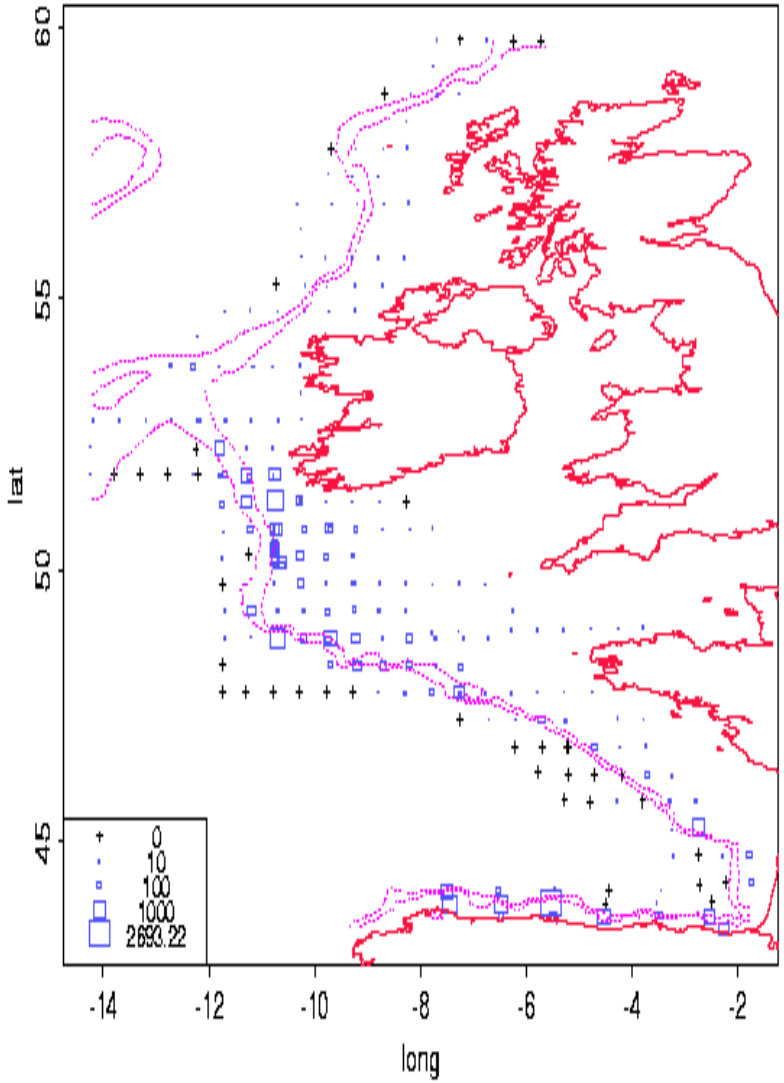
Initiation à la géostatistique

nicolas.bez@ird.fr

Estimation
Prédiction

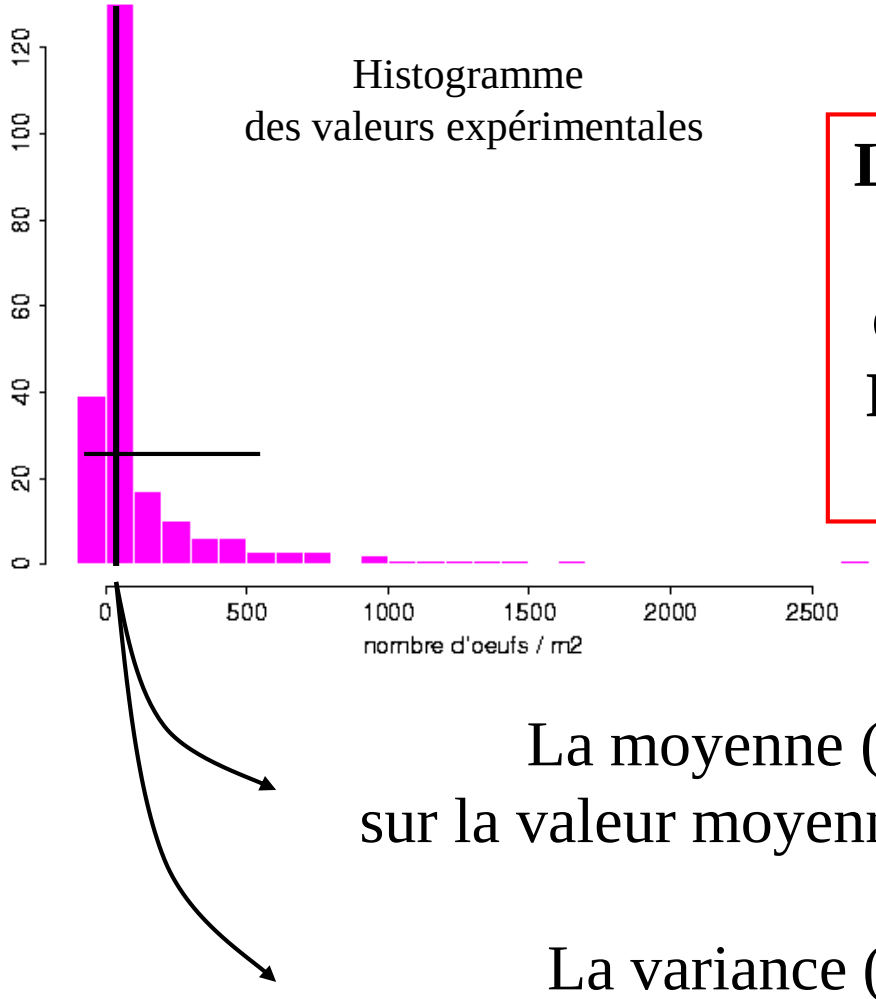
Limites de la statistique classique en situations réelles …



Un des objectifs principaux des campagnes à la mer :

Combien y-a-t-il d'individus dans l 'eau ?

Quantité (et distribution spatiale) d'œufs produits pendant une saison de ponte ?



La moyenne expérimentale *m* peut-elle être utilisée pour estimer la densité moyenne? Et si oui, quelle est la qualité de cette estimation ?

La moyenne (m) nous renseigne
sur la valeur moyenne des densités observées

La variance (s^2) nous renseigne
sur leur variabilité

La qualité d'une estimation est en général quantifiée
par la variance (de l'erreur) d'estimation.

Notion à
ne pas confondre avec la variance expérimentale
des données, même si cette dernière est utilisée pour son calcul.

$$\sigma^2_E = \frac{s^2}{N}$$

N observations x_i ➔ N random variables X_i ➔ $m^* = \frac{1}{N}\sum_{i=1}^{N} X_i$

Reminders

$var(aX) = a^2\, var(X)$

$var(X_1 + X_2) = var(X_1) + var(X_2) + 2cov(X_1, X_2)$

$$var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} var(X_i) + 2\sum_{i=1, j\neq i}^{N} cov(X_i, X_j)$$

Possible simplifications:

(1) X_i are mutually independent ➔ $cov(X_i, X_j) = 0$ ➔ $$var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} var(X_i)$$

(2) X_i are identically distributed ➔ same variance ➔ $$var\left(\sum_{i=1}^{N} X_i\right) = N\sigma^2 + 2\sum_{i=1, j\neq i}^{N} cov(X_i, X_j)$$

(1) + (2) X_i are i.i.d. ➔ $$var\left(\sum_{i=1}^{N} X_i\right) = N\sigma^2$$

Remark: « *identically distributed* » is reduced down to « same variance ». Having the same pdf is not required.

If X_i are i.i.d. then

$$var(m^*) = var\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2} var\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N}$$

$$\sigma^2{}_E = \frac{s^2}{N}$$

Variability of observations↑
⇒
Estimation variance ↑
⇒
Estimation quality ↓

Sampling size↑
⇒
Estimation variance ↓
⇒
Estimation quality ↑

Conditions de validité de cette formule

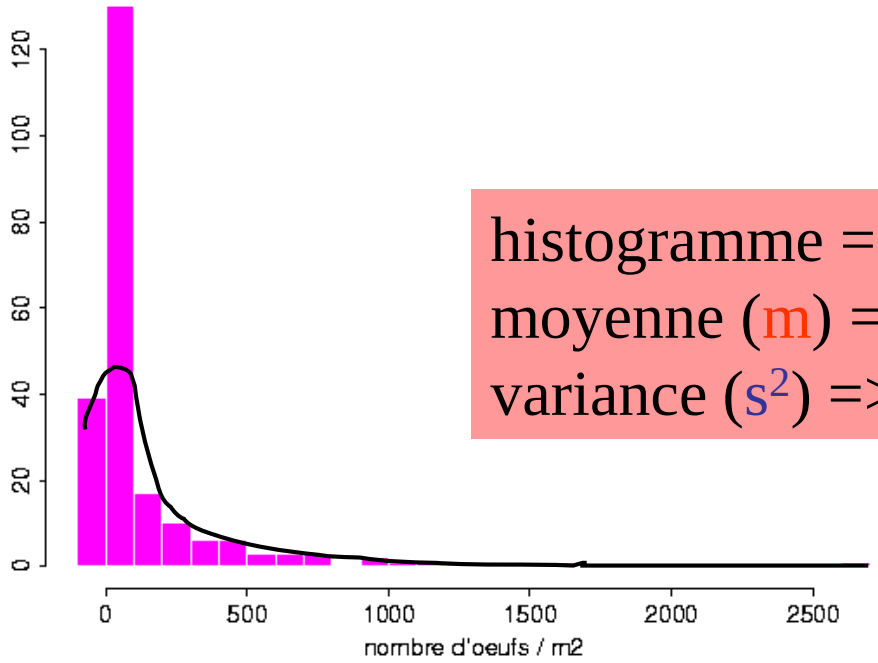
… « N échantillons
de N variables aléatoires
indépendantes
de même loi » ...

variable aléatoire indépendantes de même loi

Choix d'un modèle

Plus ou moins judicieux,
mais choix

Les valeurs prises par
la densité
peuvent être modélisées
par une **Variable Aléatoire**



histogramme => loi de probabilité
moyenne (m) => espérance mathématique (E)
variance (s^2) => variance (σ^2)

Identité en loi

On suppose que la densité de poissons a la même distribution en tous les points de l'espace.

==> homogénéité spatiale (stationnarité)

==> Recours implicite aux aspect spatiaux

Indépendance (statistique)

Permet de simplifier les formules et de rendre la démarche opératoire.
Permet d'éviter les redondances d'information

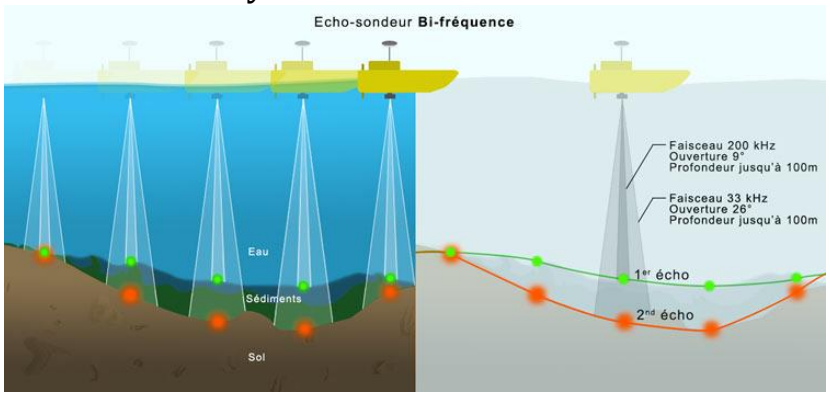
L'indépendance peut être assurée

Par un échantillonnage aléatoire pur

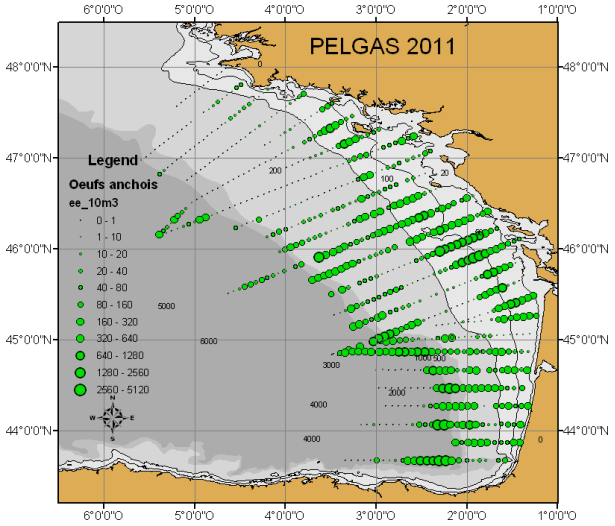
Les contre exemples sont nombreux…

Regular sampling designs

Acoustic surveys

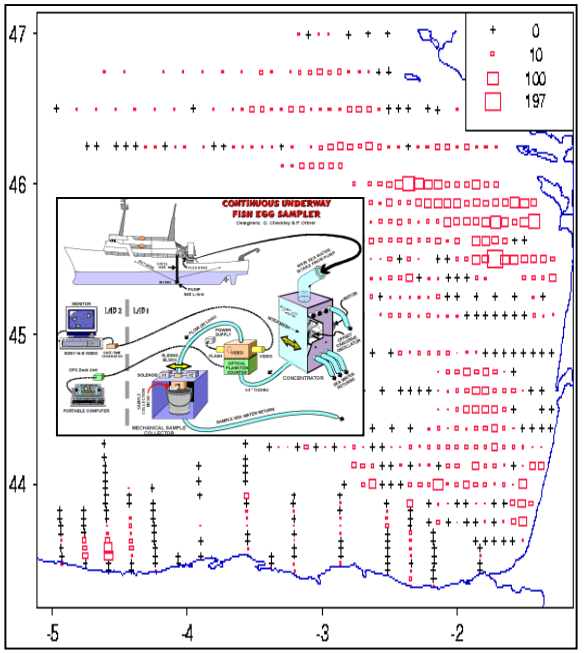


(source: Escadrone)



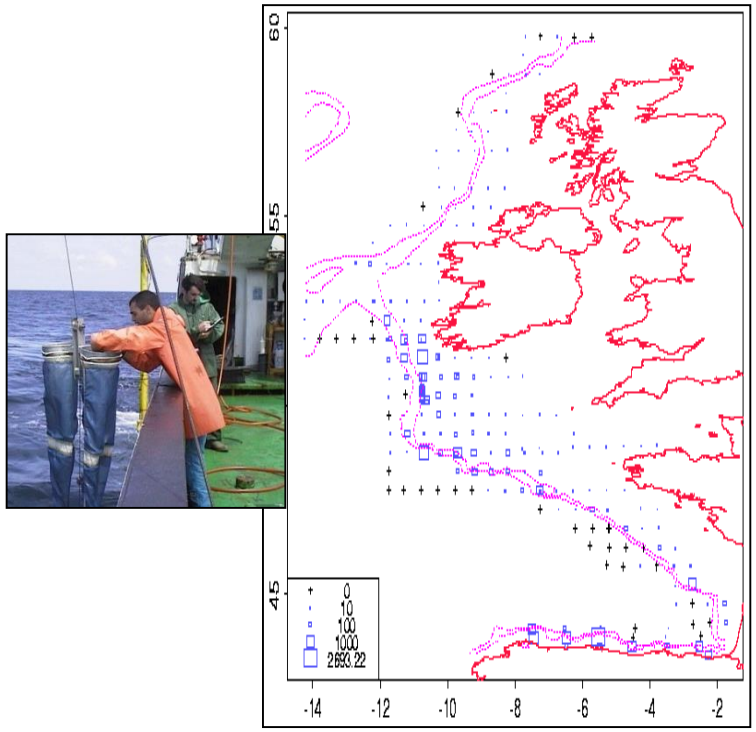
PelGas Survey
(source: Ifremer-France)

Continuous Underway Fish Egg Sampler (CUFES) surveys

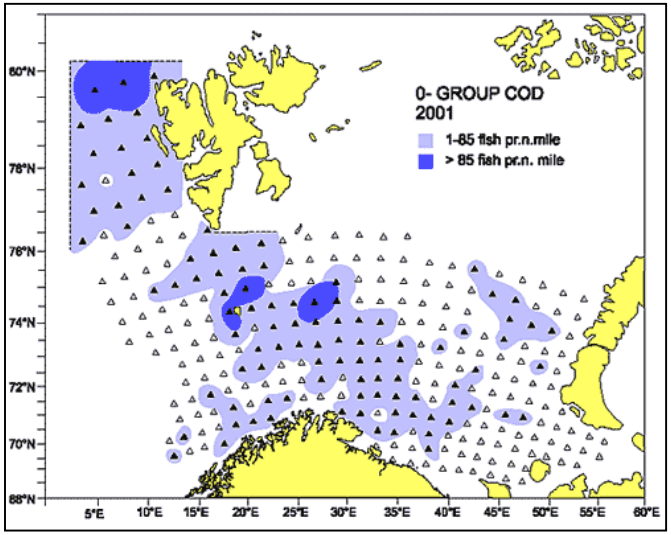


Anchovy eggs, BIOMAN Survey 1998
(source: AZTI, Spain)

Regular sampling designs



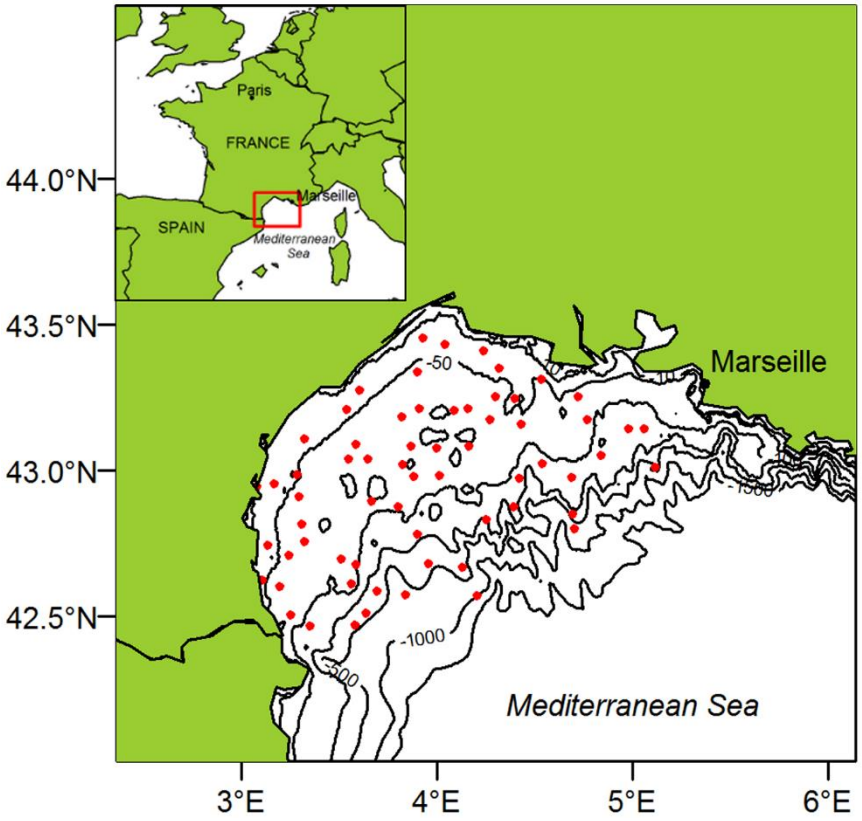
Triennal Eggs Surveys 1998 Mackerel eggs
(source: ICES)



Barents Sea Bottom trawl survey
(source: IMR-Norway)

Random stratified surveys

Several samples per strata



Bathymetric strata.
Random within a strata.
But, the full set of samples are not independent.

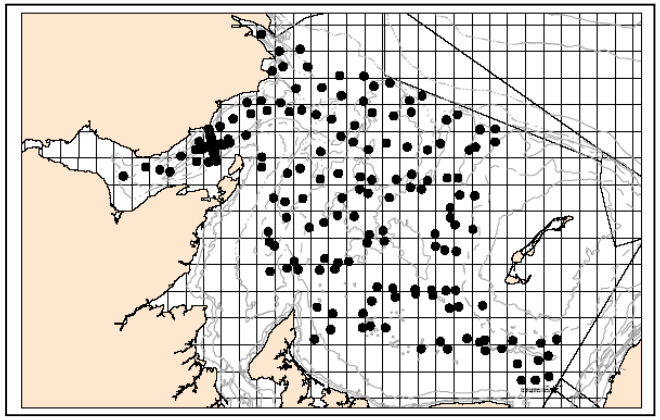
MEDITS surveys
(source:Ifremer)

Random stratified surveys

One sample per strata

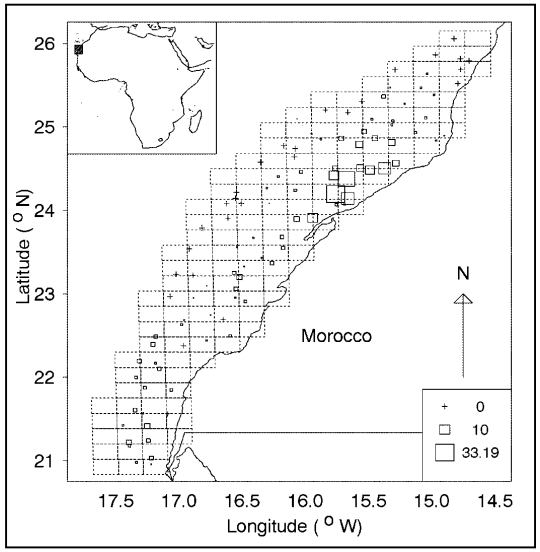
➔ The full set of samples are not independent.

What is the meaning of the variance of one point ?



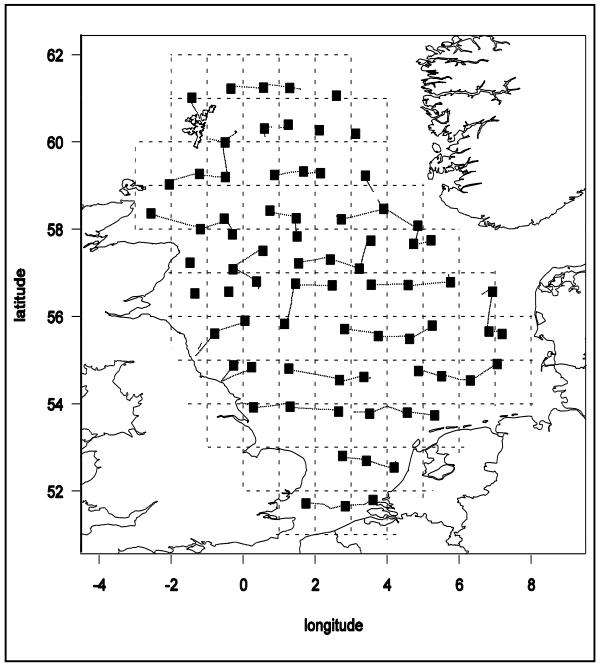
Snow crab trawl survey.

(source: Pêche et Ocean – Canada)



Cephalopod trawl survey

(source: INRH, Morroco)



IBTS surveys

(source:CIEM)

Indépendance
(statistique)

Permet de simplifier les formules et de les rendre opératoire.
Permet d'éviter les redondances d'information

L'indépendance peut être assurée

Par un échantillonnage *aléatoire* pur

Les contre exemples sont nombreux …

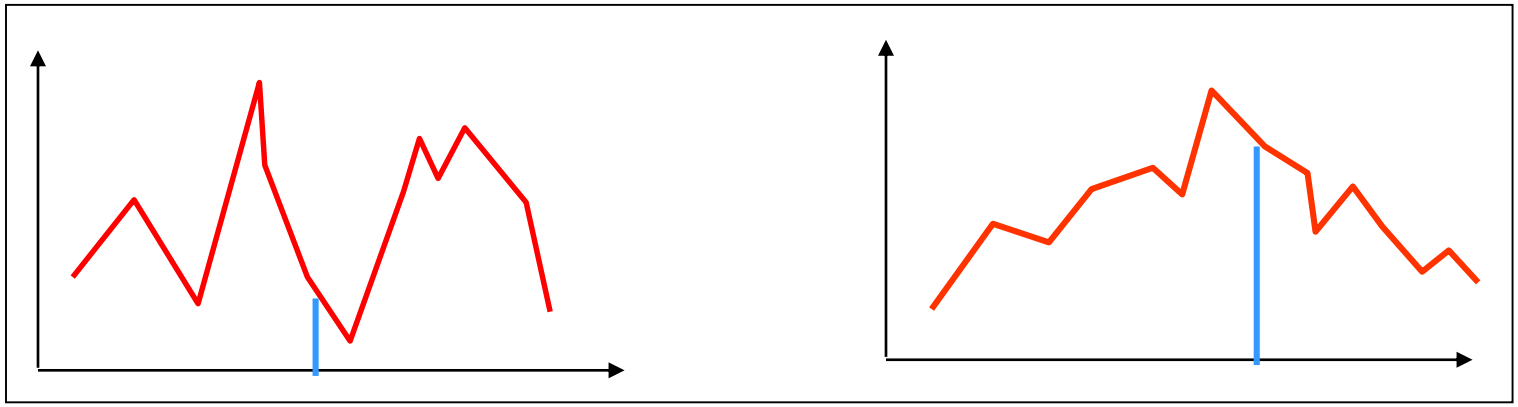
Par l'absence de structure spatiale du phénomène étudié :

Biologiquement rare ….

Profit possible de la présence d'une structuration spatiale

Pas de structure spatiale

Structure spatiale



Estimation d'un point inconnu

« à l'aveugle » :
basée sur N échantillons

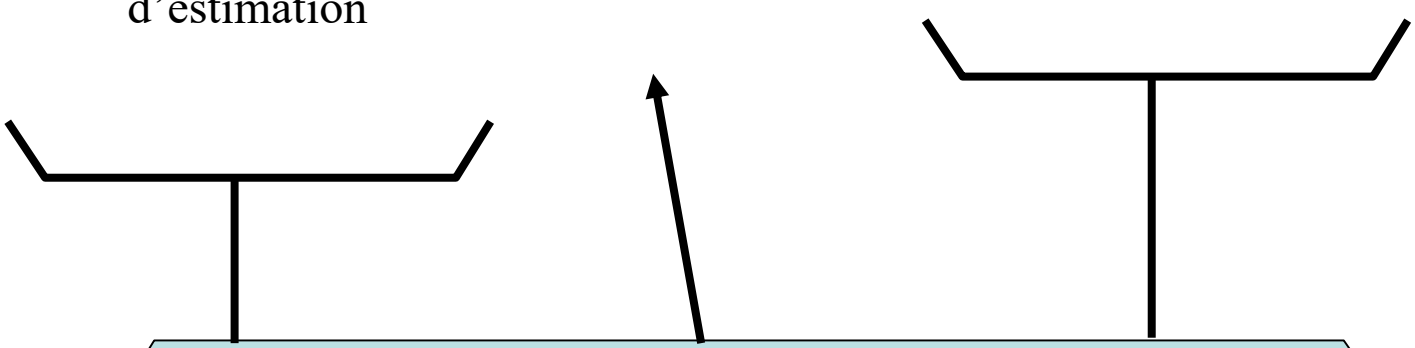
« assistée » :
basée sur N échantillons
et
sur la corrélation entre points

Autocorrélation
==>Redondance
==> perte d'information

⇩ appauvrissement de conditions d'estimation

Structure spatiale
==> information supplémentaire

⇧ augmentation des conditions d'estimation



Compensation inconnue *a priori*
qui dépend de la structure spatiale du phénomène
et de la géométrie de l'échantillonnage

L’histoire de la géostatistique en 2 diapo

Geostatistics : When ? Where ? Why ? How ?

