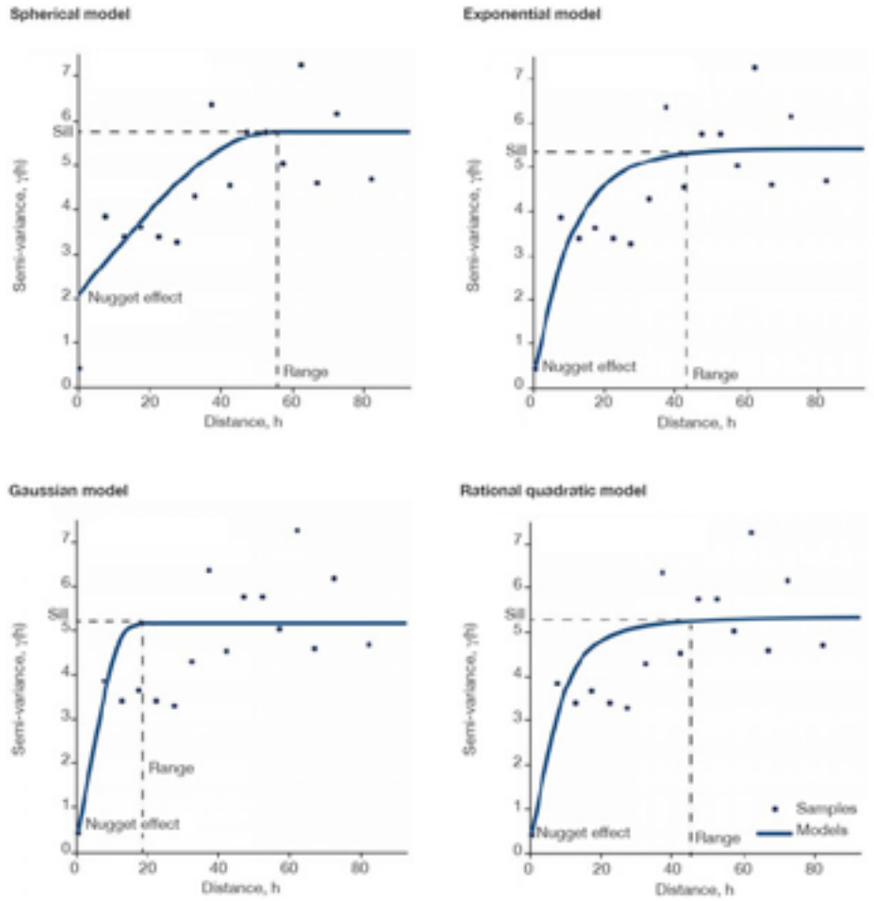


Figure 1. Example of an experimental semi-variogram with different permissible models fitted — *Exemple d'un semi-variogramme expérimental sur lequel différents modèles possibles sont ajustés.*

Interpolation statistique : Discussion

- Estimation OI (Atmosphère: Gandin 1965, Océan: Bretherton 1976) = Estimation par Kriging (Géologie, forages miniers : Krige 1951, Matheron 1963)
  => BLUE = least square estimator (th. Gauss Markov)

- kriging : une regression multiple spatiale (Wackernagel 1995)

first to give the proof

♦ **Theorem 6.2** If the general-linear-hypothesis model of full rank $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ is such that the following two conditions on the random vector $\mathbf{e}$ are met:

(1) $$E(\mathbf{e}) = \mathbf{0}$$

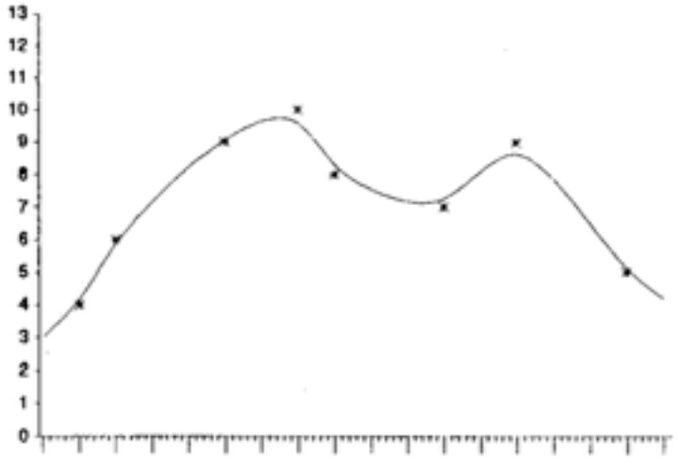
(2) $$E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}$$

the best (minimum-variance) linear (linear functions of the y_i) unbiased estimate of $\boldsymbol{\beta}$ is given by least squares; that is, $\hat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{X}'\mathbf{Y}$ is the best linear unbiased estimate of $\boldsymbol{\beta}$.