When: in the 60s.

Where: in south african (gold) mines.

Why: because there was systematic over-estimation of recoveral reserves. Selecting blocks on the basis of their sample value(s) led the keep blocks that were poorer than expected ➔ money losts ….

How: Krige (South Africa) worked on conditional bias associated to block selection and suggested to use inner and outer samples (first spatial considerations).
Matheron (France) developed rigorous estimators and founded a general discipline called geostatistics.

Fisheries Geostatistics : Where ? Why ? When ? How?

Where: in Europe.

Why: because acoustic tools provided autocorrelated samples which could not be handled properly by traditional statistical methods (non spatial).

When: in the 80s.
A precursor: Alain Laurec 1977 (fishing power).
Some other precursors : Gérard Conan (snow crab), Francis Laloë (survey design), Francis Gohin (survey design) in 1985.
1990 : Jacques Rivoirard, Ken Foote, John Simmonds, Pierre Petitgas ➔ ICES workshop & formal recommendation.
1990-date : regular spreading of the method

Some key references

In geostatistics

Chilès J.P. et Delfiner, 1999. Geostatistics, Modeling spatial uncertainties. Ed. Wiley.
Complete and deep presentation of the geostatistical theory. « The state of the art…. »

Cressie N. 1991. Statistics for Spatial data. Ed. Wiley.
A statistical approach of geostatistics.

In fisheries geostatistics

Rivoirard J., J. Simmonds, K. Foote, P. Fernandes and N. Bez, 2000. Geostatistics for estimating fish abundance. Ed. Blackwell Science.
Theory & practice of geostatistics applied to fish stock assessement from survey data.

Petitgas, P., Woillez, M., Rivoirard, J., Renard, D., and Bez, N. 2017. Handbook of geostatistics in R for fisheries and marine ecology. ICES Cooperative Research Report No. 338. 177 pp
https://www.ices.dk/sites/pub/Publication%20Reports/Cooperative%20Research%20Report%20(CRR)/CRR338.pdf

One (amongst many) R package

gstlearn (https://gstlearn.org/)

Empirical properties of spatial data

Quelques concepts à l'œuvre,
sans modélisation…

Variable régionalisée

- Phénomène régionalisé:
  - Paramètre abiotique (e.g. température, concentration en oxygène)
  - Fond marin
  - Biomasse, abondance

- Variable régionalisée $z(x)$, x à 1D, 2D ou 3D usuellement:
  - concentration en chlorophylle, oxygène (mg/l, 3D)
  - cote topographique (m, 2D)
  - Densité de biomasse (kg/m²) ou densité d'individus (ind/m²)

- Support (notion importante):
  point, volume ou surface (ex: bloc) générique, d'orientation donnée, sur lequel est définie ou mesurée la variable régionalisée
  $$z(x), z(v), z(V)$$
- Champ: zone géographique où le phénomène est présent.

  Connu/inconnu, fini/(quasi)infini.

  Correspond à la notion d'habitat si z(x) est une biomasse.

Variable sommable

v_i égaux partitionnant V

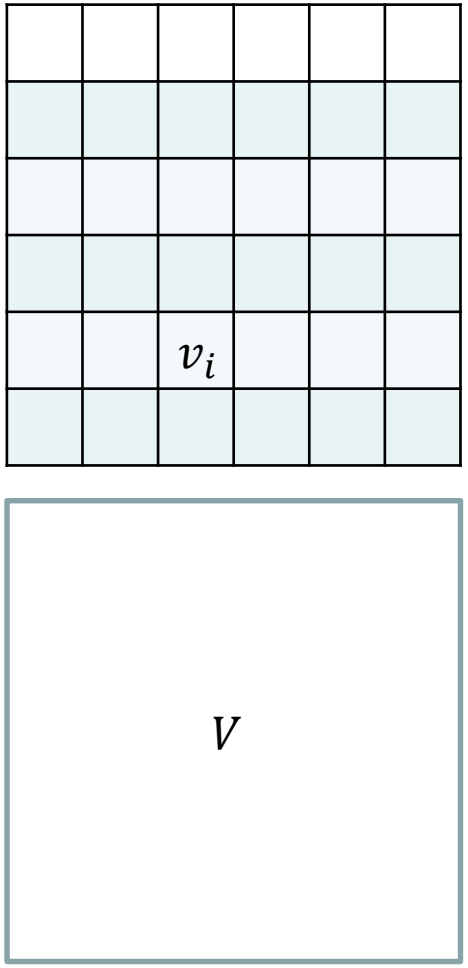
$$z(V) = \frac{1}{N}\sum_i z(v_i)$$

ou

$$z(V) = \frac{1}{V}\int_V z(x)\,dx$$

➔ valeur moyenne = moyenne des valeurs

➔ Be careful: proportions, mean length, biodiveristy indices, etc, do not suit this property.



Statistiques clefs

Valeurs $z_1, z_2, \ldots, z_i, \ldots, z_n$

Moyenne $m = \frac{1}{n} \sum_{1}^{n} z_i$

Variance $\sigma^2 = \frac{1}{n} \sum_{1}^{n} (z_i - m)^2 = \frac{1}{n} \sum_{1}^{n} z_i^2 - m^2$

Écart-type $\sigma = \sqrt{\sigma^2}$

Coefficient de variation $CV = \frac{\sigma}{m}$

Non spatiales

Variance de dispersion de v dans V $s^2(v|V) = \frac{1}{N} \sum_{i} (z(v_i) - z(V))^2$

Spatiales

Variances de dispersion par l'exemple

Variable additive z sur champ V , réunion de 6 parcelles élémentaires égales:

v_1	v_2	v_3	v_4	v_5	v_6
w_1		w_2		w_3	
V					

$v_i, i = 1, \dots, 6$

$w_j, j = 1, \dots, 3$

1	3	2	4	2	6

$z(v_i), i = 1, \dots, 6$

a) Moyenne de z sur le champ V ?
Variance des (valeurs de) parcelles v_i dans le champ ? $s^2(v|V)$

b) Le même champ est maintenant divisé en 3 blocs w_j, de 2 parcelles contiguës chacun.
Valeurs moyennes des blocs ?
Variance des blocs w_j dans le champ V ? $s^2(w|V)$

c) Variances des valeurs de parcelles v_i dans chacun des 3 blocs w_j ?
Moyenne de ces variances dite variance de dispersion des parcelles dans un bloc?
$s^2(v|w)$

Solution: Variances de dispersion

1	3	2	4	2	6

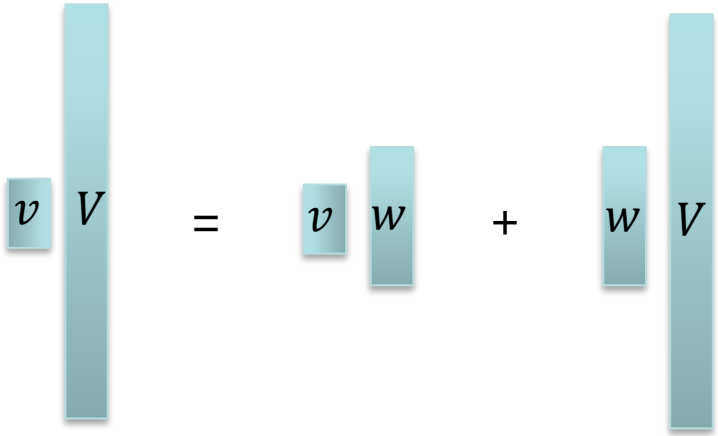
2	3	4

- Variance des valeurs de parcelles dans chacun des 3 blocs:
  1 1 4
- Moyenne de ces variances, dite variance des parcelles dans un bloc: (1+1+4)/3 = 2
- Additivité des variances de dispersion: 8/3 = 2 + 2/3

Variances de dispersion

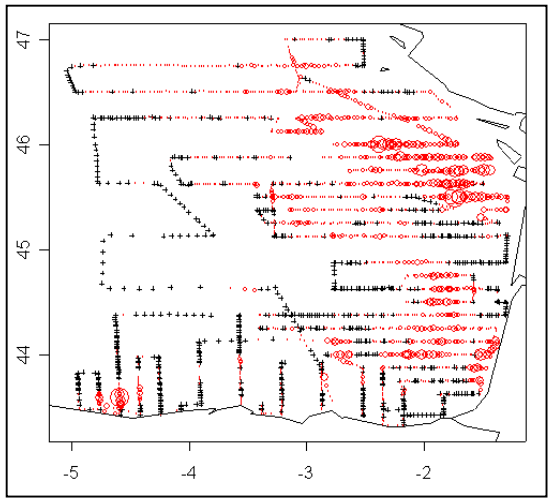
- Additivité:

$$s^2(v|V) = s^2(v|w) + s^2(w|V)\ avec\ v\ \subset w\ \subset V$$

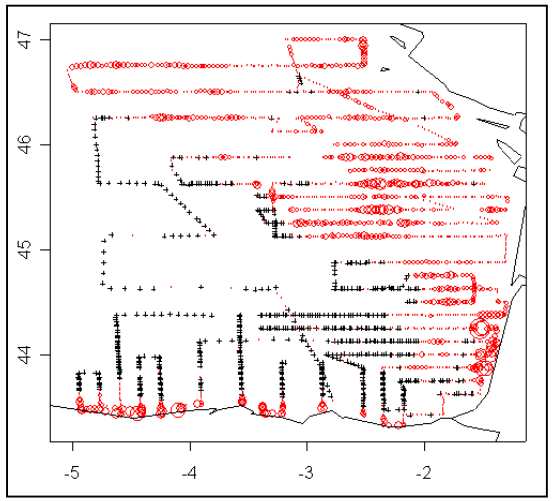


- **REGULARISATION :** La variance diminue quand le support augmente (pour des supports multiples l'un de l 'autre)

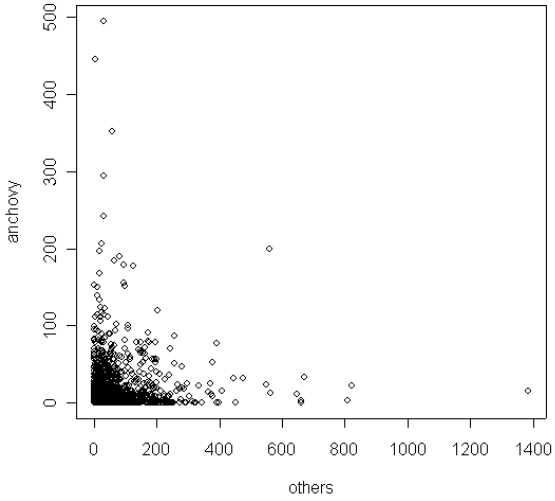
$$s^2(v|V)\ \geq s^2(w|V)\ avec\ v\ \subset w\ \subset V$$



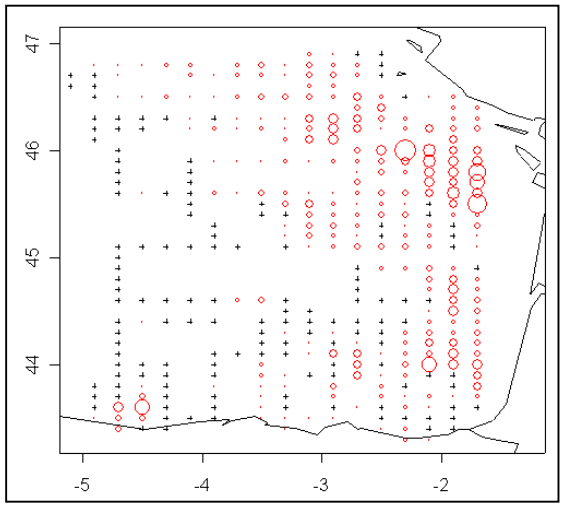
Anchois, œufs de stade I



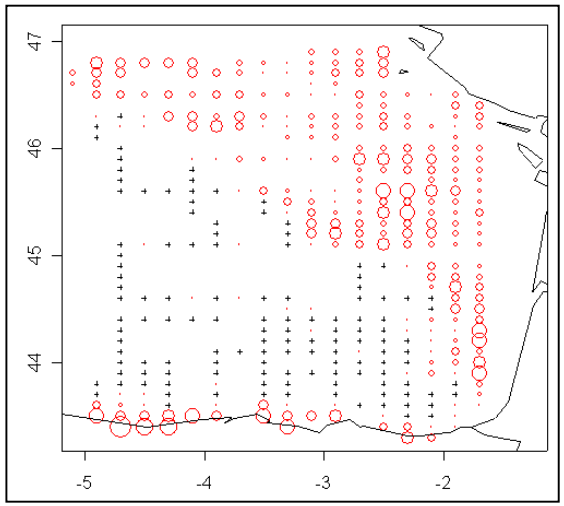
Autres œufs de stade I



$var_1 = 5854$
$var_2 = 774$
$\rho = 0.129$

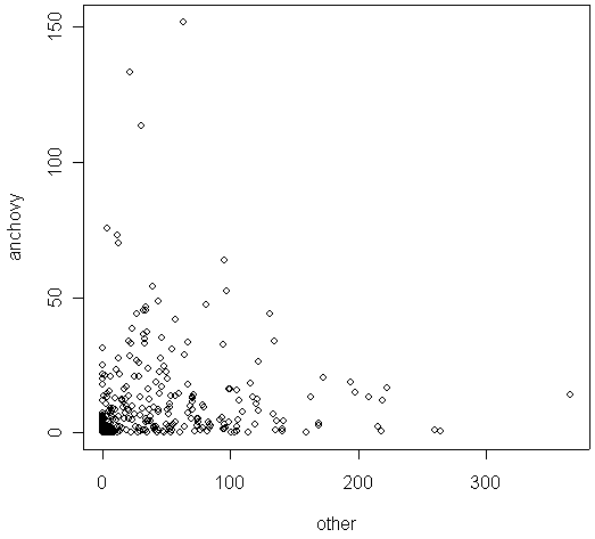


Anchois, œufs de stade I



Autres œufs de stade I

Same data with a larger support (resolution)



$var_1 = 2722$
$var_2 = 283$
$\rho = 0.134$

Up-scaling

Passage possible des supports petits au supports plus grands avec intégration progressive et contrôlée des structures spatiales

Down-scaling

Impossible de passer des grands supports aux petits sans hypothèses sur les schémas d'organisation aux résolutions inférieures. Problème inverse. Déconvolution.

Simulation (conditionnelles) au support petit conditionnellement au grand possible.

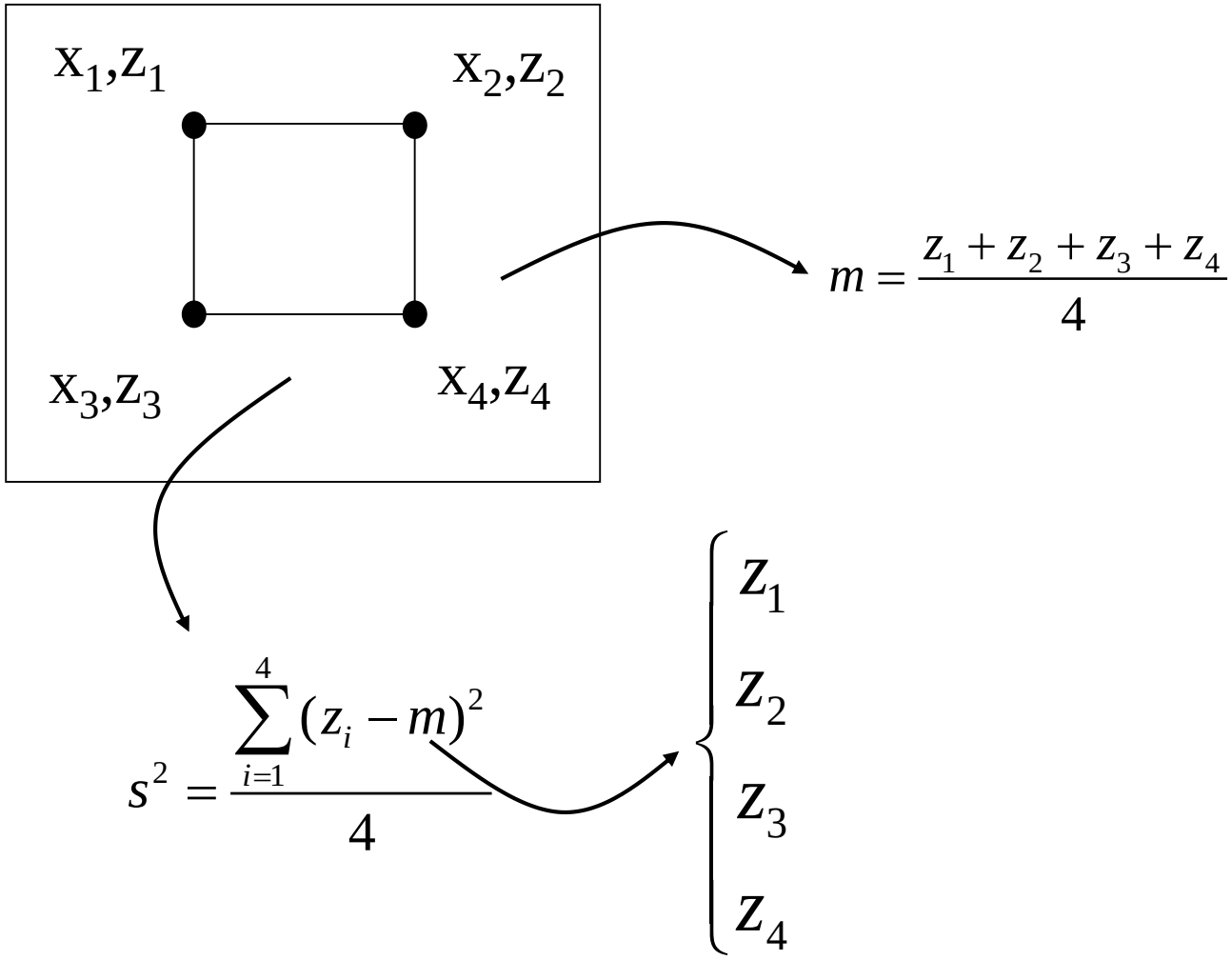
From Dungan et al., 2002. Ecography.

	Phenomenom	Observations	Analysis	Fuzzyness
Extent	X	X	X	hight
Grain	X	X	X	hight
Resolution		X	X	Medium
Lag		X	X	Medium
Support		X		Low
Cartographic ratio			X	Low
Scale	X	X	X	high

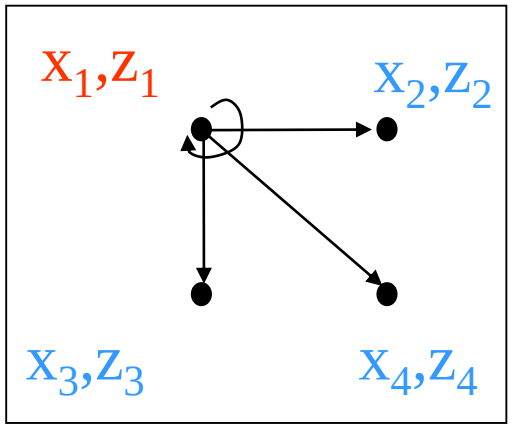
De la variance au variogramme

Analyse exploratoire

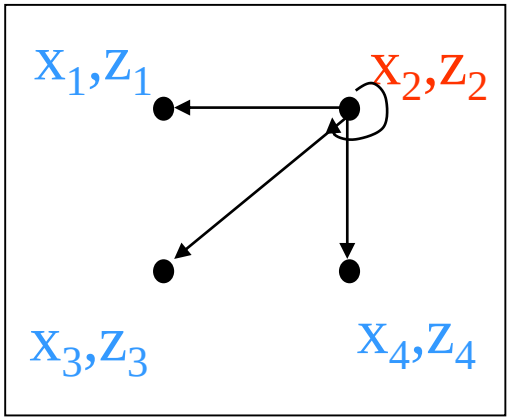
Toujours sans modélisation…


x1,z1
x2,z2
x3,z3
x4,z4
$m = \frac{z_1 + z_2 + z_3 + z_4}{4}$
$s^2 = \frac{\sum_{i=1}^{4} (z_i - m)^2}{4}$
z_1
z_2
z_3
z_4

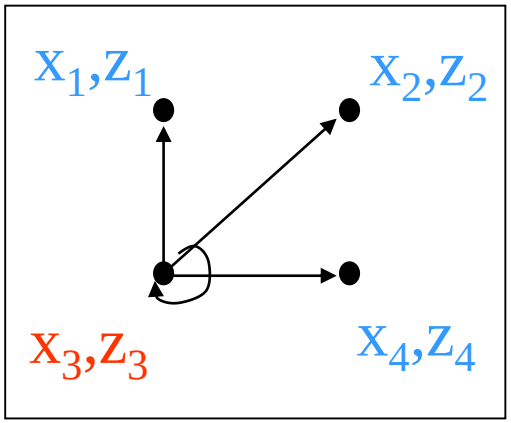
i=1, j= 1,2,3,4



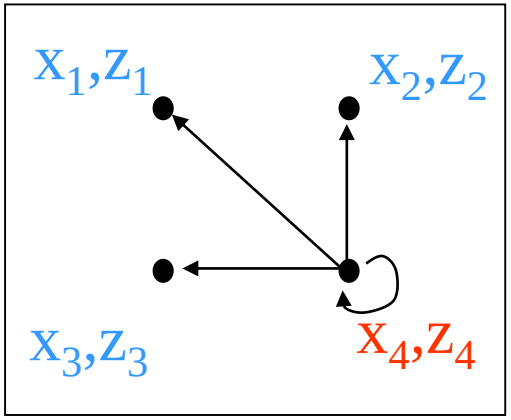
i=2, j= 1,2,3,4



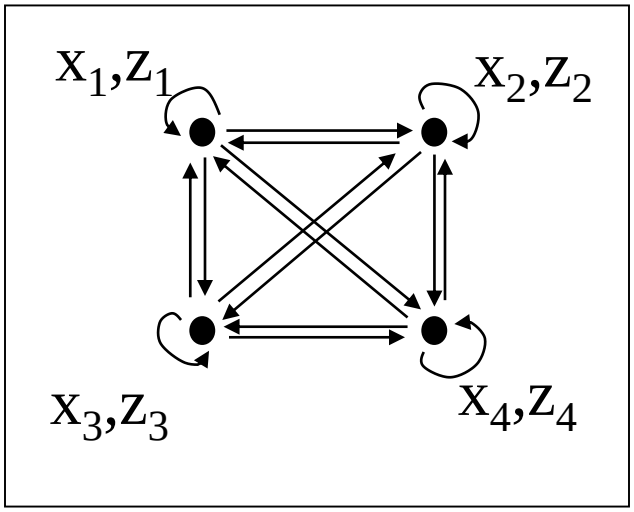
i=3, j= 1,2,3,4



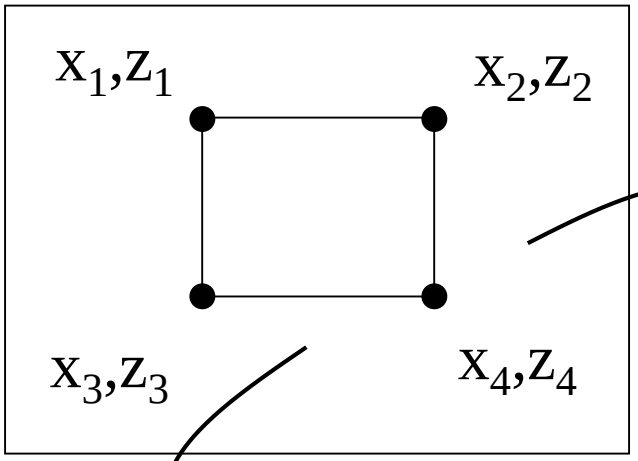
i=4, j= 1,2,3,4



Bilan du nombre de paires



Total = 16 = 4^2



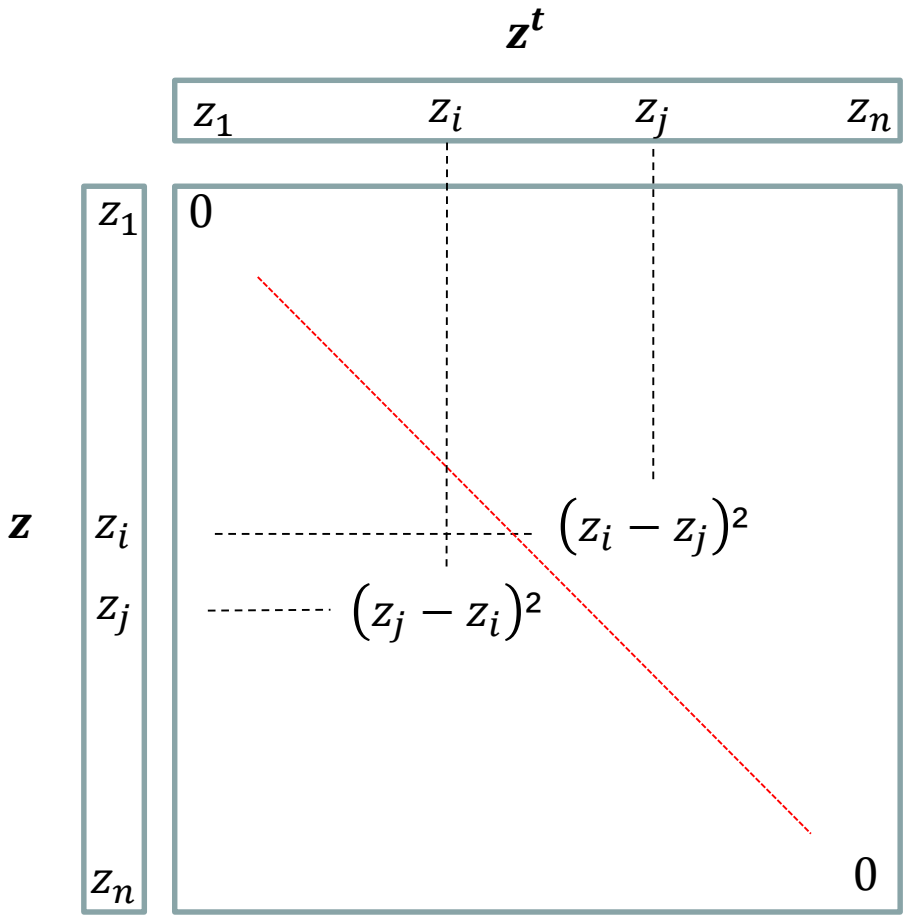
$$m = \frac{z_1 + z_2 + z_3 + z_4}{4}$$

$$s^2 = \frac{\sum_{i=1}^{4} (z_i - m)^2}{4}$$

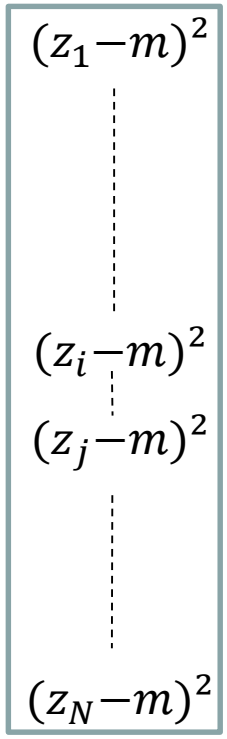
$$\begin{cases} z_1 \\ z_2 \\ z_3 \\ z_4 \end{cases}$$

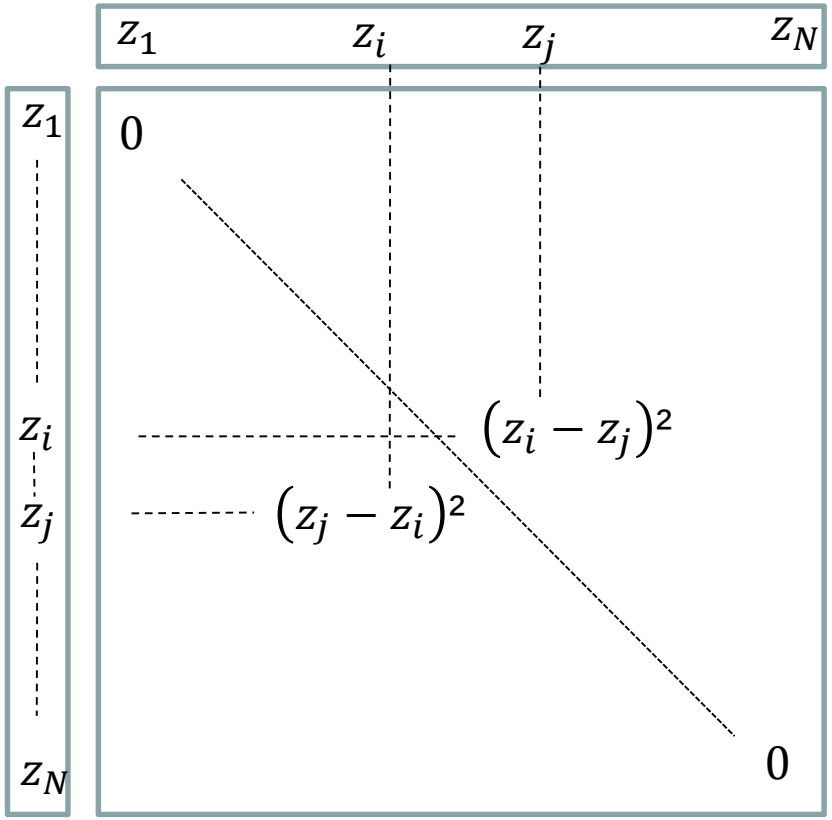
==>

$$s^2 = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} (z_i - z_j)^2}{2 * 4^2}$$



$$s^2 = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (z_i - z_j)^2}{2 * N^2}$$





$$s^2 = \frac{\sum_{i=1}^{N} (z_i - m)^2}{N}$$

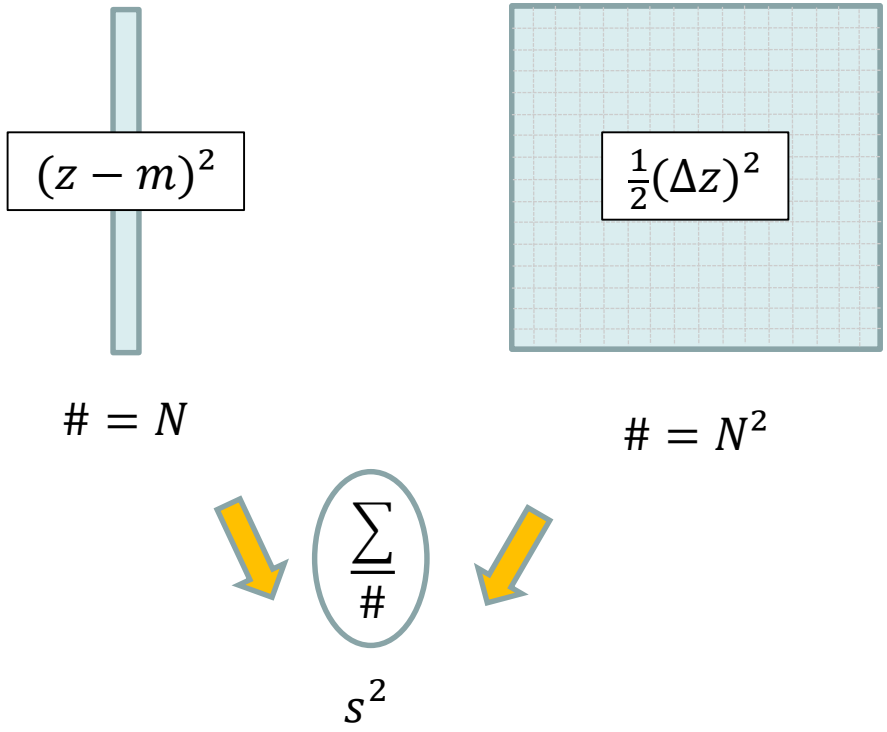
$$s^2 = \frac{1}{2} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (z_i - z_j)^2}{N^2}$$

$$s^2 = \frac{\sum_{i=1}^{N_{obs}} (z_i - m)^2}{N_{obs}} = \frac{1}{2} \frac{\sum_{i=1}^{N_{obs}} \sum_{j=1}^{N_{obs}} (z_i - z_j)^2}{N_{obs}^2} = \frac{1}{2} \frac{\sum_{i=1}^{N_{obs}} \sum_{j=1}^{N_{obs}} (z_i - z_j)^2}{N_{paires}} = \frac{\sum_{k=1}^{N_{paires}} \frac{1}{2}(\Delta z_k)^2}{N_{paires}}$$

$$\text{mean} = \frac{\text{Sum of elements}}{\text{Number of elements}} = \frac{\sum}{\#}$$

$$s^2 = \frac{\sum}{\#} (z - m)^2 = \frac{\sum}{\#} \frac{1}{2} \Delta z^2$$

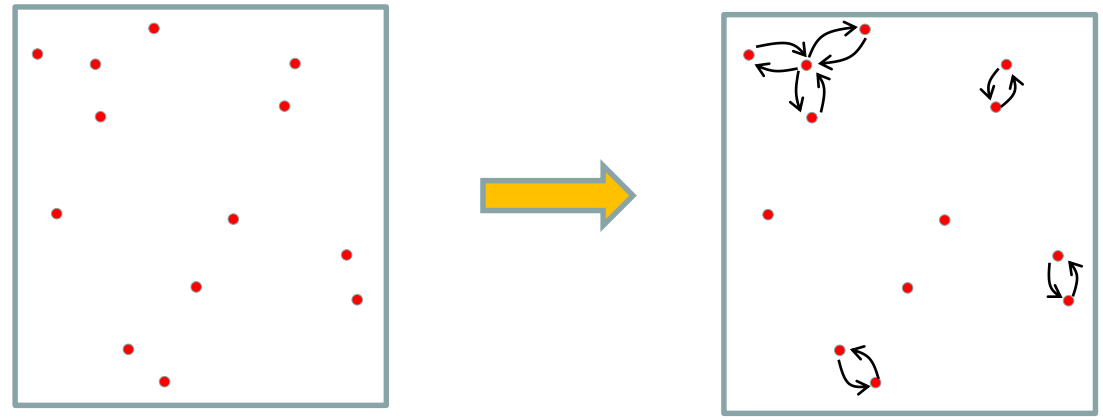
Variance:
mean square difference between observations and their mean
or
mean of the half square differences between observations

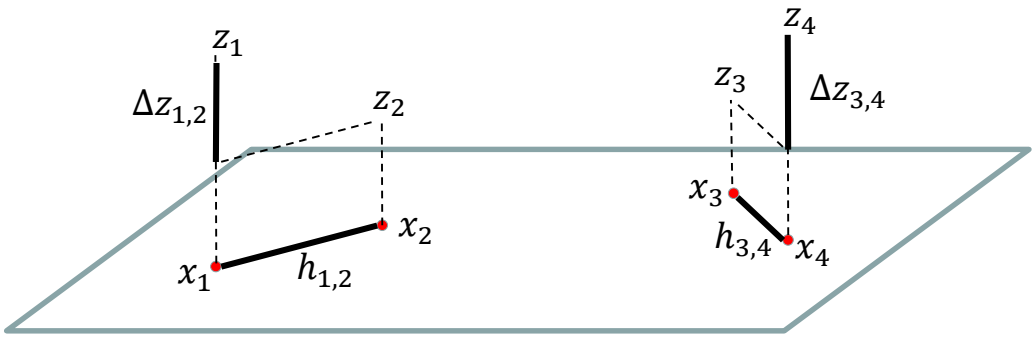


Prise en compte de la distance géographique

N points

N² paires de points





N² couples $(h_{i,j}, \Delta z_{i,j})$

$$h_{i,j} = |x_i - x_j|$$

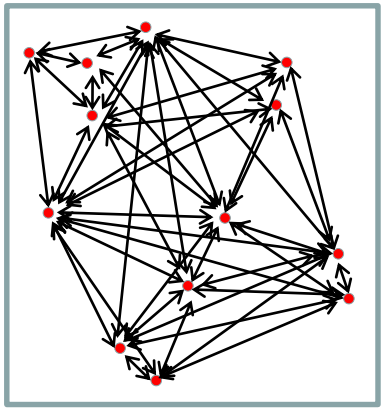
$$\Delta_{i,j} = z_i - z_j$$

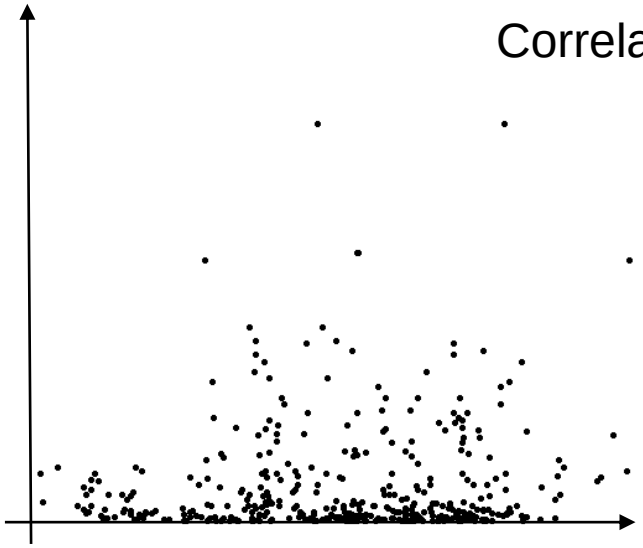
$\frac{1}{2}(\Delta z)^2$

Semi square differences

Variogram cloud

Correlation plot





h

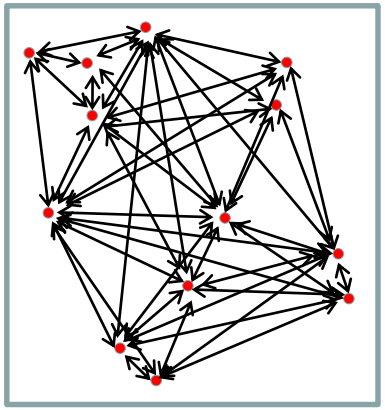
Distance

$\frac{1}{2}(\Delta z)^2$

Semi square differences

Variogram cloud

Correlation plot



$s^2 = \frac{\sum}{\#}$

Variance

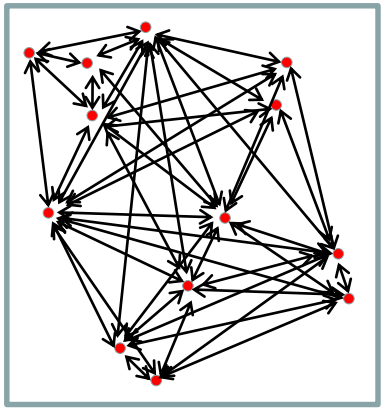
h

Distance

Regroupement des paires par classes de distance géographique : le variogramme

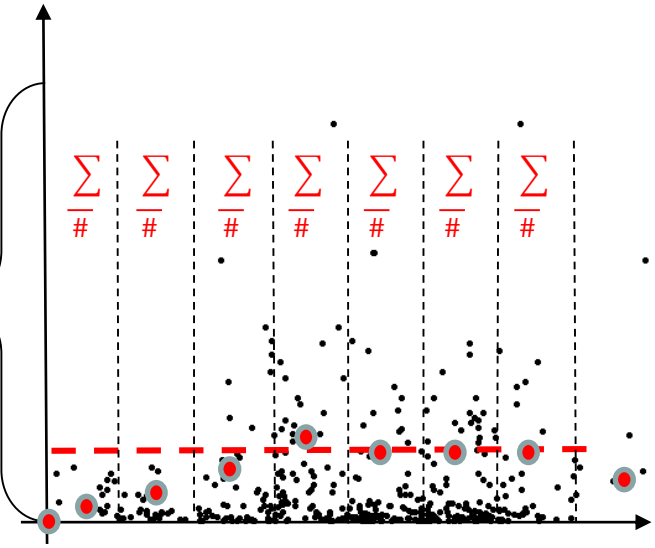
$\frac{1}{2}(\Delta z)^2$

Semi square differences



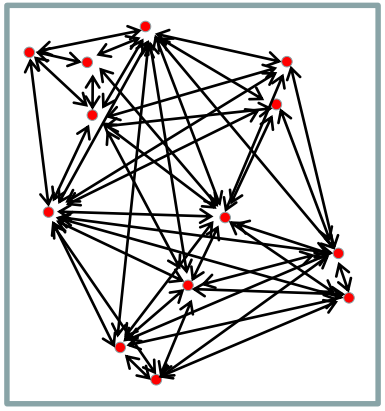
$s^2 = \frac{\sum}{\#}$

Variance



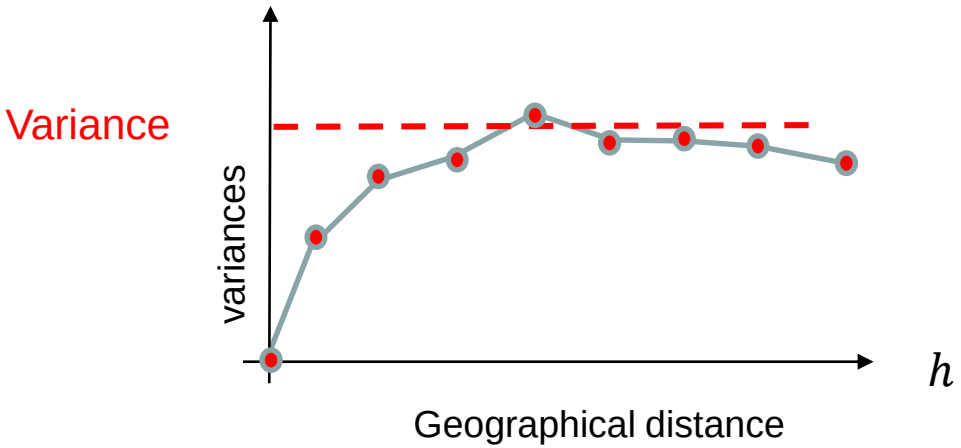
h

Distance



$$\gamma(h) = \frac{\sum_{h_{i,j} \approx h} (z_i - z_j)^2}{2 \cdot N(h)}$$

Variogram plot



Rq: Despite the ½ in the formulae, the y-axis does *not* correspond to a semi-variance as it is often mentioned in the literature. It is a variance.

Variogramme

$$\gamma(h) = \frac{\sum_{x_i - x_j = h} (z_i - z_j)^2}{2.N(h)}$$

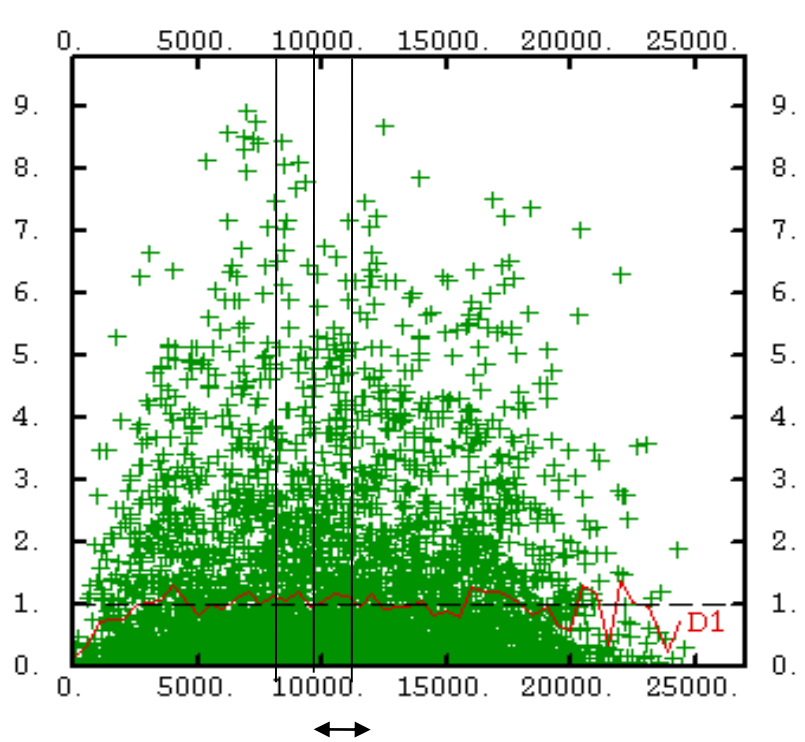
Demi écart quadratique moyen des points distants de h

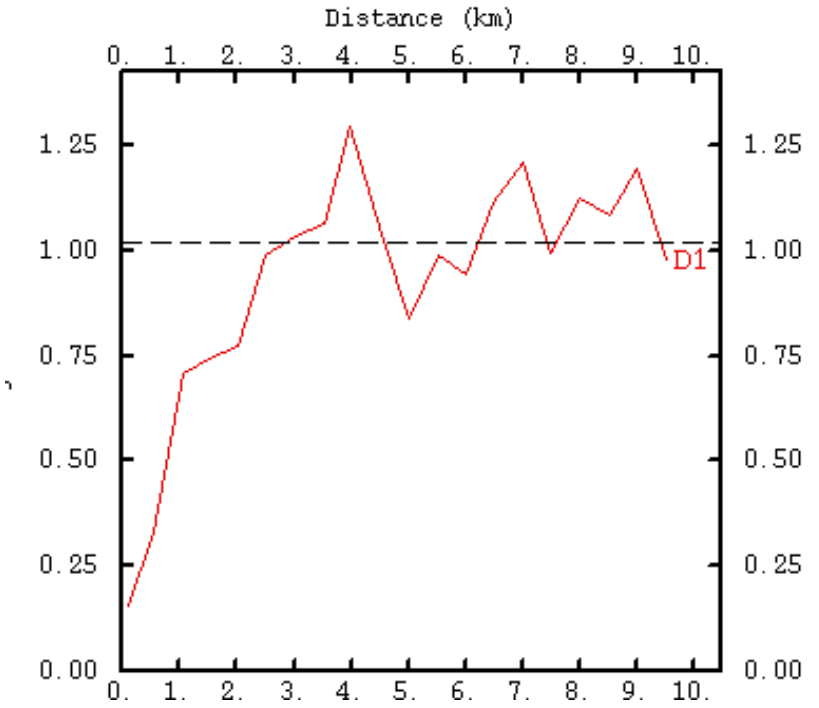
Variance

$$s^2 = \frac{\sum_h N(h)\gamma(h)}{\sum_h N(h)}$$

Données 2D – irrégulières

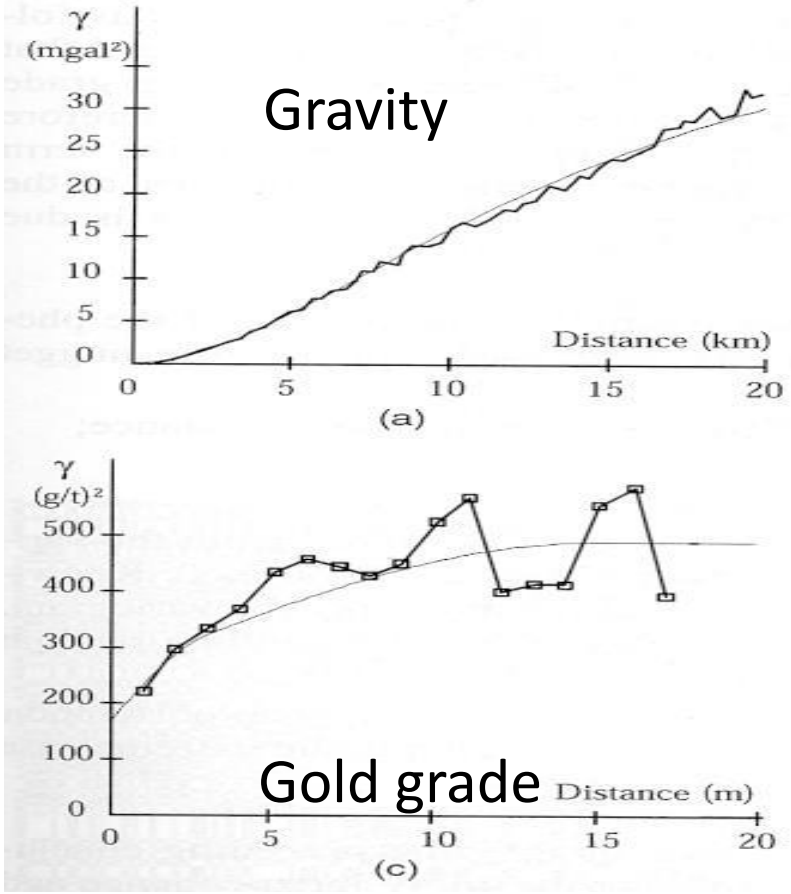
Variogramme expérimental



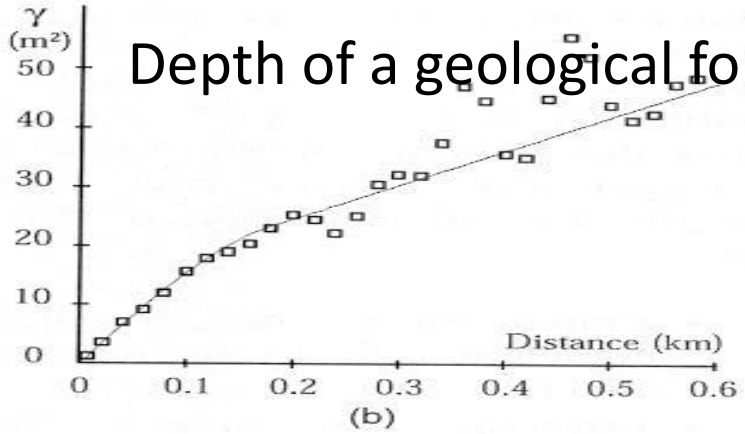


distance

Variogram example

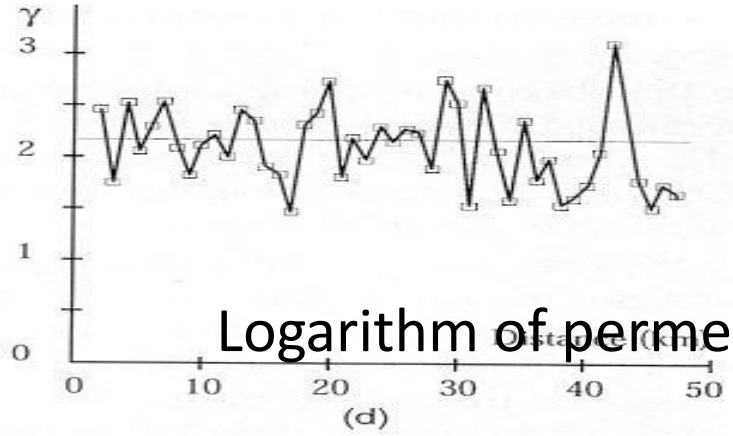


Gravity



Depth of a geological formation

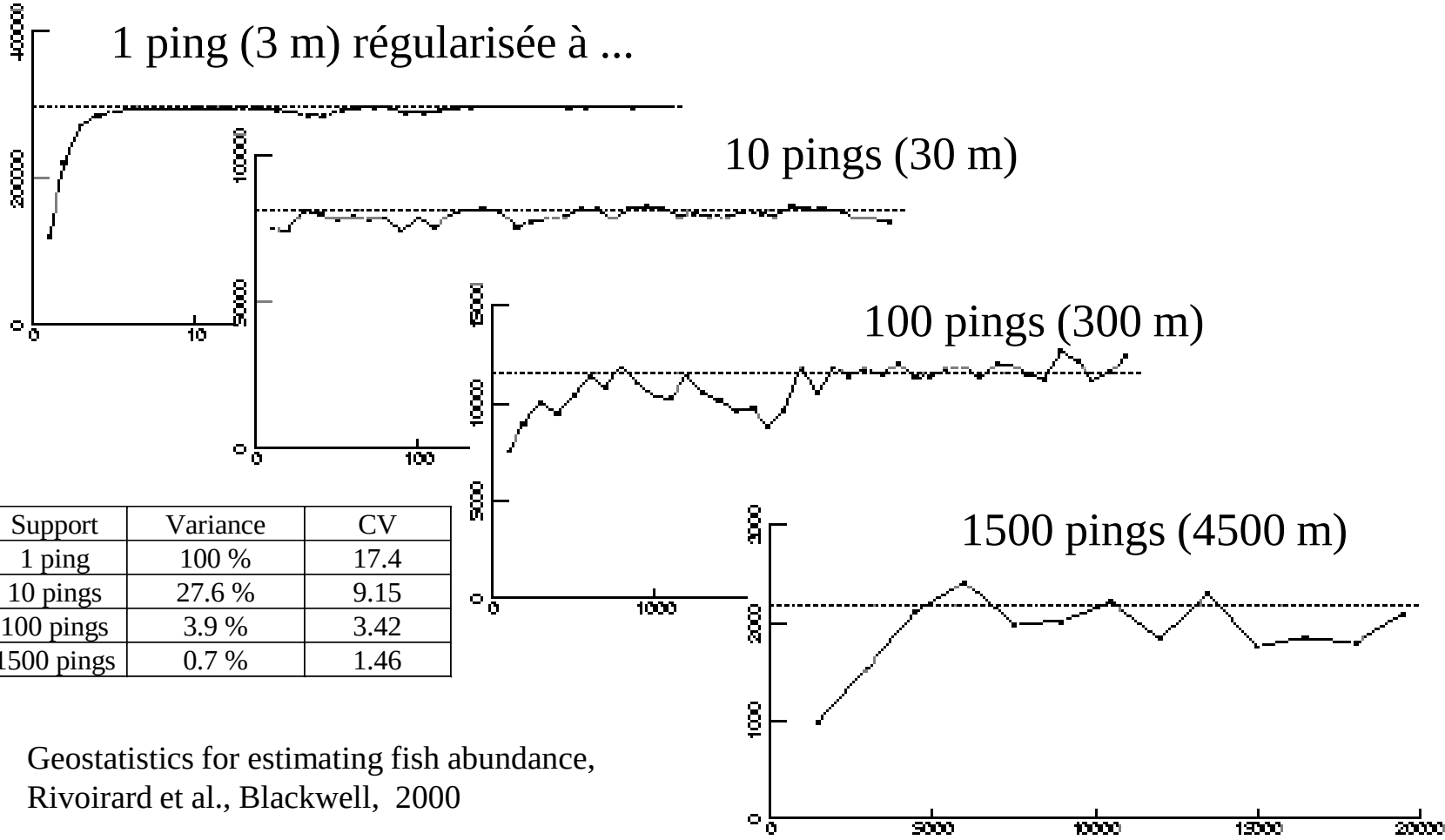
Gold grade



Logarithm of permeability

Extrait de “Geostatistics: modeling spatial uncertainty” J.P. Chilès, P. Delfiner (Wiley 1999)