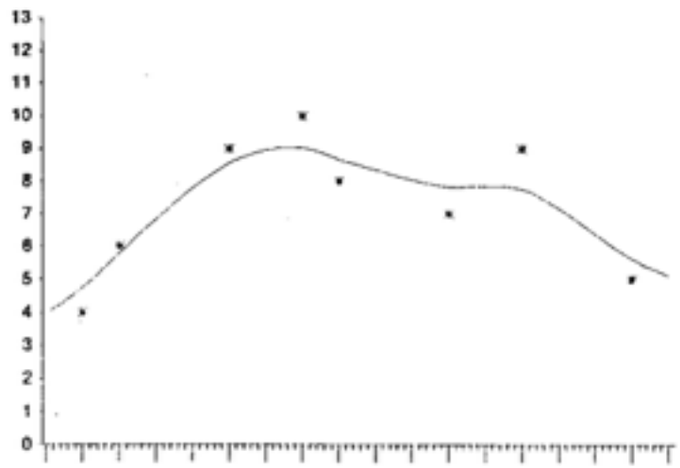
Proof: Let $\mathbf{A}$ be any $p \times n$ constant matrix and let $\boldsymbol{\beta}^* = \mathbf{A}\mathbf{Y}$; $\boldsymbol{\beta}^*$ is a general linear function of $\mathbf{Y}$, which we shall take as an estimate of $\boldsymbol{\beta}$. We must specify the elements of $\mathbf{A}$ so that $\boldsymbol{\beta}^*$ will be the best unbiased estimate of $\boldsymbol{\beta}$. Let $\mathbf{A} = \mathbf{S}^{-1}\mathbf{X}' + \mathbf{B}$. Since we must find $\mathbf{B}$ in order to be able to specify $\mathbf{A}$.

erreur OI ≈ nugget kriging

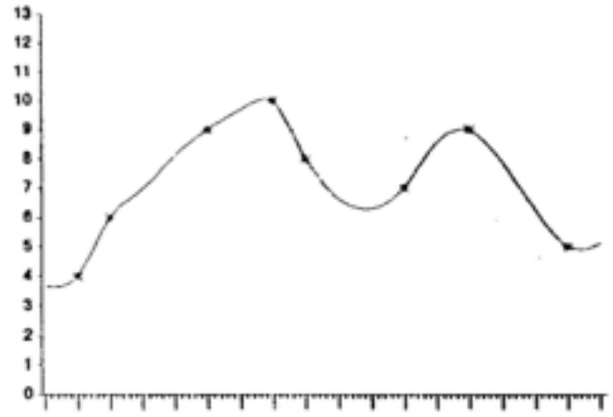
(a) Nugget effect for P = 0.01



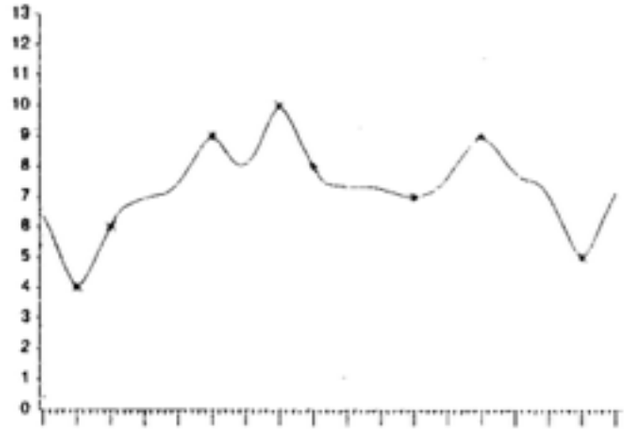
(b) Nugget effect for P = 0.1



(d) Distance of influence d = 0.3



(b) Distance of influence d = 0.1



Rayon de correlation / distance d'influence (Trochu 1993)

(f) Distance of influence d = 0.01