Support effect

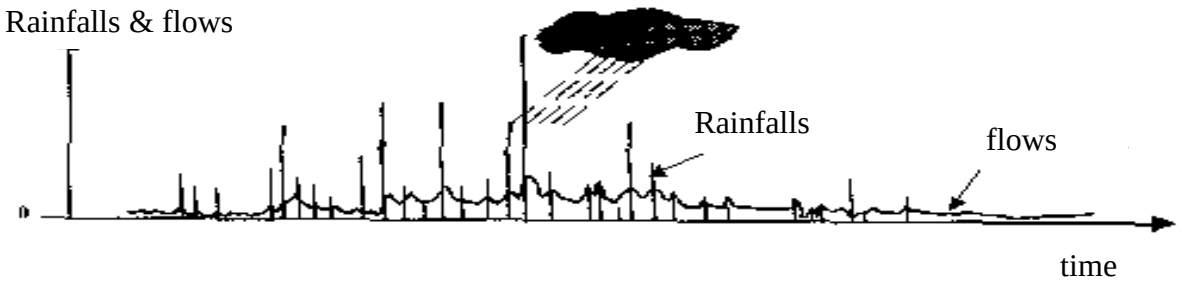


Support	Variance	CV
1 ping	100 %	17.4
10 pings	27.6 %	9.15
100 pings	3.9 %	3.42
1500 pings	0.7 %	1.46

Geostatistics for estimating fish abundance,
Rivoirard et al., Blackwell, 2000

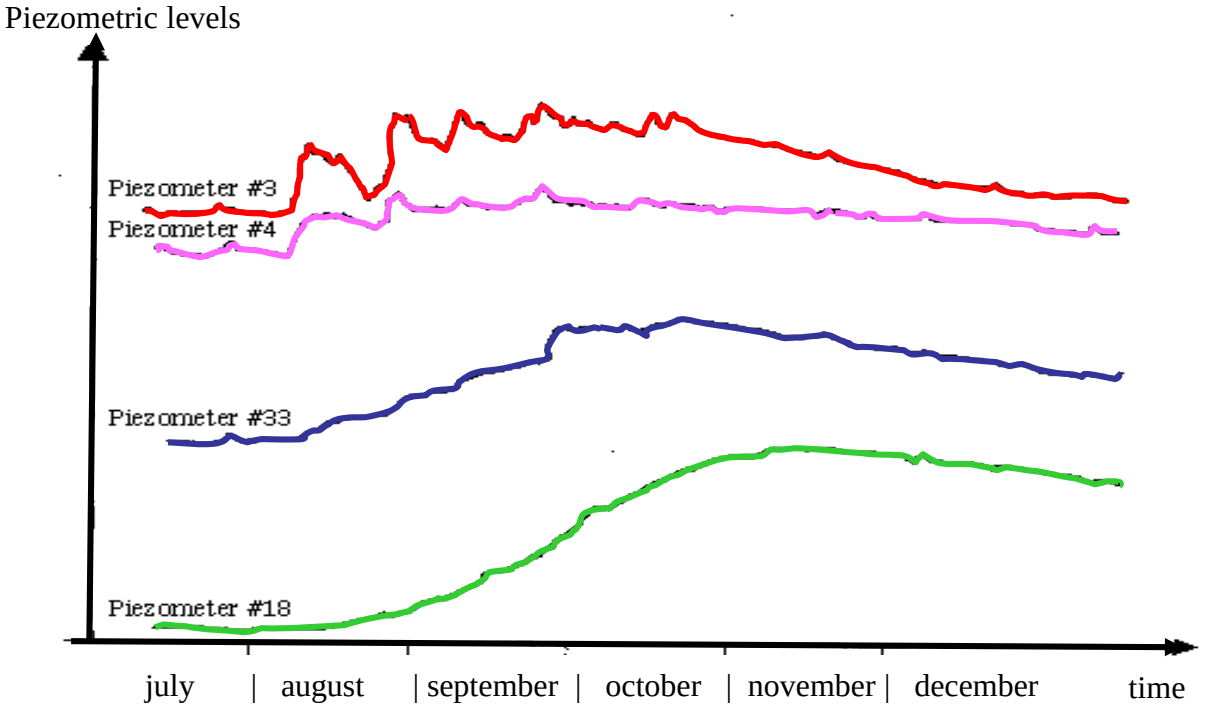
Marine Lab, Aberdeen

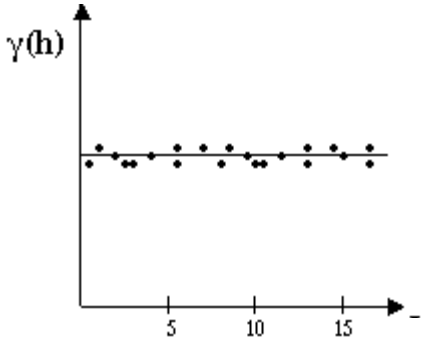
Interpretation of variograms



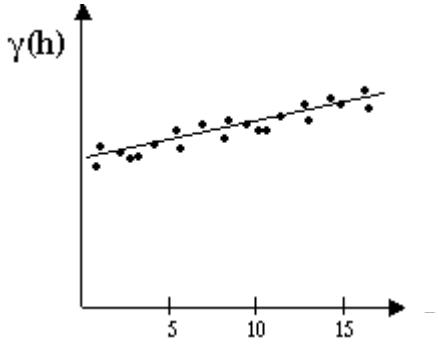
Piezometric level for an aquifer measured from July to Decembre.

Korhogo Bassin (Ivory Coast)

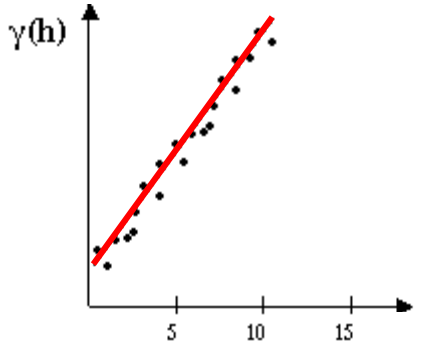




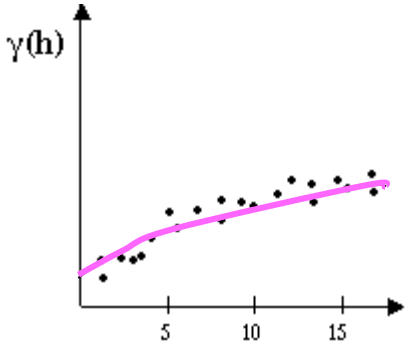
Pluie Days



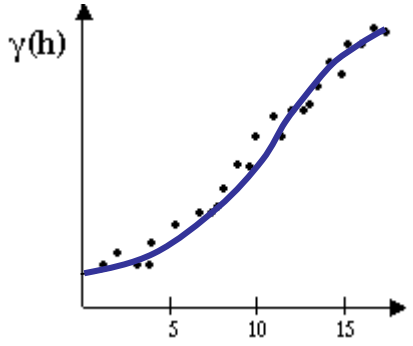
Ruissellement Days



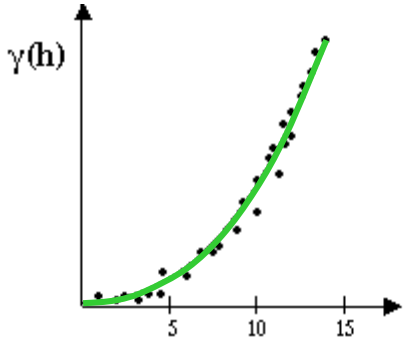
Piézomètre n°3 Days



Piézomètre n°4 Days



Piézomètre n°33 Days

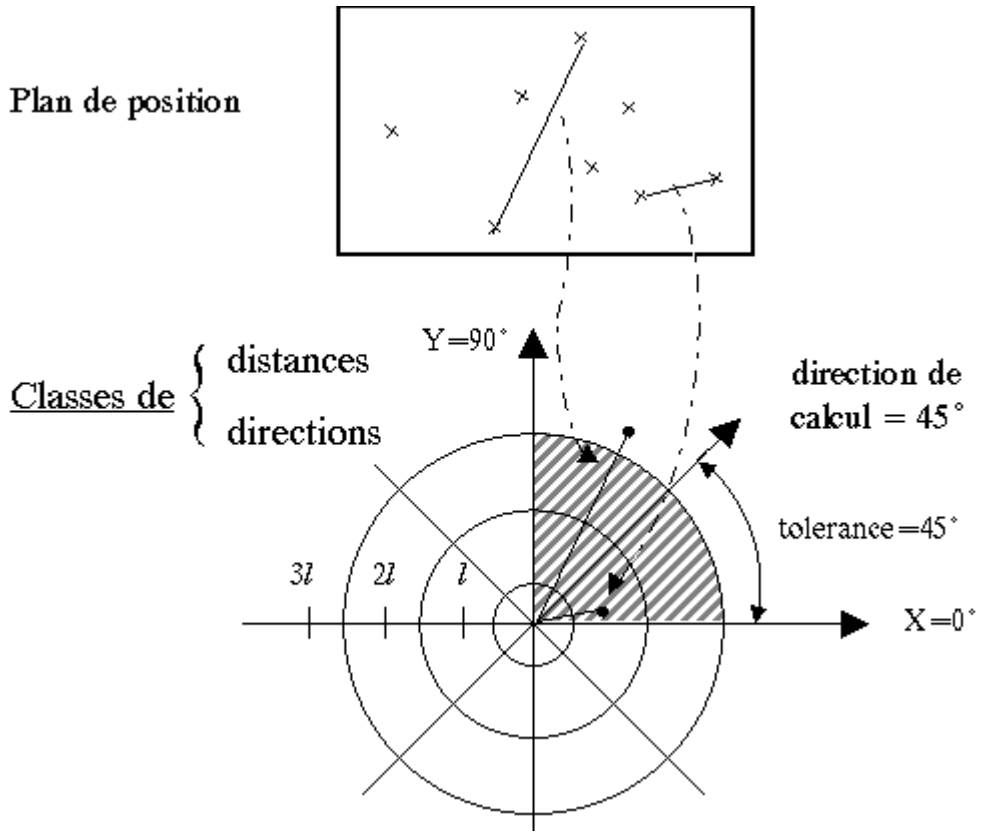


Piézomètre n°18 Days

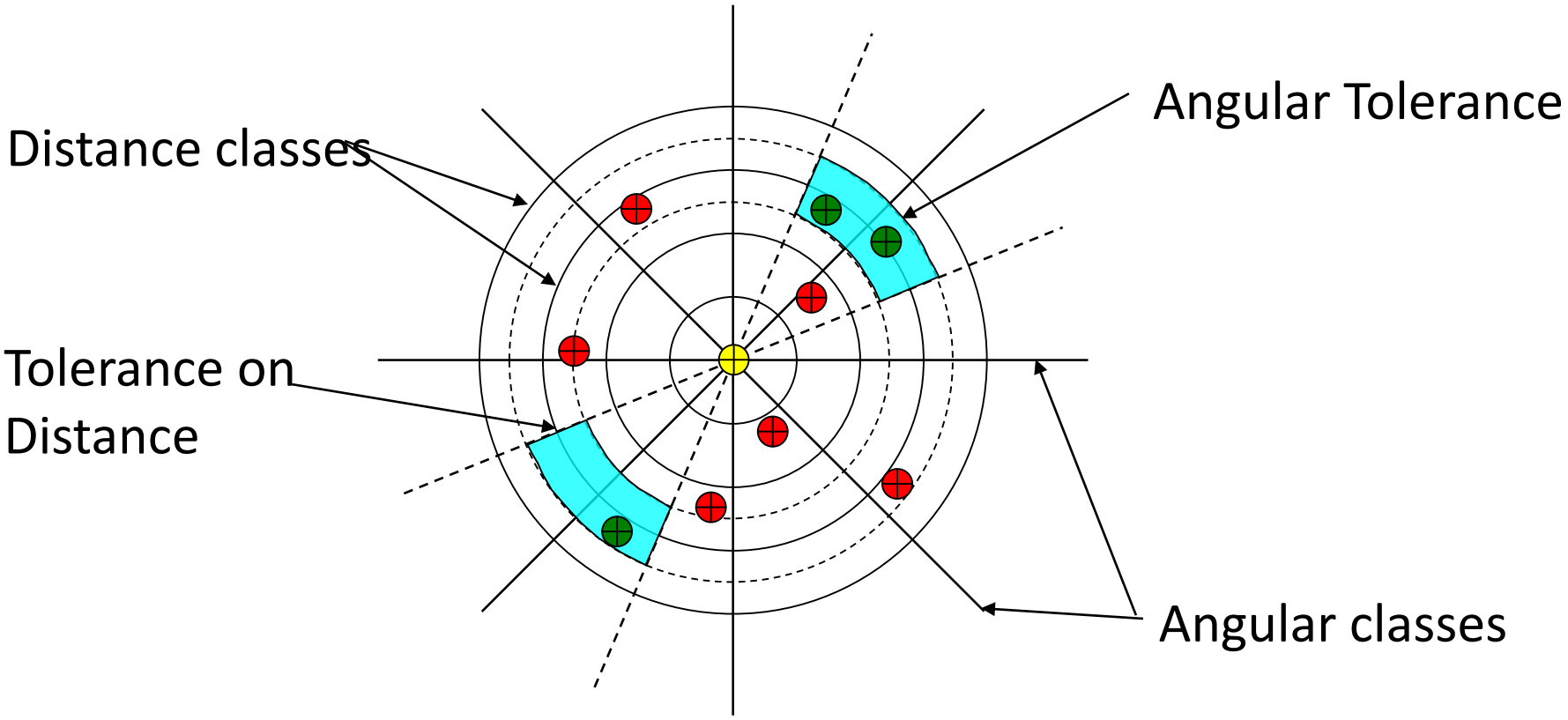
Echantillonnage irrégulier à 2D

Classes de distances:
- pas du variogramme
- nombre de pas
- tolérance sur le pas

Classes de directions:
- angle de référence
- nombre de secteurs
- tolérance angulaire

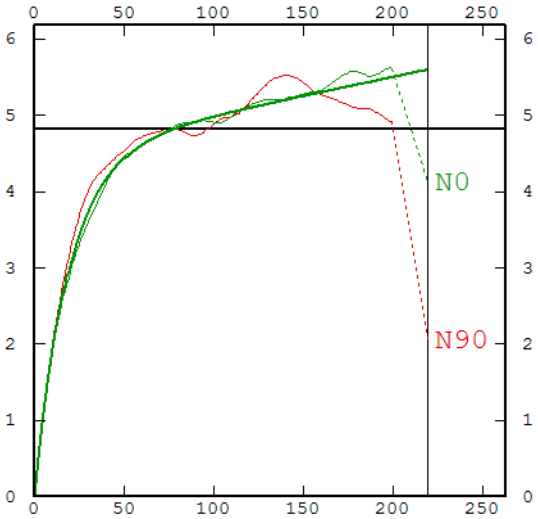


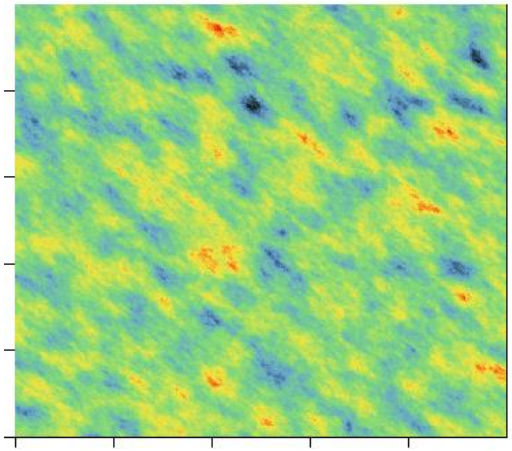
Directional Variograms



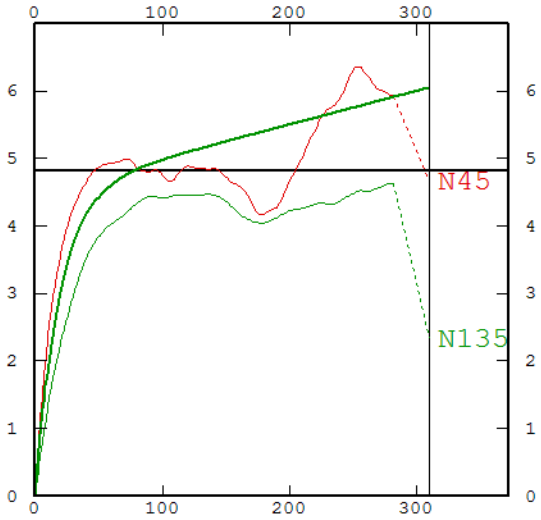
Anisotropy

N0 and N90



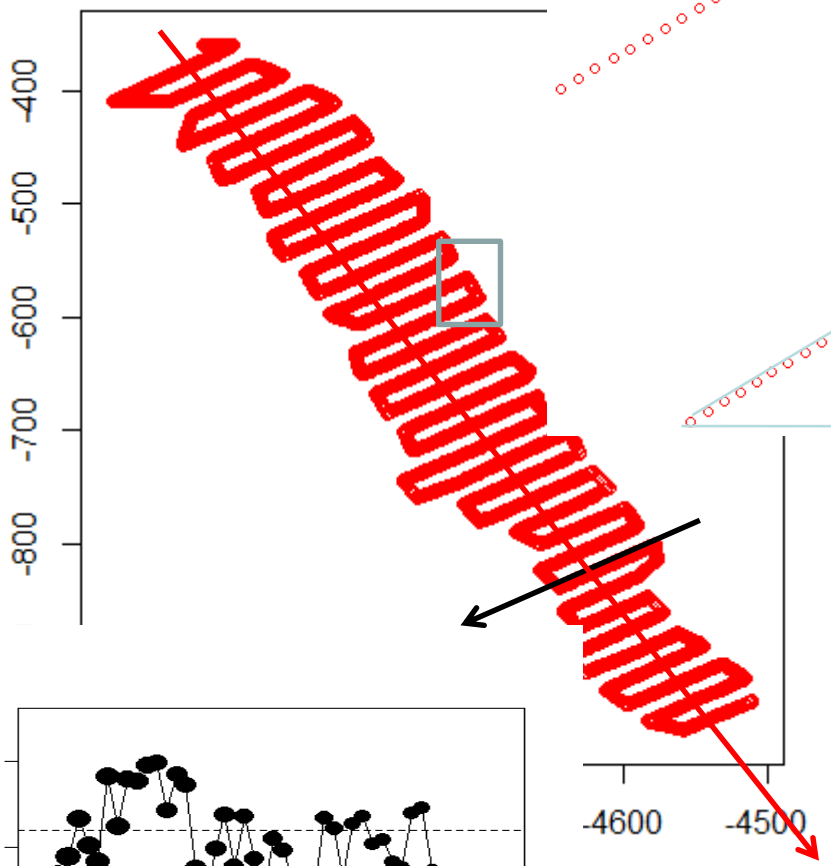


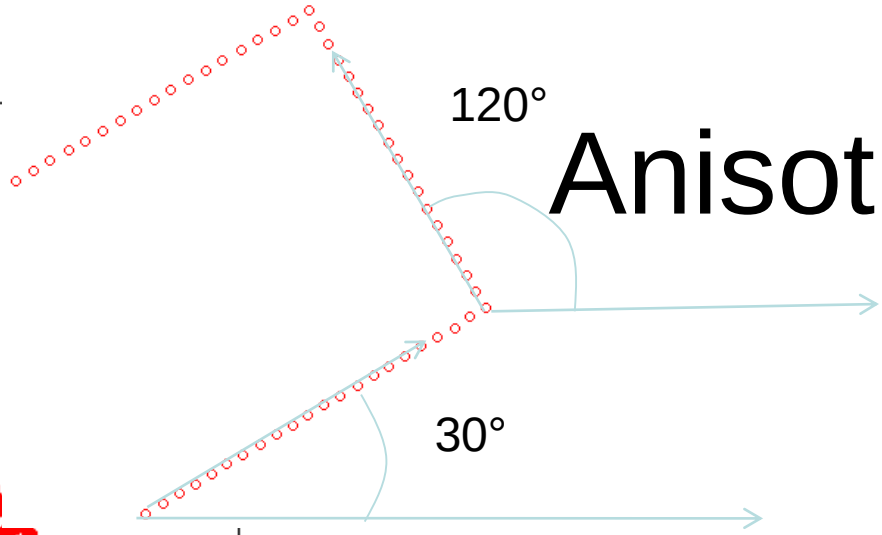
N45 and N135

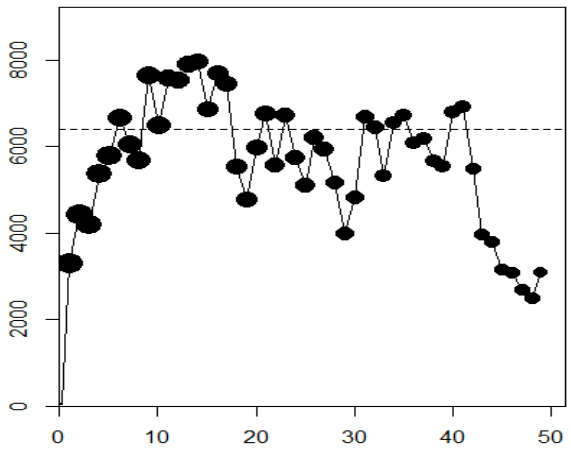


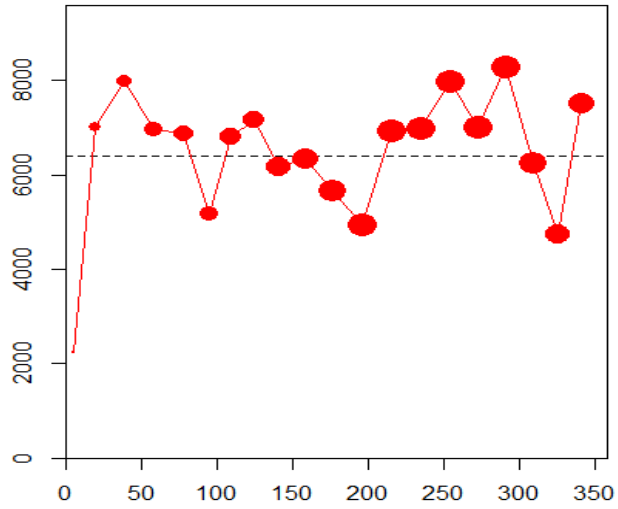
Need to compute at least 4 directions to detect the anisotropy

Anisotropy

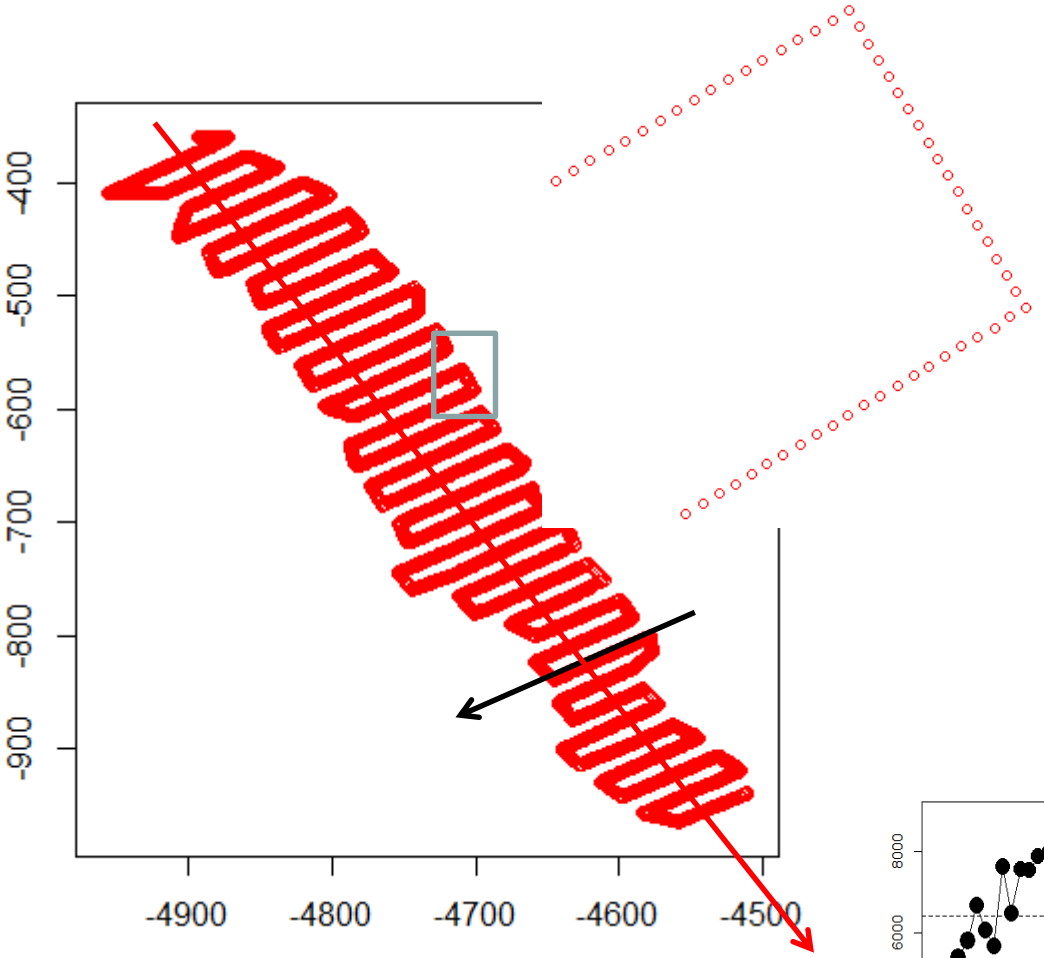

-400
-500
-600
-700
-800
-4600
-4500

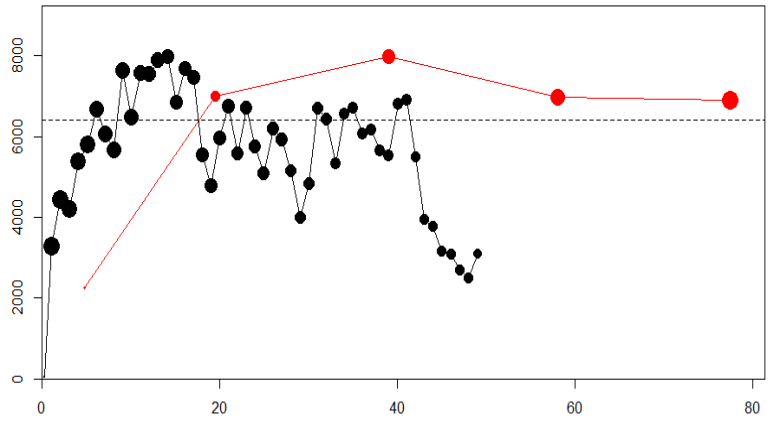

120°
30°

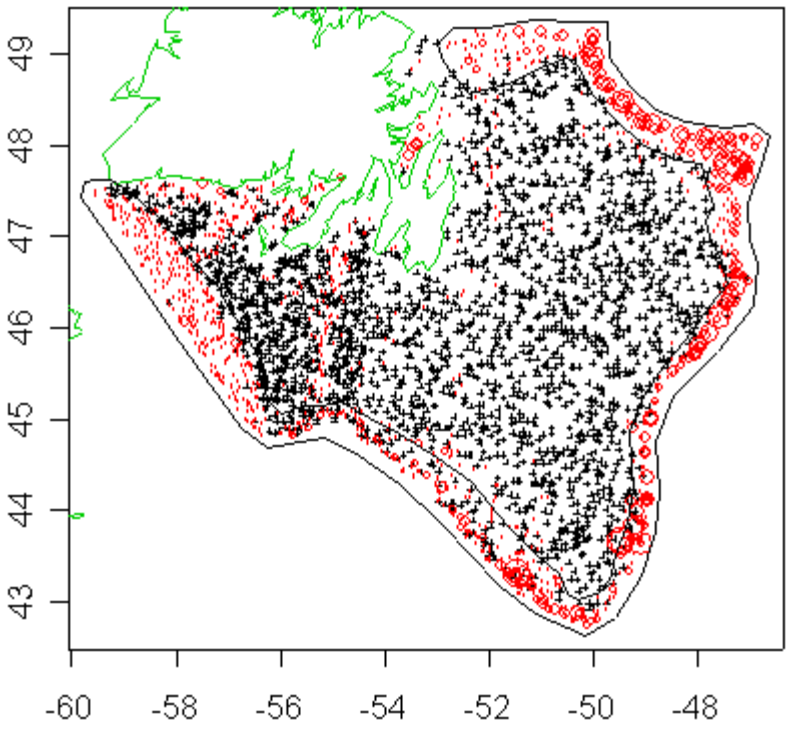

0
2000
4000
6000
8000
0
10
20
30
40
50

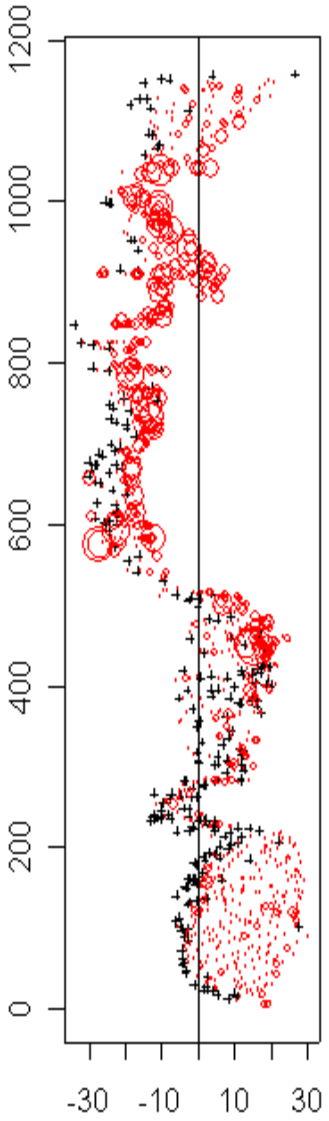

0
2000
4000
6000
8000
0
50
100
150
200
250
300
350

Anisotropy


-400
-500
-600
-700
-800
-900
-4900
-4800
-4700
-4600
-4500

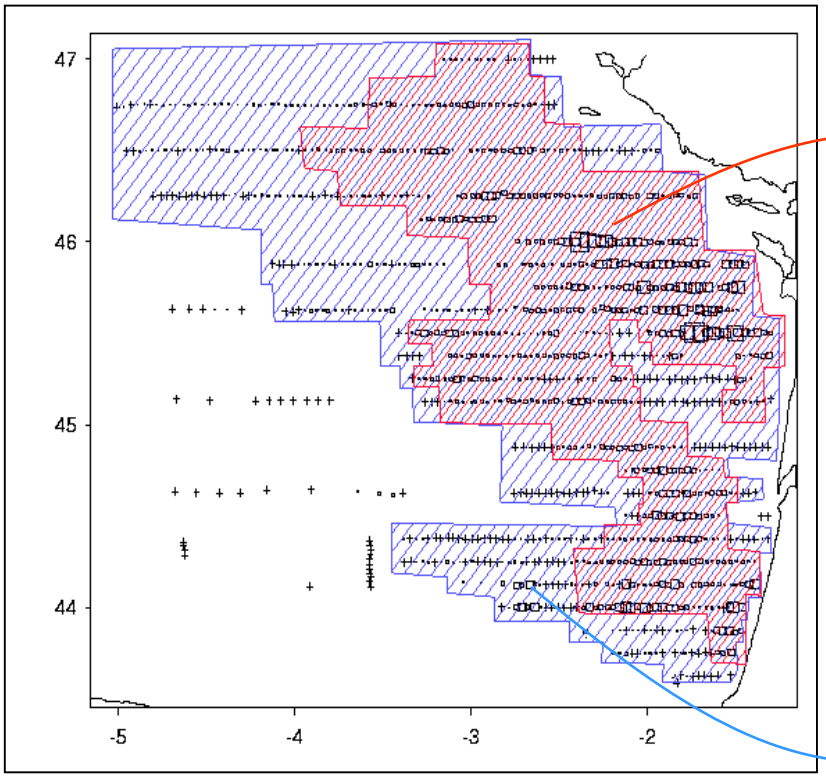

0
2000
4000
6000
8000
0
20
40
60
80

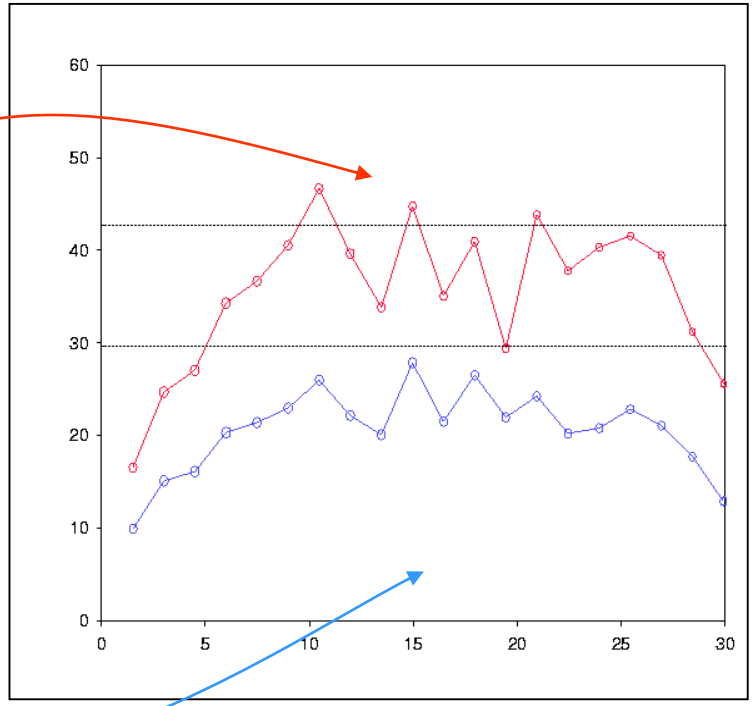




Greeland halibut

Field definition





Distances (n. mi.)

Variograms in the East-West direction

(direction of the transect)

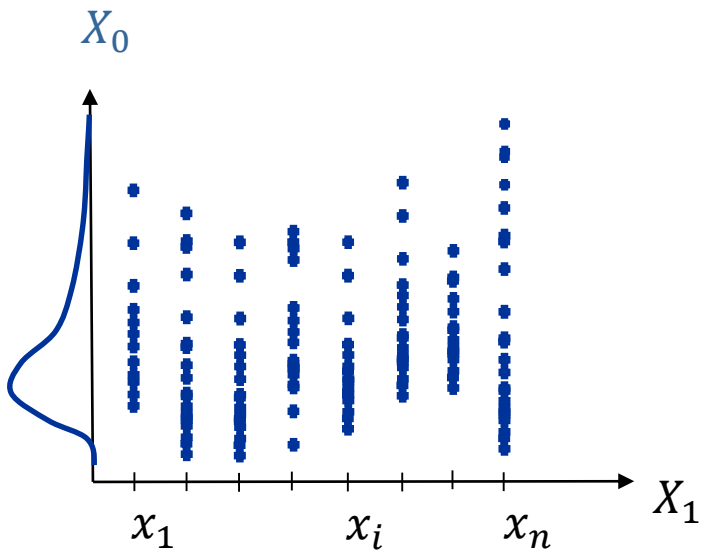
Variogram calculations: hints

- Use ad hoc projection systems
- Choose the lag and check sensitivity of the results ; check the homogeneity of the number of pairs for all lags
- If it is possible, compute the variograms among different directions (at least 4 in 2D) to detect a possible anisotropy

Estimation

Du modèle linéaire au krigeage

Conditional expectation: graphical definition

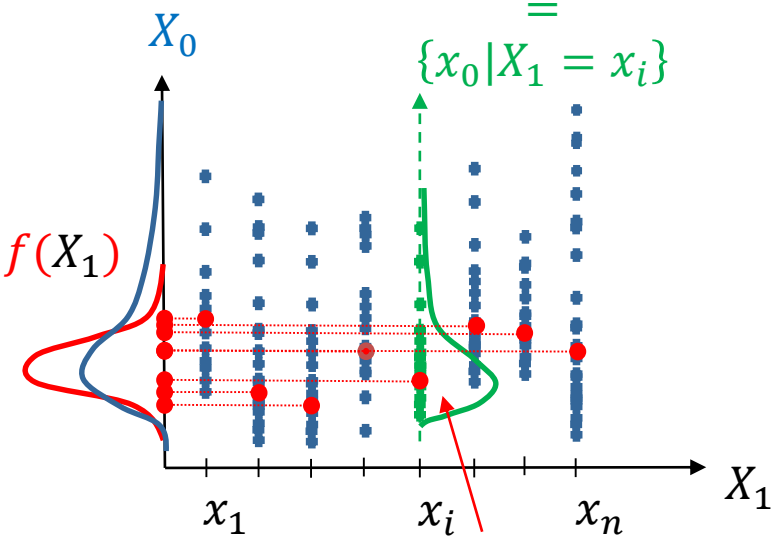


Conditional expectation: graphical definition

$X_0|X_1 = x_i \ (\neq \ X_0)$

$\equiv$

$\{x_0|X_1 = x_i\}$



$X_{0_c} = E(X_0|X_1) = f(X_1)$

Random variable

$E(X_0|X_1 = x_i)$

Scalar variable

Conditional expectation:
best estimate of a random variable by a function of (an)other random variable(s)

X_0 and X_1 two random variables.

$X_0^* = E(X_0|X_1)$ is the best approximation of X_0 by a *function* of X_1
best in the sens of minimum mean square difference $E\big((X_0 - X_0^*)^2\big)$

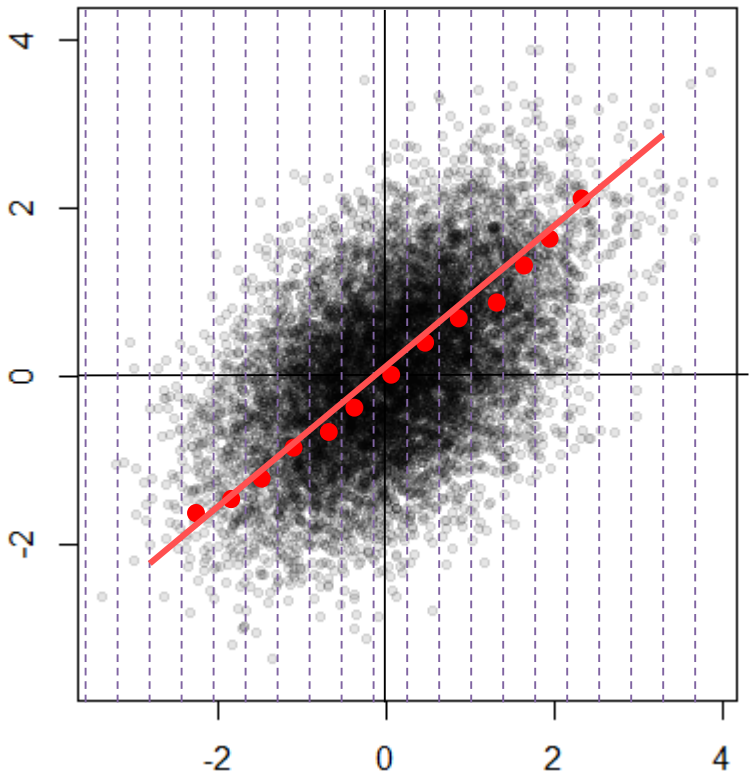
$$E\big((X_0 - E[X_0|X_1])^2\big) \leq E\big((X_0 - \varphi(X_1))^2\big)$$

As $E\big(E(X_0|X_1)\big) = E(X_0)$ (no bias), and we get that

$$var\big(X_0 - E(X_0|X_1)\big) \leq var(X_0 - \varphi(X_1))$$

$X_0{}^* = E(X_0|X_1)$ is a random variable (estimator)
$x_0{}^* = E(X_0|X_1 = x_1)$ is a scalar (estimation)

Conditional expectation: THE (bi)Gaussian case



Conditional expectation: THE (bi)Gaussian case and the linear model

If (X_0, X_1) is bi-gaussian, then the conditional expectation is linear: $X_0^* = E(X_0|X_1) = \lambda X_1 + \beta$

When the variable are centered, we can remove the intercept: $X_0{}^* = E(X_0|X_1) = \lambda X_1$

This can be generalized to several explanatory variable

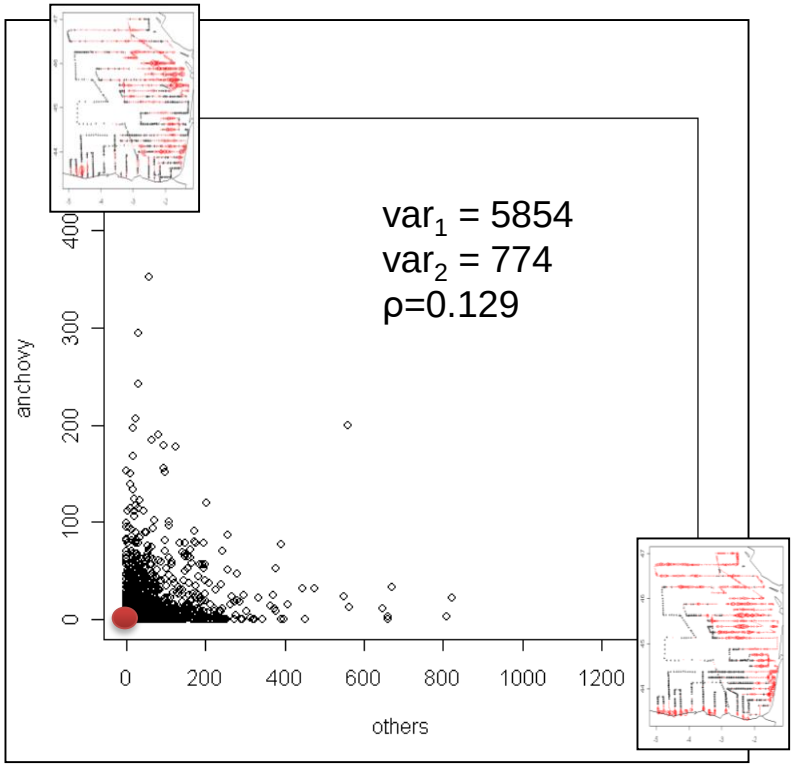
$$X_0{}^* = E(X_0|X_1, \dots, X_N) = \lambda_1 X_1 + \cdots + \lambda_N X_N = \sum_{i=1}^{N} \lambda_i X_i$$

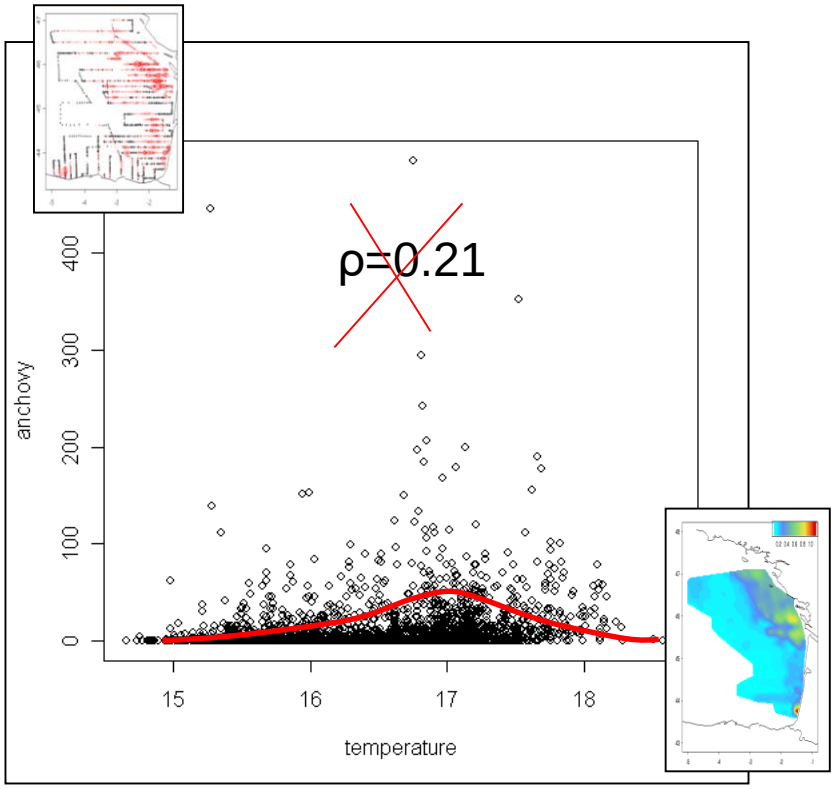
Nota: The explanatory variables not need to be independent

This assumes that $(X_0, X_1, \dots, X_N)$ is multi-Gaussian.

In all the other situations, i.e. in non-Gaussian situations, the linear solution is postulated and sub-optimal. One can however look for the Best Linear Unbiased Estimator (BLUE) eventhough sub-optimal.

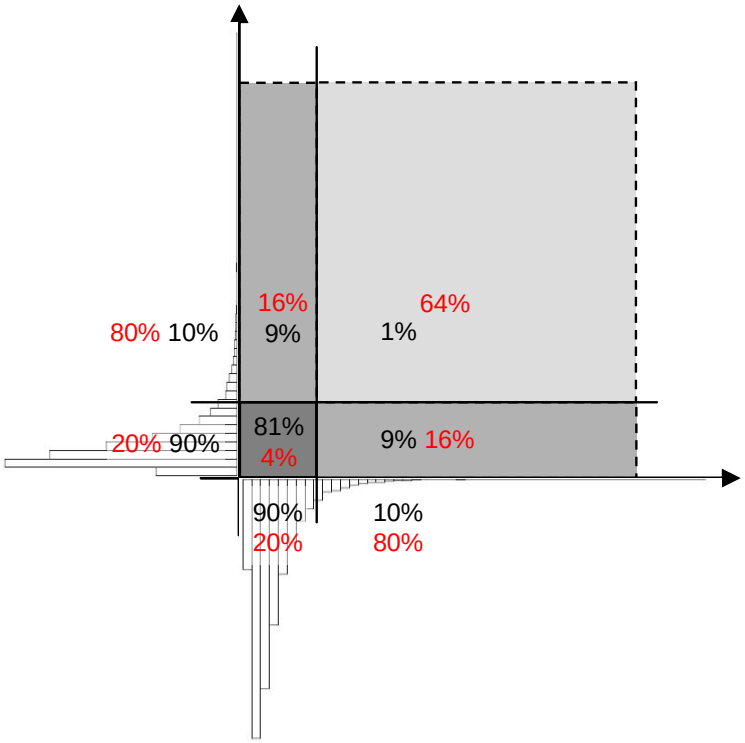
Remarques sur des lois non bigaussiennes et sur le coefficient de corrélation



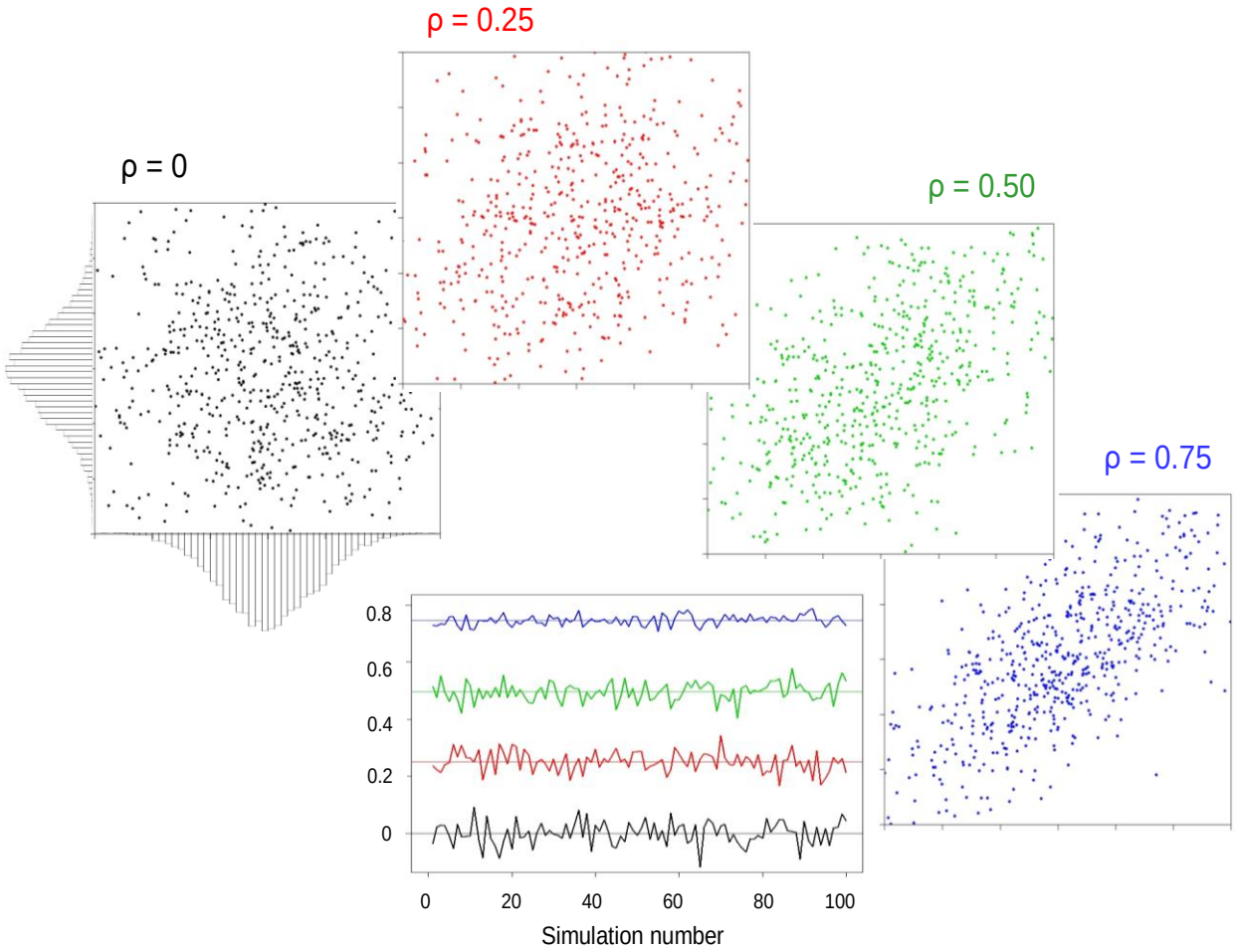


ρ=coefficient de corrélation = $\frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$

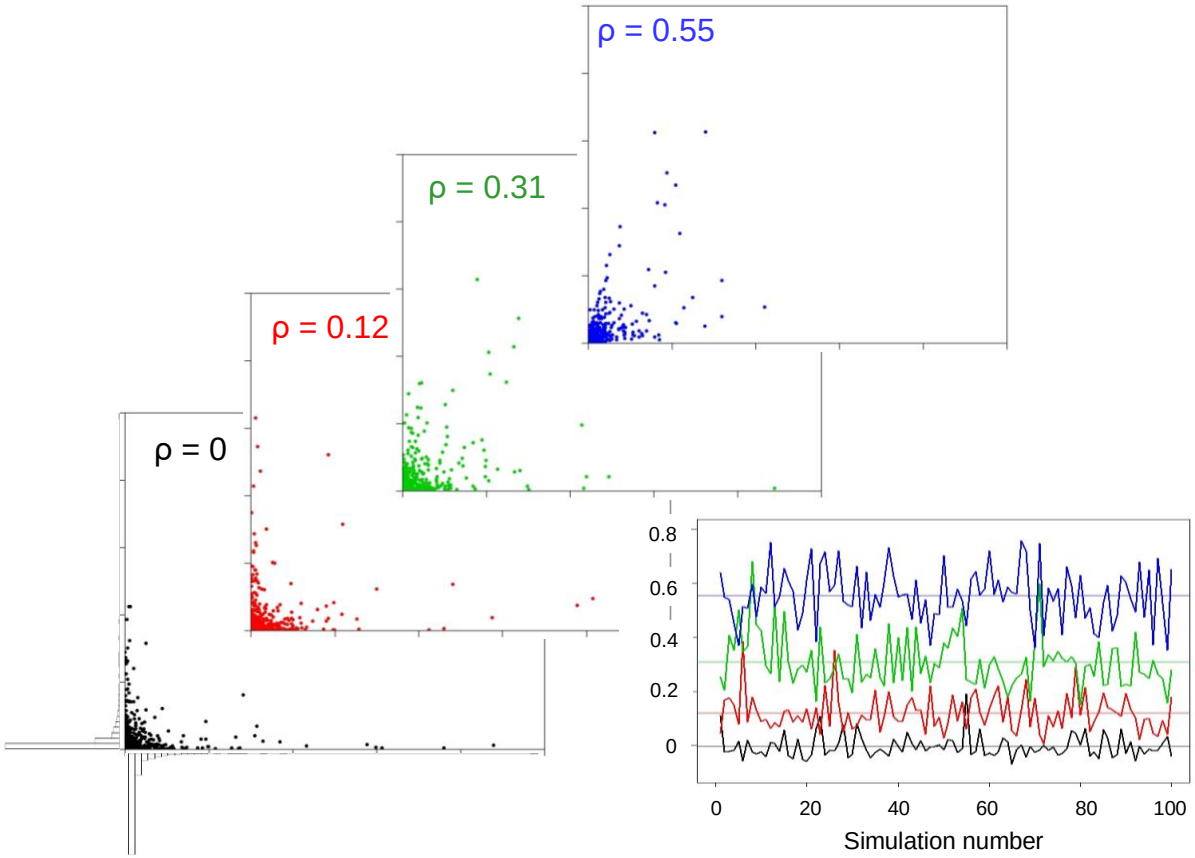
Outil symétrique en X et Y



Bi-lognormal scatter plot in case of independence (ro=0).
In black, the proportion of the values per interval (for each variable) or area (for the scatter plot).
In red, the relative importance of the intervals and of the areas.
The majority of the scatter plot (64%) accounts for a very small proportion of the points (1%).
Reciprocally, the majority of the scatter points (81%) are condensed in only 4% of the scatter plot area.



Scatter plots of pairs of Normal variables with increasing levels of linear correlations. The bottom panel represents the fluctuations observed for 100 simulations of each level of correlation



Scatter plots of pairs of log-normal variables with increasing levels of linear correlations. The bottom right panel represents the fluctuations observed for 100 simulations of each level of correlation. This gives an idea of the variation one may expect in practice when computing coefficient of correlation on real density variables.

Conditional expectation ; Linear model

We assume the conditional expectation is a linear function

$$X_0{}^* = E(X_0|X_1) \approx \lambda X_1 + \beta$$

Linear approximation of the approximation of X_0 by a function of X_1 !

When the variable are centered we can remove the intercept

$$X_0{}^* = E(X_0|X_1) \approx \lambda X_1$$

This can be generalized to several explanatory variable

$$X_0{}^* = E(X_0|X_1, \dots, X_N) \approx \lambda_1 X_1 + \cdots + \lambda_N X_N$$

$$X_0{}^* = E(X_0|X_1, \dots, X_N) \approx \sum_{i=1}^{N} \lambda_i X_i$$

Best Linear Unbiased Estimator (BLUE)

Simple centered monovariate case $X_0^* = E(X_0|X_1) \approx \lambda_1 X_1$

The estimation variance is equal to:

$$var(X_0 - X_0{}^*) = var(X_0 - \lambda_1 X_1) = var(X_0) + \lambda_1^2 var(X_1) - 2\lambda_1 cov(X_0, X_1) = F(\lambda_1)$$

Unknow is chosen so that to minimise the estimation variance.

Minimization ➔ (partial) derivative wrt to the unkown is null: $\frac{dF(\lambda_1)}{d\lambda_1} = 0$

$$\frac{dF(\lambda_1)}{d\lambda_1} = 2\lambda_1\, var(X_1) - 2cov(X_0, X_1) = 0$$

So: $\lambda_1 = \frac{cov(X_1, X_0)}{var(X_1)}$

Rq: the pdfs of X_0 and/or of X_1 must not be known/provided. Studying the residuals neither. The pdfs are only required for statistical testings.

Regression, covariance & variogram : two sides of the same coin

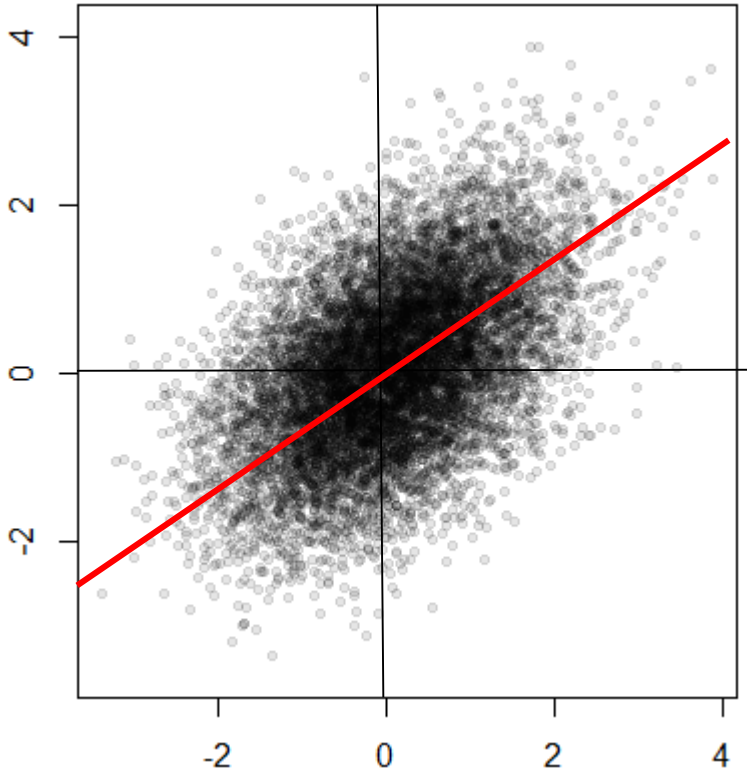
$$var(X_1 - X_0) = \sigma_{X_1}^2 + \sigma_{X_0}^2 - 2cov(X_1, X_0)$$
$$\cong 2\sigma^2 - 2cov(X_1, X_0)$$
$$\propto 1 - slope$$

Linear regression

$$\frac{X_0 - m_{X_0}}{\sigma_{X_0}} = \rho \frac{X_1 - m_{X_1}}{\sigma_{X_1}}$$

$$X_0 \propto \rho \frac{\sigma_{X_0}}{\sigma_{X_1}} X_1$$

$$slope = \frac{cov(X_1, X_0)}{\sigma_{X_1}^2}$$



$\propto var(X_1 - X_0)$
$\propto dissimilarity$

$\propto cov(X_1, X_0)$
$\propto similarity$

Slope ↑ when $cov(X_0, X_1)$ ↑ or when $var(X_0 - X_1)$ ↓

Best Linear Unbiased Estimator (BLUE): towards the general case

Bivariate case: $X_0^* = E(X_0|X_1, X_2) \approx \lambda_1 X_1 + \lambda_2 X_2 = \sum_{i=1}^{i=2} \lambda_i X_i$

Estimation error: $X_0^* - X_0 = \sum_{i=0}^{i=2} \lambda_i X_i$ with $\lambda_0 = -1$

Notations:
$C_{i,i} = cov(X_i, X_i) = var(X_i)$
$C_{i,j} = cov(X_i, X_i)$

Estimation variance:

$$var(X_0^* - X_0) = var\left(\sum_{i=0}^{i=2} \lambda_i X_i\right) = \sum_{i=0}^{2}\sum_{j=0}^{2} \lambda_i \lambda_j cov(X_i, X_j) = \sum_{i=0}^{2}\sum_{j=0}^{2} \lambda_i \lambda_j \, C_{i,j} = F(\lambda_1, \lambda_2)$$

Reminder:
$var(X_0 - X_1) = var(X_0) + var(X_1) - 2cov(X_0, X_1)$
$var(X_0 - X_1) = covar(X_0, X_0) + covar(X_1, X_1) - cov(X_0, X_1) - \mathrm{cov}(X_1, X_0)$
$\boldsymbol{var(\Sigma) = \Sigma\Sigma\, cov}$

Particular case:
$var(X_0 - X_1) = var(X_0) + var(X_1)$ if $X_0 \perp X_1$
$\boldsymbol{var(\Sigma) = \Sigma(var)}$ **if $\boldsymbol{X_i}$ mutually independent**

Best Linear Unbiased Estimator (BLUE): towards the general case

Minimizing the estimation variance: $\frac{\partial F(\lambda_1,\lambda_2)}{\partial \lambda_1} = \frac{\partial F(\lambda_1,\lambda_2)}{\partial \lambda_2} = 0$

$$\left.\begin{array}{l} \frac{\partial F(\lambda_1,\lambda_2)}{\partial \lambda_1} \Rightarrow \lambda_2 C_{1,2} + \lambda_1 C_{1,1} = C_{1,0} \\ \frac{\partial F(\lambda_1,\lambda_2)}{\partial \lambda_2} \Rightarrow \lambda_1 C_{2,1} + \lambda_2 C_{2,2} = C_{2,0} \end{array}\right\} \text{2 unknows; 2 linear equations}$$

System: $\begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} C_{1,0} \\ C_{2,0} \end{bmatrix}$

Inversion: $\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} C_{1,1} & C_{1,2} \\ C_{2,1} & C_{2,2} \end{bmatrix}^{-1} \cdot \begin{bmatrix} C_{1,0} \\ C_{2,0} \end{bmatrix} = \frac{1}{C_{1,1}C_{2,2} - C_{1,2}^2} \begin{bmatrix} C_{2,2} & -C_{1,2} \\ -C_{2,1} & C_{1,1} \end{bmatrix} \cdot \begin{bmatrix} C_{1,0} \\ C_{2,0} \end{bmatrix}$

Best Linear Unbiased Estimator (BLUE): towards the general case

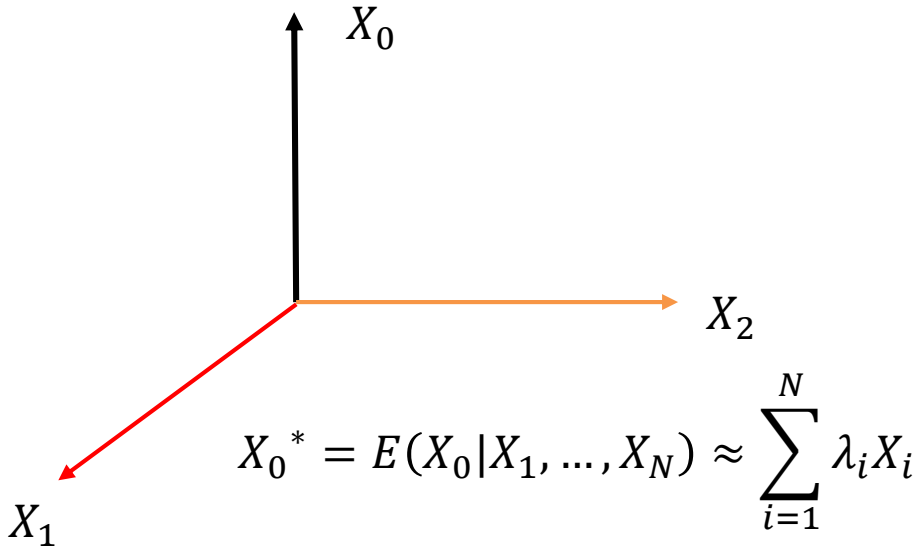
Nota 1: relation between variables ares considered 2 by 2 and, because we are interested in linear regressions, the covariances suffice to calibrate the regression.

Nota 2: X_1 and X_2do not need to be independent. One can use covariates that are linearly dependent to make a regression. However, the realizations $(x_{1,i}, x_{2,i}, y_i)$ have to be independent outcomes of (X_1, X_2, Y) to correctly estimate the covariances by mean square errors.

Nota 3: no parametric assumptions required to compute $\lambda_1$and λ_2. There are needed to perform statistical tests in the model.

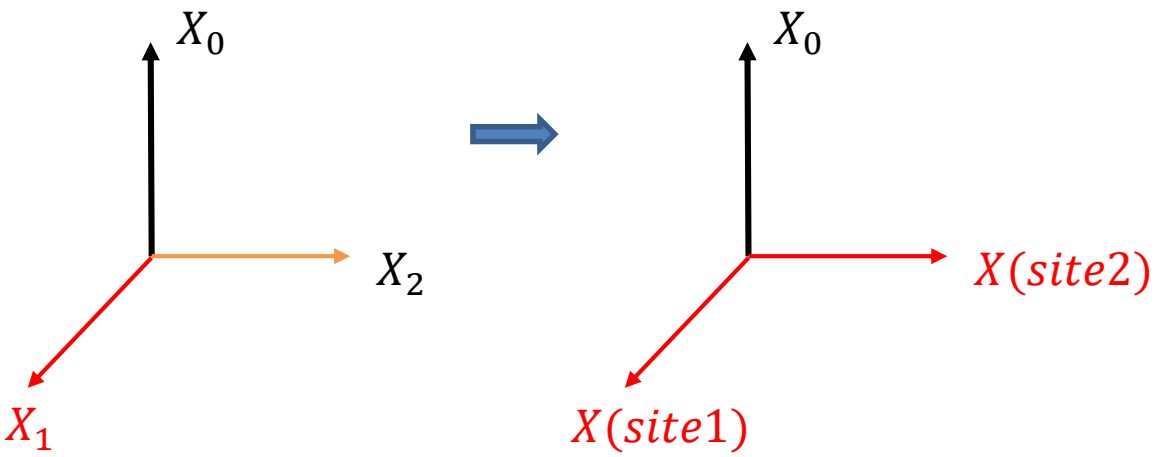
Nota 4: 2 unknows ; 2 linear equations.
Generalization comes easily: regression with N variables leeds to a system with N equations, with an $N \times N$ matrix to invert.

Make it spatial: from Linear Models to Kriging



$$X_0{}^* = E(X_0|X_1, \ldots, X_N) \approx \sum_{i=1}^{N} \lambda_i X_i$$

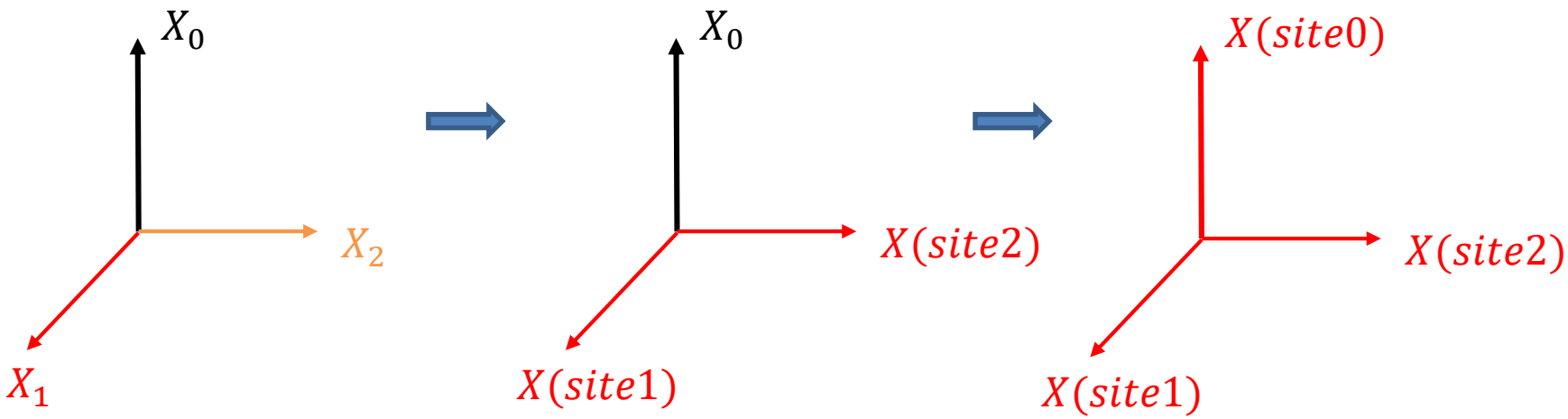
From Linear Models to Kriging



$$X_0{}^* = E(X_0|X_1, \ldots, X_N) \approx \sum_{i=1}^{N} \lambda_i X_i$$

Explanatory variables are the same variable at different geographical locations.

From Linear Models to Kriging

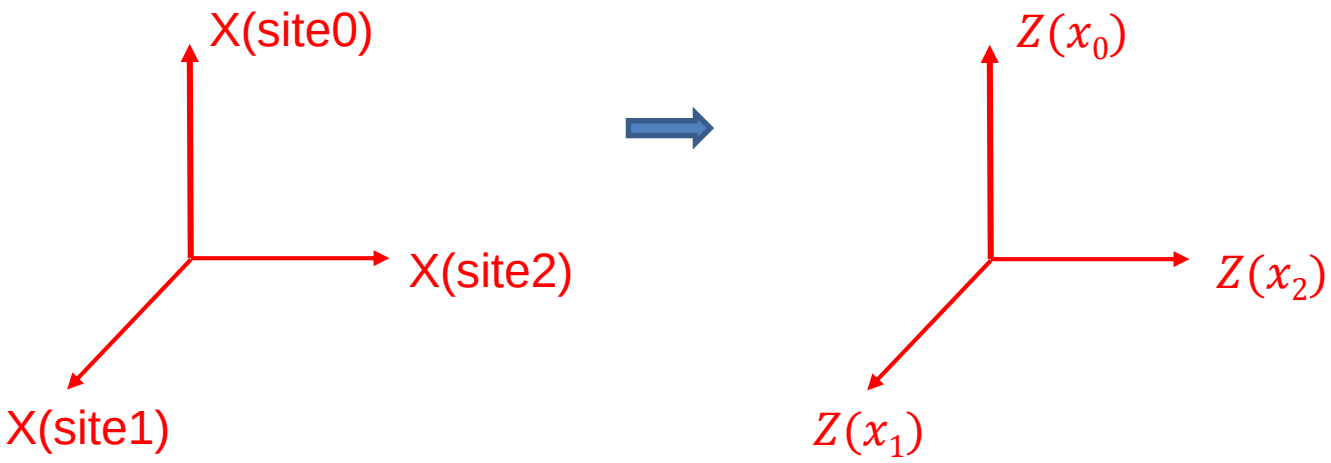


$$Y^* = E(Y|X_1, \dots, X_N) \approx \sum_{i=1}^{N} \lambda_i X_i$$

Explanatory variables are the same variable at different geographical locations.

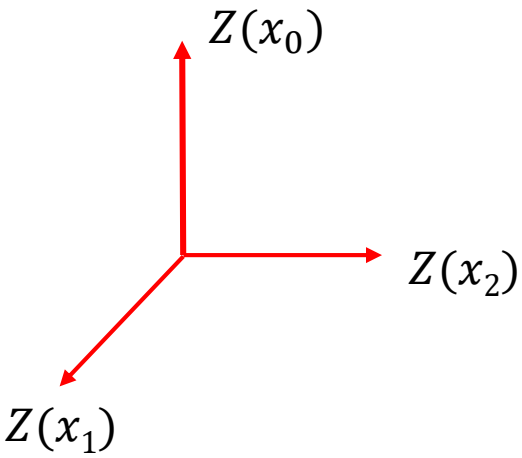
The explained variable is also of the same type than the explanatory variables.

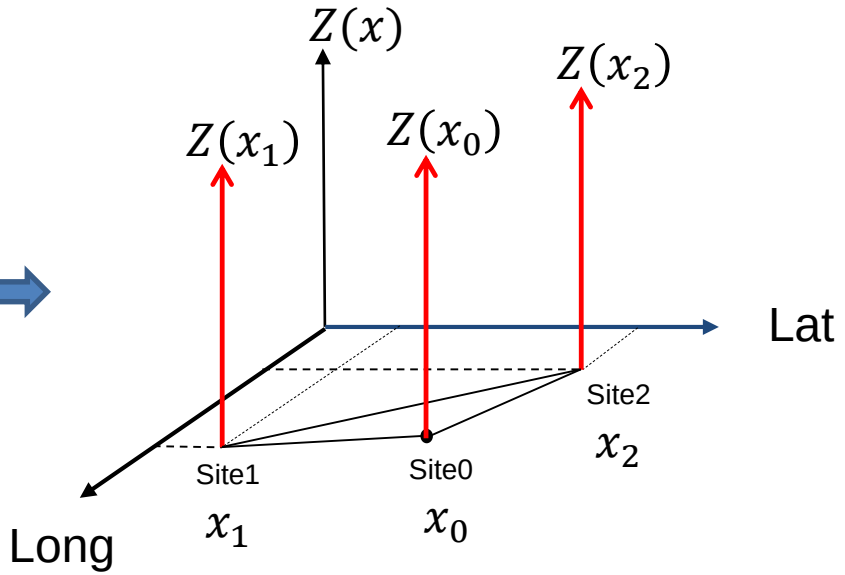
From Linear Models to Kriging



Geostatistical notations.

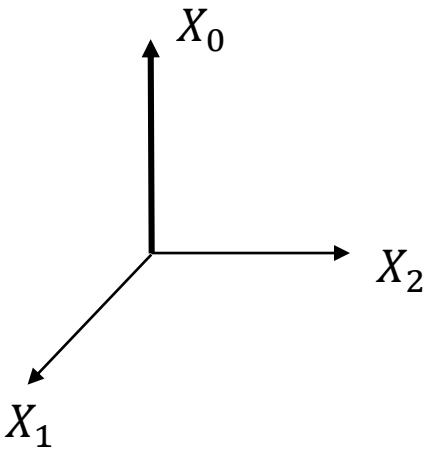
From Linear Models to Kriging

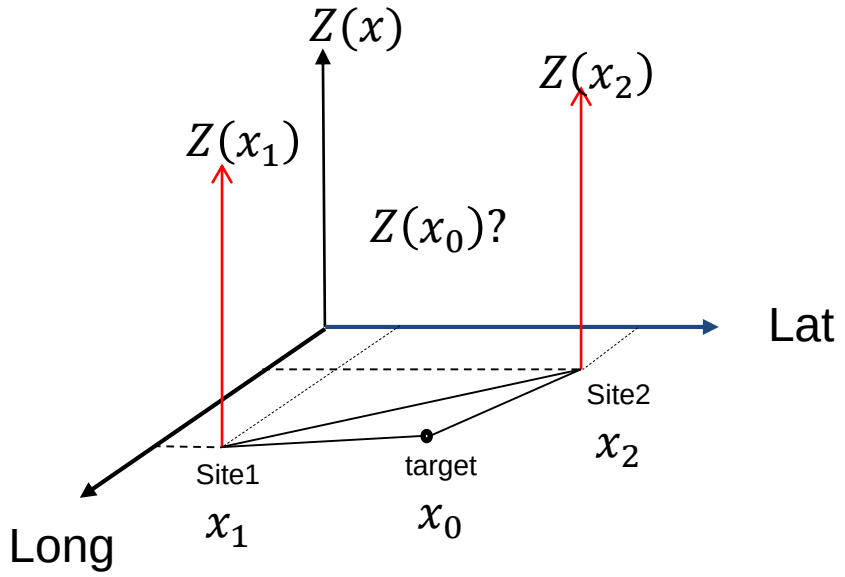




Nota: $Z(x)$ is now a random function
a random vector of infinite dimension,
one random variable per location

From Linear Models to Kriging

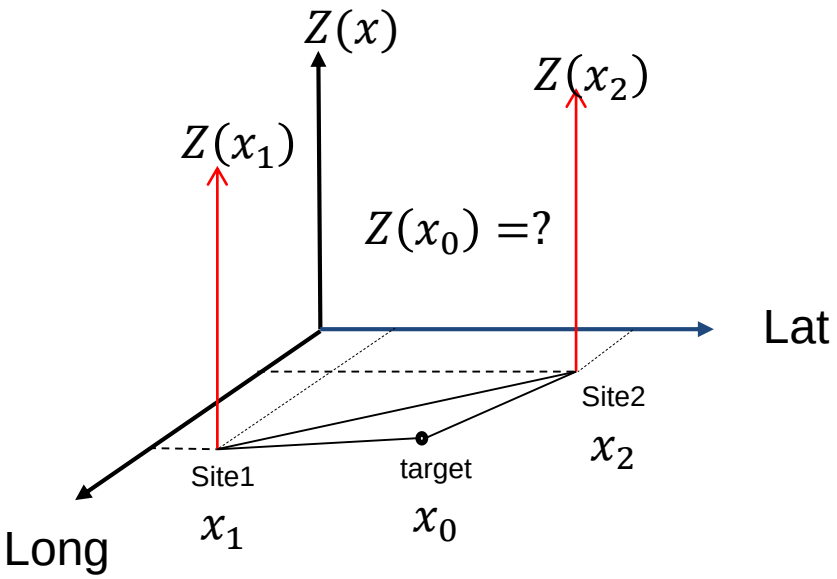




$$X_0{}^* = E(X_0|X_1, \dots, X_N) \approx \sum_{i=1}^{N} \lambda_i X_i$$

$$Z_0^{\boldsymbol{Kriging}} = E(Z_0|Z_1, \dots, Z_N) \approx \sum_{i=1}^{N} \lambda_i Z_i$$

Ponctual Kriging



$$Z_0^{\boldsymbol{K}} = E(Z_0|Z_1, \dots, Z_N) \approx \sum_{i=1}^{N} \lambda_i Z_i$$

Similar but different …

We are no longer explaining Z_0 by a function of the Z_i, but predicting it.

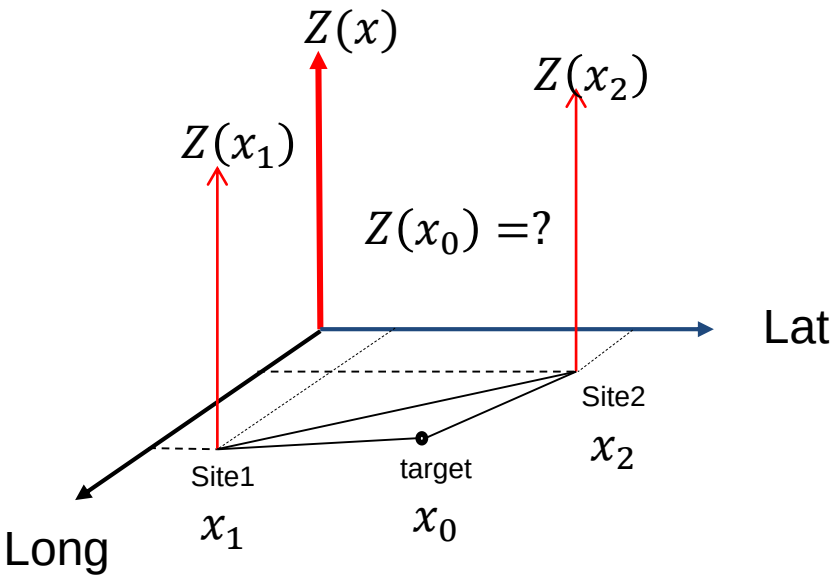
The objective is not to estimate λ_i but to chose λ_i that best estimate Z_0

Same framework than linear models
Same algorithm than linear models
➔ minimum estimation variance (BLUE)
➔ partial derivatives of the estimation variance

Estimator: $Z_0^{\boldsymbol{K}} = E(Z_0|Z_1, \dots, Z_N)$

Estimation: $z_0^{\boldsymbol{K}} = E(Z_0|Z_1 = z_1, \dots, Z_N = z_N)$

Ponctual Kriging



$$Z_0^{\boldsymbol{K}} = E(Z_0|Z_1, \dots, Z_N) \approx \sum_{i=1}^{N} \lambda_i Z_i$$

In multivariate linear regression:

$$\lambda_1 = \frac{C_{1,0}C_{2,2} - C_{1,2}C_{2,0}}{C_{1,1}C_{2,2} - C_{1,2}^2}$$

the solution is known as long as the covariances of all pairs of variables are known.

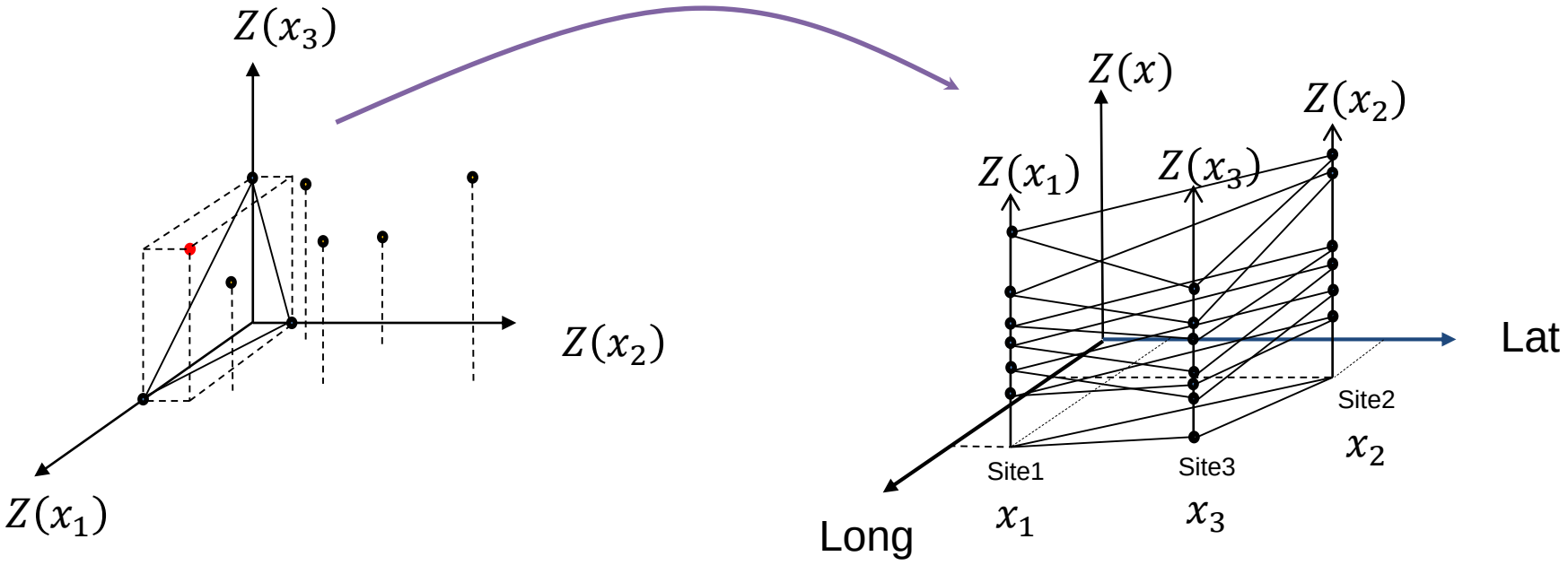
In the spatial context, to perform kriging we need to know all the spatial correlations between samples points

$$cov(Z_i, Z_j)$$

and all the spatial correlation between target and samples points

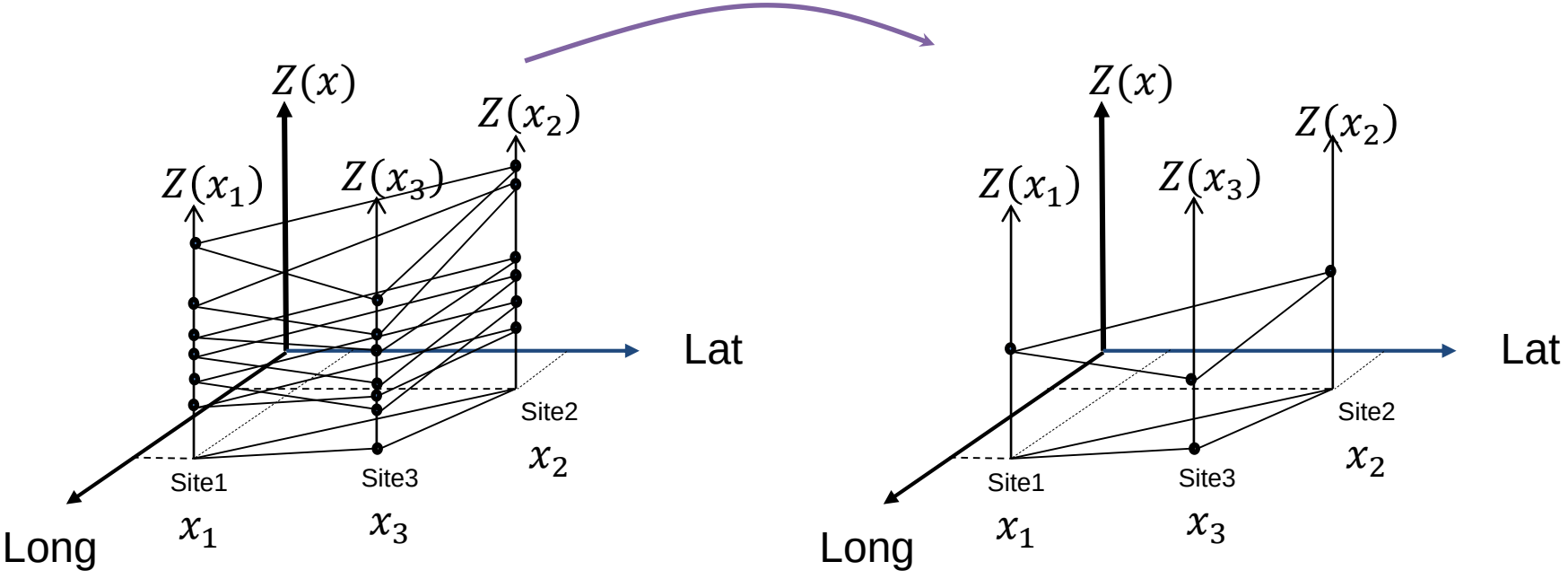
$$cov(Z_i, Z_0)$$

What is available for real ?

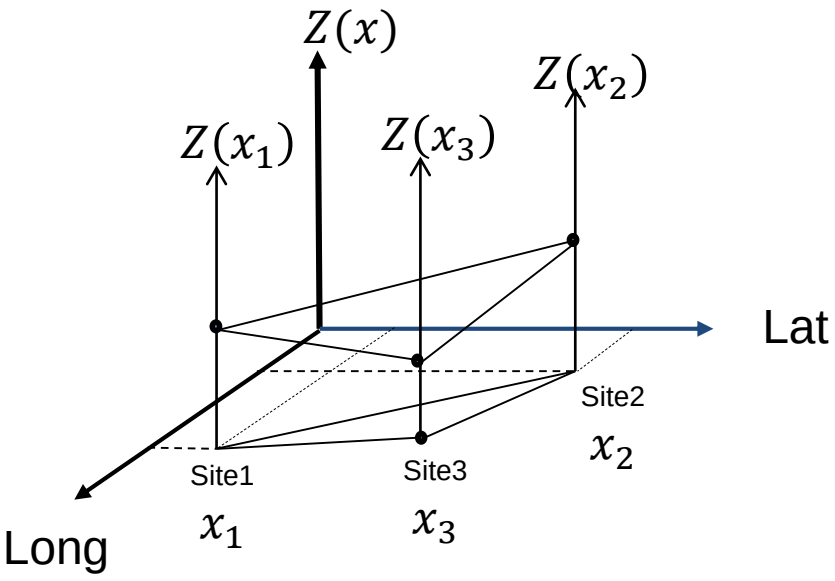

$Z(x_3)$
$Z(x_2)$
$Z(x_1)$
$Z(x)$
$Z(x_2)$
$Z(x_1)$
$Z(x_3)$
Lat
Site2
x_2
Site1
x_1
Site3
x_3
Long

From several …

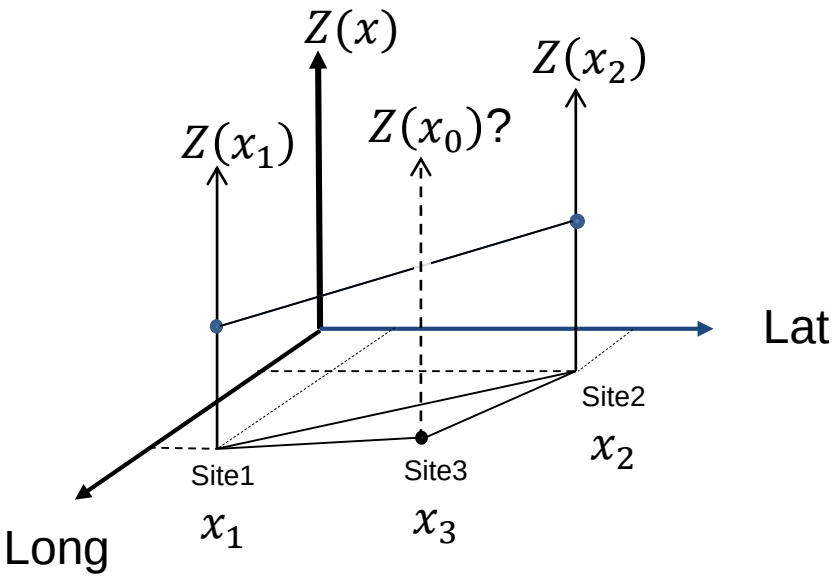
… to one realization per variable



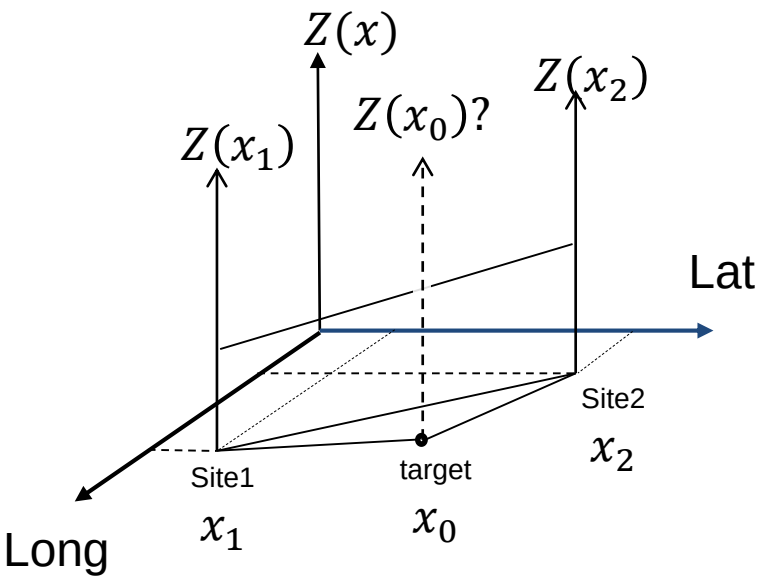
From regression $Z_3|Z_1, Z_2$ …



… to prediction $Z_0|Z_1, Z_2$



Ponctual Kriging



Spatial correlations between samples points

➔ (z_i, z_j) is observed once.

➔ Can not access to $cov(Z_i, Z_j)$ experimentally

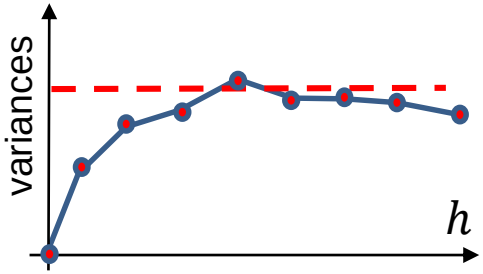
➔ Repetition is coming from other pairs *elsewhere*

➔ **Stationarity = unavoidable assumption**

$$Z_0^{\boldsymbol{K}} = E[Z_0|Z_1, \dots, Z_N] \approx \sum_{\alpha=1}^{N} \lambda_\alpha Z_\alpha$$

Spatial covariance and variogram: first steps from data to model

Reminder: $\gamma = \frac{1}{2}\overline{(\Delta z)^2}$



If#1randomization $z(x_i)$➔ $Z(x_i) = Z_i$

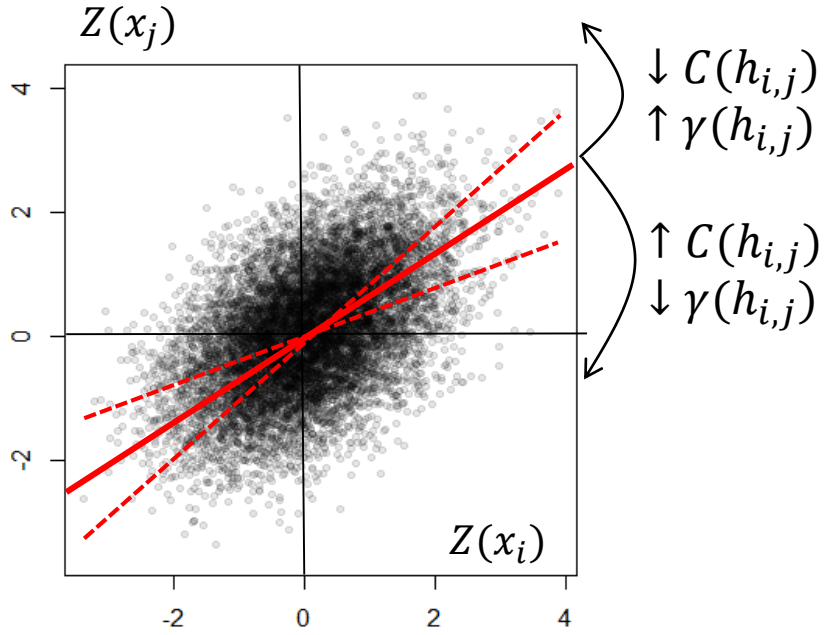
If#2 assumption $E(Z_i - Z_j) = 0$

Then: $\text{var}(Z_i - Z_j) = \text{E}\left((Z_i - Z_j)^2\right) = 2 \cdot \gamma(h_{i,j})$

$var(Z_i - Z_j) = var(Z_i) + var(Z_j) - 2cov(Z_i, Z_j)$

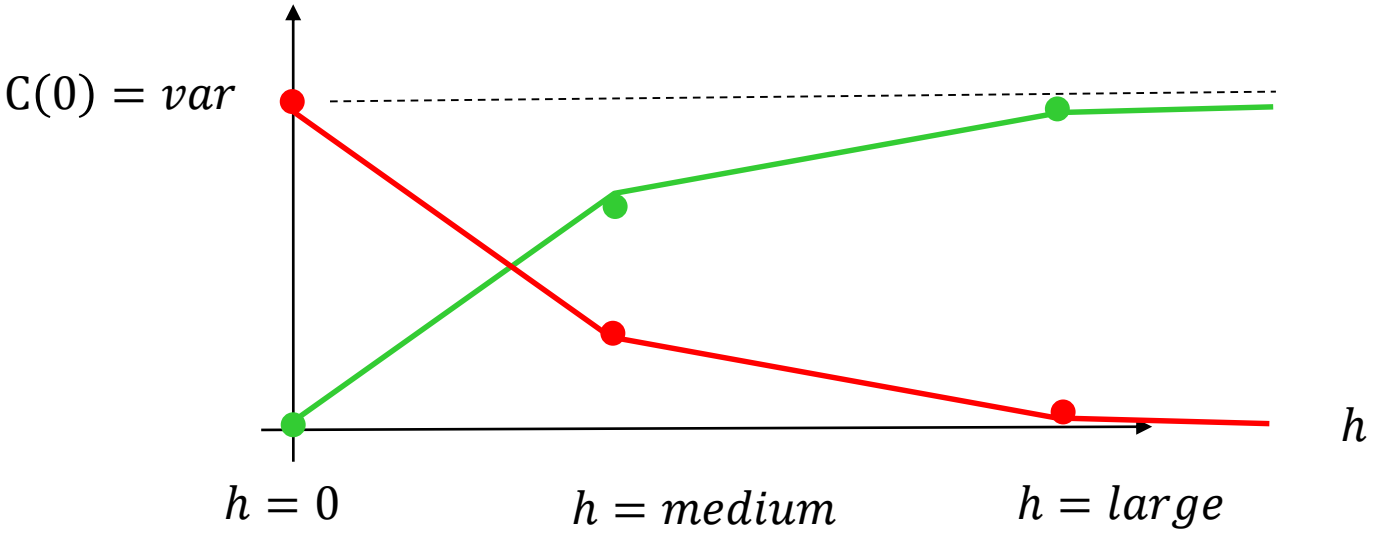
$var(Z_i - Z_j) = 2var(Z) - 2cov(Z_i, Z_j)$

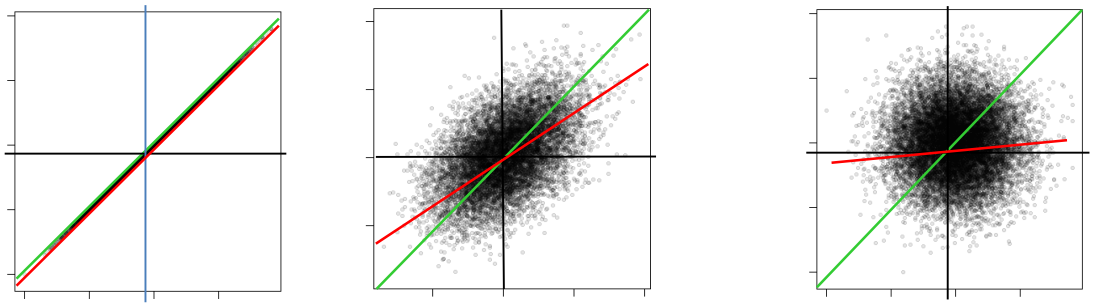
$\gamma(h_{i,j}) = C(0) - C(h_{i,j})$



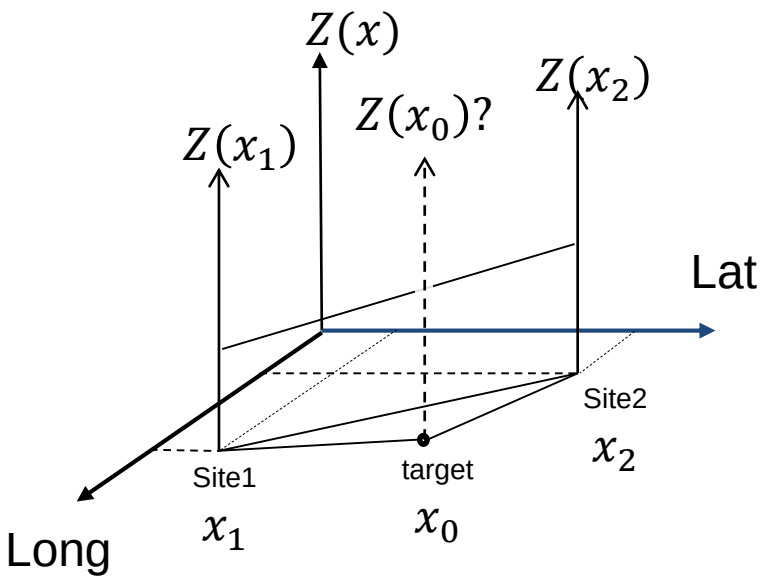
Variogram and covariance

$$\gamma(h) = C(0) - C(h)$$





Ponctual Kriging



Spatial correlations between samples points

➔ (z_i, z_j) is observed once.

➔ Can not access to $cov(Z_i, Z_j)$ experimentally

➔ Repetition is coming from other pairs *elsewhere*

➔ **Stationarity = unavoidable assumption**

Spatial correlation between target and samples points

➔ (z_i, z_0) is NOT observed.

➔Can not access to $cov(Z_i, Z_0)$

➔ Need for a model of stationary spatial covariance

$$Z_0^{\boldsymbol{K}} = E[Z_0|Z_1, \dots, Z_N] \approx \sum_{\alpha=1}^{N} \lambda_\alpha Z_\alpha$$

Construction de champs aléatoires gaussiens sous R et étude de leurs variogrammes

Approche analytique de processus aléatoires (1D) :

le processus auto-régressif d'ordre1 et la marche aléatoire

AutoRegressive process of order1

$$Z(x+1) = \rho Z(x) + \sqrt{1-\rho^2} \cdot U_{x+1}$$

with

- $\rho < 1$
- $U_x \; i.i.d., E[U_x] = 0, var(U_x) = 1$
- $Z(1) = U_1$

Exercice : verify that:

- $E\big(Z(x)\big) = 0$
- $var\big(Z(x)\big) = 1$
- $cov(Z(x), Z(x+h) = \rho^h$
- $\gamma(h) = \frac{1}{2} var(Z_x - Z_{x+h}) = 1 - \rho^h$

AutoRegressive process of order1

$$E[Z_{x+1}] = \rho E[Z_x] + \sqrt{1-\rho^2} E[U_{x+1}] = \rho E[Z_x]$$

Initialization

$E[Z(1)] = E[U_1] = 0$

Heridity

$E[Z_x] = 0$

Expected value is constant and thus

independent of x ➔ stationarity of (moment of) order 1

AutoRegressive process of order1

$$var[Z_{x+1}] = \rho^2 var[Z_x] + (1-\rho^2) var[U_{x+1}] + 2\rho\sqrt{1-\rho^2}\, cov(Z_x, U_{x+1})$$

The U_x being independent ➔ $cov(Z_x, U_{x+1}) = 0$

Then:

$$var[Z_{x+1}] = \rho^2 var[Z_x] + (1-\rho^2)$$

Initialization

$$var[Z_2] = \rho^2 var[Z_1] + (1-\rho^2) var[U_2] = \rho^2 + (1-\rho^2) = 1$$

Heridity

$$var[Z_x] = 1$$

Variance is constant and thus

independent of x ➔ stationarity of (moment of) order 2

AutoRegressive process of order1

$$covar(Z_x, Z_{x+1}) = covar\left(Z_x, \rho Z_x + \sqrt{1-\rho^2}U_{x+1}\right)$$
$$= \rho \cdot covar(Z_x, Z_x) + \sqrt{1-\rho^2}covar(Z_x, U_{x+1})$$
$$= \rho$$

Heridity

$$covar(Z_x, Z_{x+h}) = \rho^h$$
$$covar(Z_x, Z_{x+h+1}) = covar\left(Z_x, \rho Z_{x+h} + \sqrt{1-\rho^2}U_{x+h+1}\right)$$
$$= \rho \cdot covar(Z_x, Z_{x+h}) + \sqrt{1-\rho^2}covar(Z_x, U_{x+h+1})$$
$$= \rho^{h+1}$$

$$\Longrightarrow\ covar(Z_x, Z_{x+h}) = \rho^h = C(h)$$

For an AR(1),
the covariance only depends on h and is independent of x ➔ stationarity of the covariance

AutoRegressive process of order1

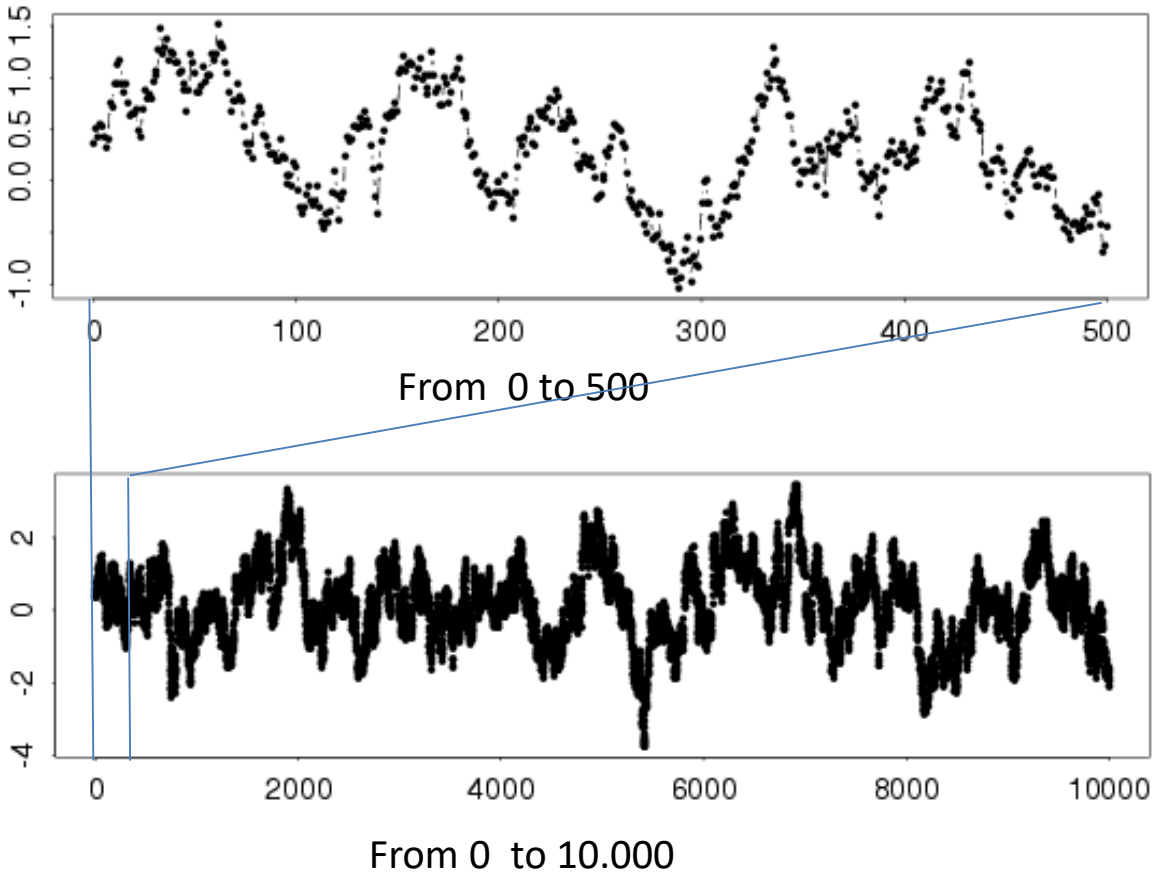
$$var(Z_x - Z_{x+h}) = var\big(Z(x)\big) + var\big(Z(x+h)\big) - 2covar(Z_x, Z_{x+h})$$
$$= 2 - 2\rho^h$$

So
$$\frac{1}{2}var(Z_x - Z_{x+h}) = 1 - \rho^h = \gamma(h)$$

For an AR(1),
the variogram only depends on h and is independent of x ➔ stationarity of the variogram

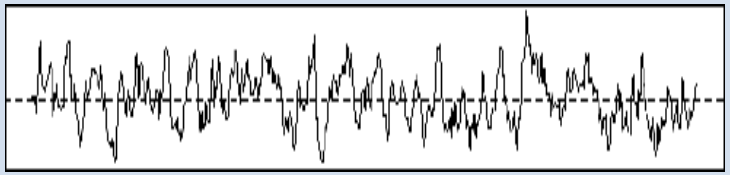
Rq: the pdf of U_x and thus of $Z(x)$ has not been specified. What matters are only the moments of order 1 and 2.

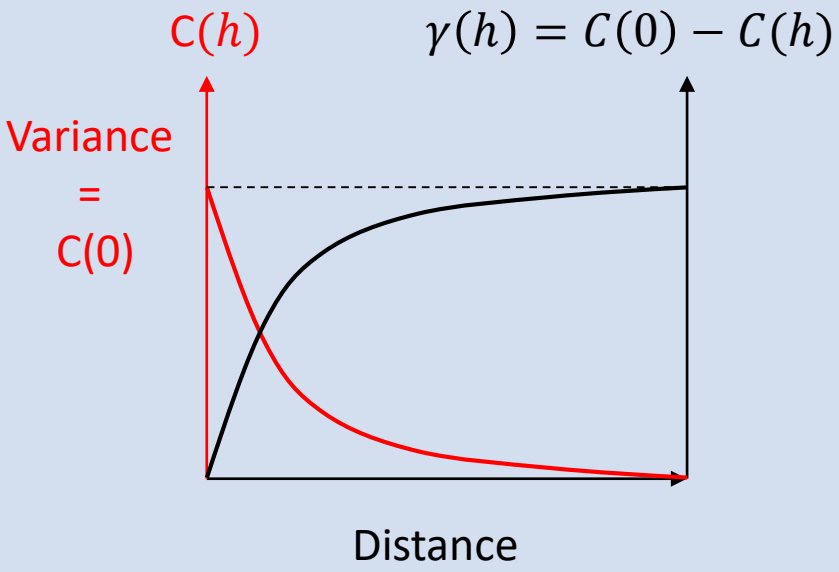
Stationary case: AR1



AR(1)

AR1





$(for\ 0 < \rho < 1)$

Random walk

$x = 1, 2, \ldots$

$U_x = +1 \; or \; -1$ with probability $p = 0.5$

$U_x \; i.i.d., E[U_x] = 0, var(U_x) = 1$

$$Z(x) = \sum_{i=1}^{x} U_i$$

Exercice : verify that:

- $E\big(Z(x)\big) = 0$
- $var\big(Z(x)\big) = x$
- $cov(Z(x), Z(x+h) = \rho^h$
- $\gamma(h) = \frac{1}{2} var(Z_x - Z_{x+h}) = \frac{h}{2}$

Random walk

$$E[Z_x] = E\left[\sum_{i=1}^{x} U_i\right] = \sum_{i=1}^{x} E[U_i] = 0$$

Expected value is constant ➔ stationarity of (moment of) order 1

$$var[Z_x] = var\left[\sum_{i=1}^{x} U_i\right] = \sum_{i=1}^{x}\sum_{j=1}^{x} cov(U_i, U_j) = \sum_{i=1}^{x} cov(U_i, U_i)$$

So

$$var[Z_x] = x$$

Variance is NOT constant and depends on x ➔ NON stationarity of (moment of) order 2

Random walk

$$covar[Z_x, Z_{x+h}] = covar\left(Z_x, Z_x + \left[\sum_{i=x+1}^{x+h} U_i\right]\right) = var(Z_x) + \sum_{i=x+1}^{x+h} cov(Z_x, U_i)$$

So

$$covar[Z_x, Z_{x+h}] = x$$

Covariance is depends on x ➔ NON stationarity of the covariance

$$var[Z_x - Z_{x+h}] = var\left[\sum_{i=x+1}^{x+h} U_i\right] = \sum_{i=x+1}^{x+h} cov(U_i, U_i)$$

So

$$var[Z_x - Z_{x+h}] = h$$

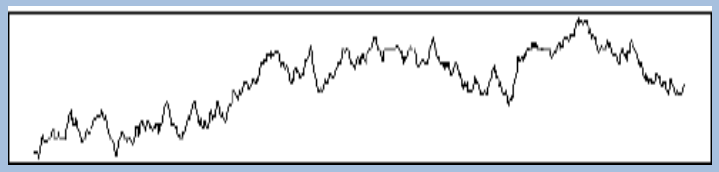
And

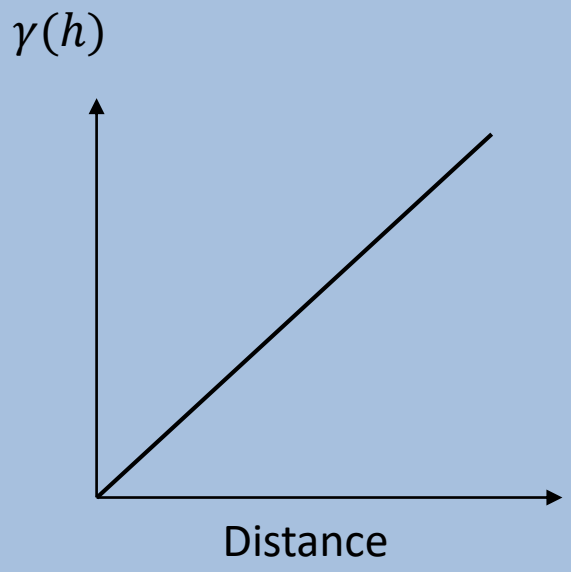
$$\gamma(h) = \frac{1}{2} \cdot var[Z_x - Z_{x+h}] = \frac{h}{2}$$

Variogram is not depends on x ➔ stationarity of the variogram

Random walk

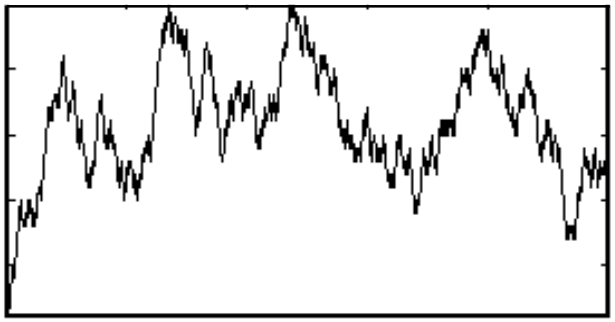
Random walk



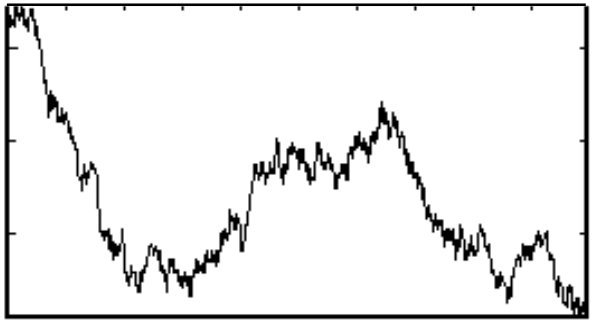


Intrinsic case

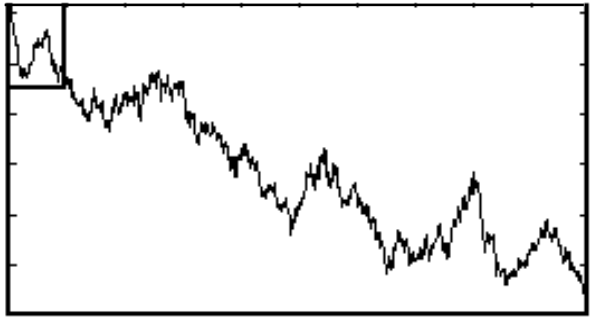
RF with a linear variogram : random walk with -1 and +1 valuations



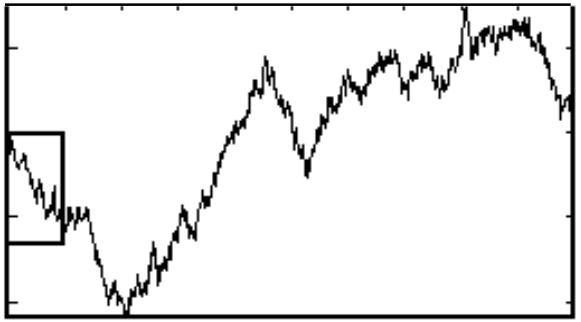
From 0 to 500



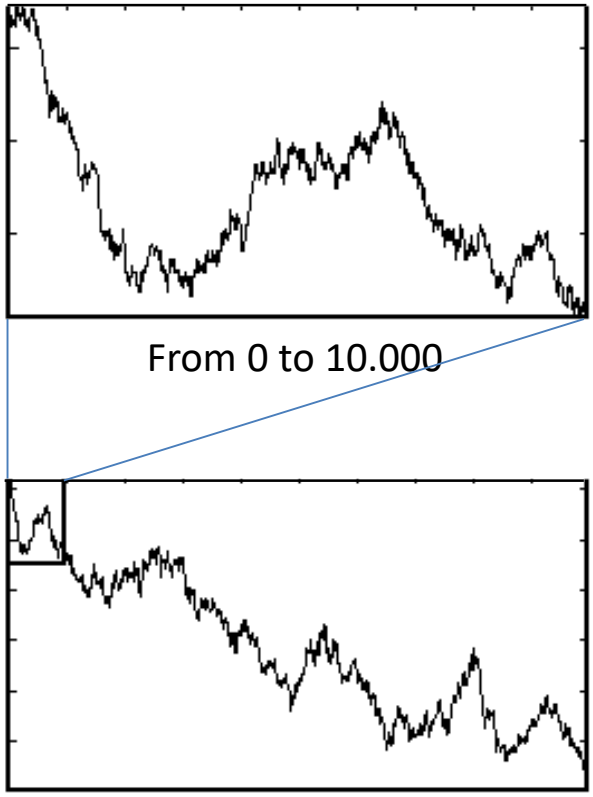
From 0 to 10.000



From 0 to 100.000



From 0 to 1.000.000


From 0 to 10.000
From 0 to 100.000

Generalization to continuous space:

the random functions

Random Function

A random function (RF) $Z = \left(Z(x), x \in R^2\right)$ is an infinite family of random variables.

A random variable is implemented everywhere. So a RF is defined by

- the characteristics of the Random Variables: $E(Z_x), \mathrm{var}(Z_x)$
- the correlations between them, 2 by 2 : $\mathrm{cov}(Z_x, Z_y)$

  3 by 3 :

  etc ...

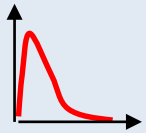
Spatial law: distributions of all finite possible combinations:

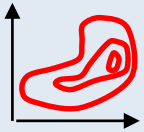
$$F_{x_1,\ldots,x_n}(z_1,\ldots,z_n) = P\left\{Z(x_1) < z_1,\ldots,Z(x_n) < z_n\right\}$$

$$\forall n, \quad \forall x_1,\ldots,x_n, \quad \text{and} \quad \forall z_1,\ldots,z_n$$

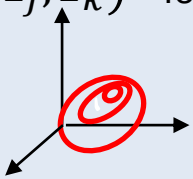
Spatial probability law

$P\{Z_{x_1} < z_1, \dots, Z_{x_n} < z_n\}$
$\forall n, \forall x_1, \dots, x_n$

$F_{x_i}(z_i)$  for all x_i

$F_{x_i,x_j}(z_i, z_j)$  for all x_i, x_j

$F_{x_i,x_j,x_k}(z_i, z_j, z_k)$ for all x_i, x_j, x_k



.....

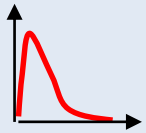
Moments of orders 1 & 2

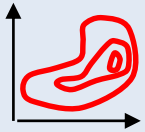
The full (spatial) law can be reduced down to the first two moments. In particular :

$$E\big(Z(x)\big)$$
$$var\big(Z(x)\big)$$
$$covar\big(Z(x), Z(y)\big) = C(x, y)$$

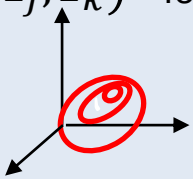
Spatial probability law → **Bivariate probability law** → **Order 2 moments**

$P\{Z_{x_1} < z_1, \dots, Z_{x_n} < z_n\}$
$\forall n, \forall x_1, \dots, x_n$

$F_{x_i}(z_i)$  for all x_i

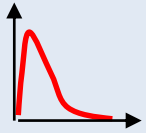
$F_{x_i,x_j}(z_i, z_j)$  for all x_i, x_j

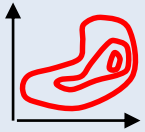
$F_{x_i,x_j,x_k}(z_i, z_j, z_k)$ for all x_i, x_j, x_k



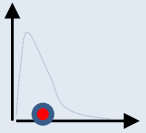
.....

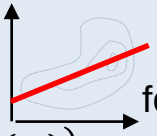
$P\{Z_{x_1} < z_1, Z_{x_2} < z_2\}$
$\forall x_1, x_2$

$F_{x_i}(z_i)$  for all x_i

$F_{x_i,x_j}(z_i, z_j)$  for all x_i, x_j

$E(Z_{x_1})$; $cov(Z_{x_1}, Z_{x_2})$
$\forall x_1, x_2$

Order 1
$E[Z(x_i)]$  for all x_i

Order 2
$\mathrm{cov}(Z(x_i), Z(x_j))$  for all x_i, x_j

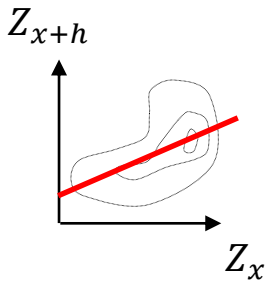
Stationarity (of order 2)

A random function is stationary when its spatial law is independent of the location.
In particular :

$$\left.\begin{array}{l} E(Z_x) = E(Z_y) = m \\ Var(Z_x) = \mathrm{var}(Z_y) = \sigma^2 \\ \mathrm{cov}(Z_x, Z_y) = C(x-y) = C(h) \end{array}\right\} \rightarrow \text{Order 2 stationarity}$$

...

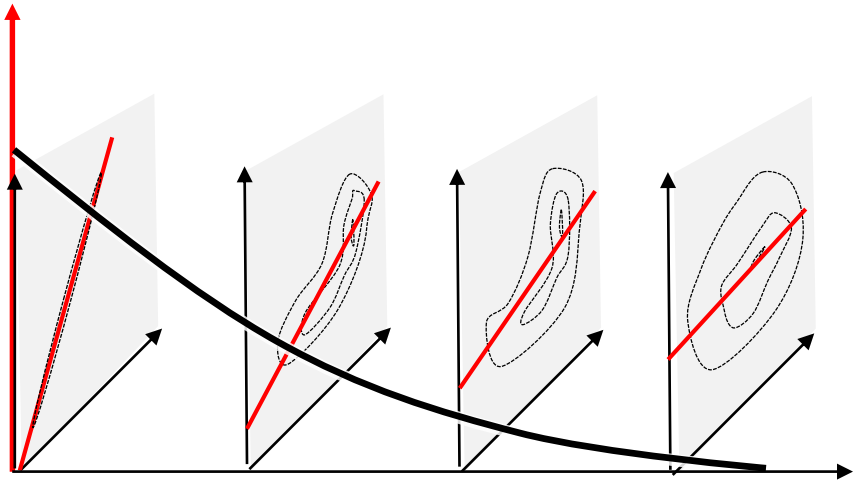
Spatial covariance



$$slope = \frac{cov(Z_x, Z_{x+h})}{C(0)} = \frac{C(h)}{C(0)}$$

$C(h) = cov(Z_x, Z_{x+h})$

$C(0) = cov(Z_x, Z_x) = var(Z_x)$



distance (h)

Intrinsic Random Function (IRF):
a random function $Z(x)$ is an
intrinsic random function (of order 2)
if and only if
its **increments $Z(x) - Z(x+h)$ are stationarity** (of order 2)

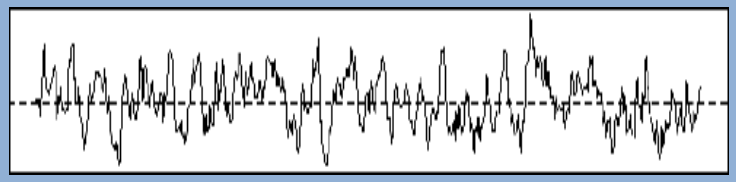
$$E(Z_x - Z_{x+h}) = 0$$

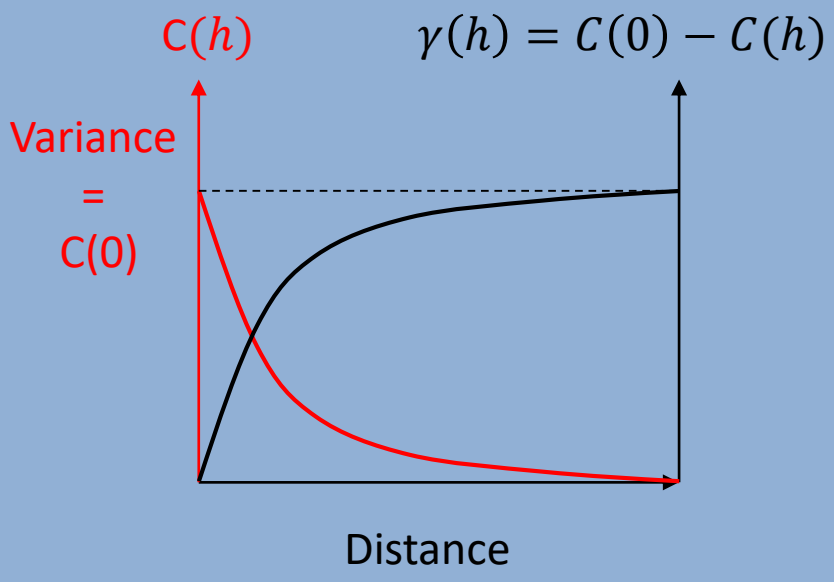
so that

$$\text{var}(Z_x - Z_{x+h}) = E\left((Z_x - Z_{x+h})^2\right) = 2\gamma(h)$$

Stationary Random Function (SRF)

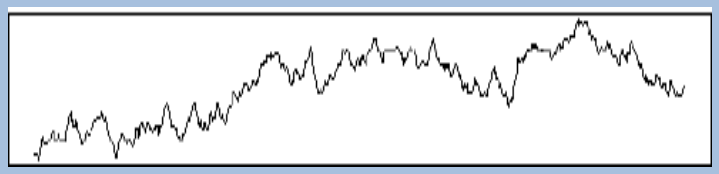
AR1

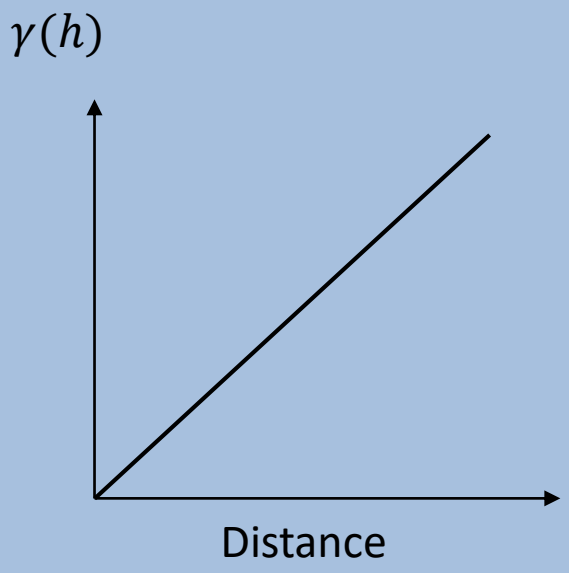




Intrinsic Random Function (IRF)

Random walk





Cas des F.A. stationnaires
avec covariances spatiales

$$Cov\big(Z(x_i), Z(x_j)\big) = C\big(x_i - x_j\big) = C_{i,j}$$

$$var\left(\sum_i \lambda_i Z(x_i)\right) = \sum_i \sum_j \lambda_i \lambda_j C_{i,j}$$

Cas des F.A. intrinsèques
avec variogrammes

$$\gamma\big(Z(x_i), Z(x_j)\big) = \gamma\big(x_i - x_j\big) = \gamma_{i,j}$$

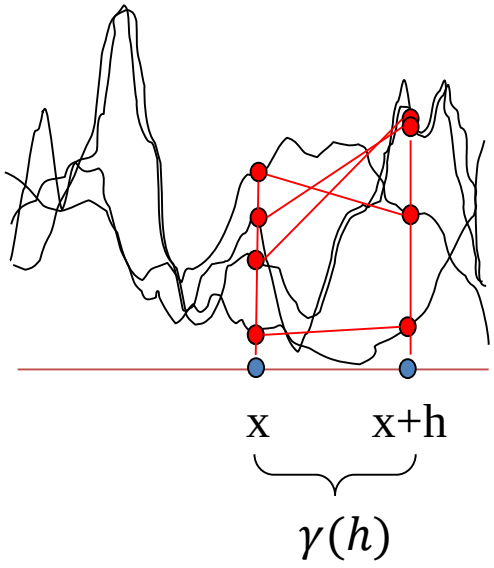
$$\sum_i \lambda_i = 0 \quad \Rightarrow \quad var\left(\sum_i \lambda_i Z(x_i)\right) = -\sum_i \sum_j \lambda_i \lambda_j \gamma_{i,j}$$

On prendre le cas le plus général, i.e. le cas des FAI.

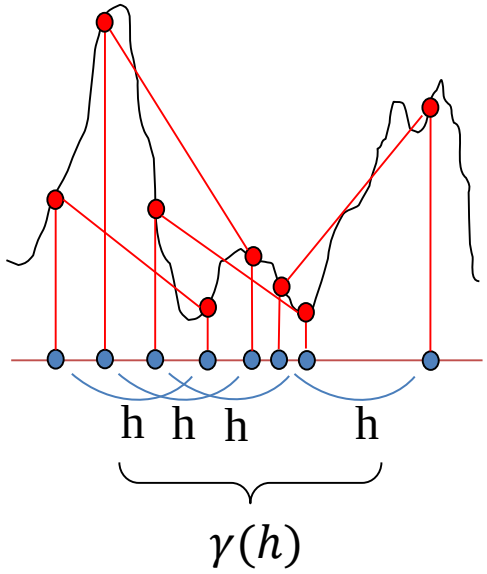
MODEL

TRUTH

Several realizations of an SRF



One regionalized variable



Statistics over
all the realisations of the RF
on (1 or 2) known points.

=
stationnarity

Statistics over
spatial repetitions
over one realisation of the RF

Comparison: stationary and intrinsic

- A stationary RF is also intrinsic, but an intrinsic RF is not necessarily stationary.
  Ex: Random walk with a linear variogram

- No recall towards an expected value...
  No systematic behavior either (like the sea bed that decreases from the shore)

Exploratory data analysis

$$s^2 = \frac{\sum\sum(z_i - z_i)^2}{2N^2}$$

$$\gamma^*(h) = \frac{\sum(z_i - z_i)^2}{2N(h)}$$

Assume: $\begin{cases} z_i \Rightarrow Z_{x_i} \\ E[Z_x - Z_{x+h}] = 0 \end{cases}$

$$\gamma(h) = \frac{1}{2} var(Z_x - Z_{x+h})$$

Prediction

LM: $E[Y|X_1, \dots, X_N] \approx \sum_{\alpha=1}^{N} \lambda_\alpha X_\alpha$

Kriging: $E[Z_0|Z_1, \dots, Z_N] \approx \sum_{\alpha=1}^{N} \lambda_\alpha Z_\alpha$

Random function (RF): $Z(x)$
defined by its spatial probability law (pdf)
$P\{Z_{x_1} < z_1, \dots, Z_{x_n} < z_n\}, \quad \forall n, \forall x_1, \dots, x_n$

$\begin{cases} E[Z_x] = E[Z_y] = m \\ cov(Z_x, Z_{x+h}) = C(h) \end{cases}$ Stationary random function - SRF (of order 2)

$\begin{cases} E[Z_x - Z_{x+h}] = 0 \\ var(Z_x - Z_{x+h}) = 2\gamma(h) \end{cases}$ Intrinsic Random Function - IRF

SRF $\subset$ **IRF**

If $C(h)$ exists, then $\boldsymbol{C(0) - C(h) = \gamma(h)}$

Working with variograms is more general
than working with covariances

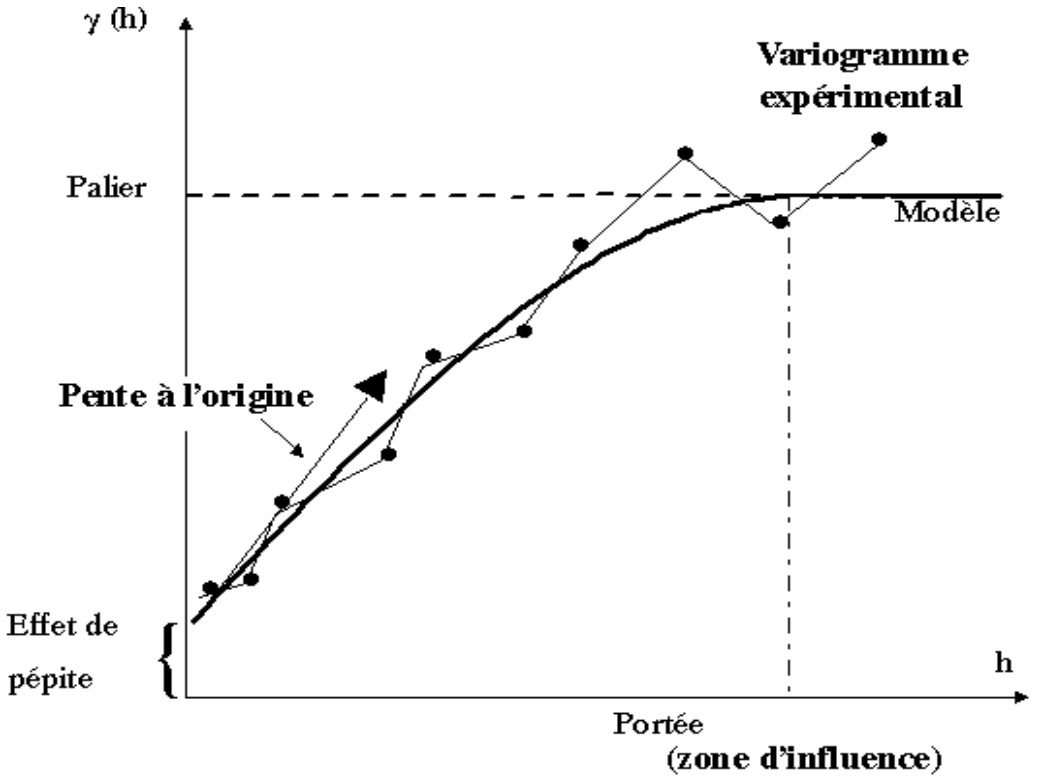
De la variance au variogramme

Passage au modèle…

Phase 2 : choix d'un modèle

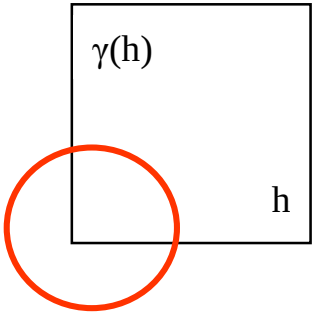
Propriétés structurales du variogramme

- comportement à l'origine
- portée, palier (comportement aux grandes distances)
- anisotropie
- structure gigogne



Comportement du modèle à l'origine

- Le comportement du variogramme est directement lié au degré de continuité de la variable



Variable différentiable

$\gamma(h) \approx h^2$
$|h| \rightarrow 0$

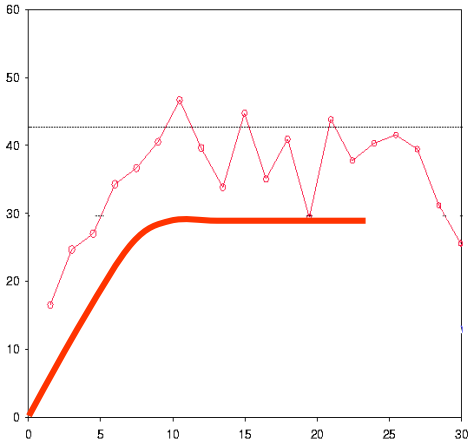
Variable continue

$\gamma(h) \approx h$
$|h| \rightarrow 0$

Variable discontinue

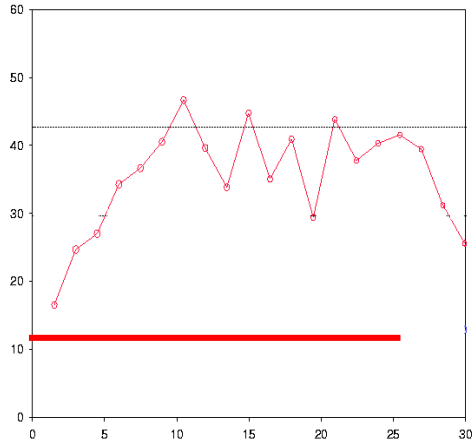
↑ Effet de pépite

Nested structures (1)



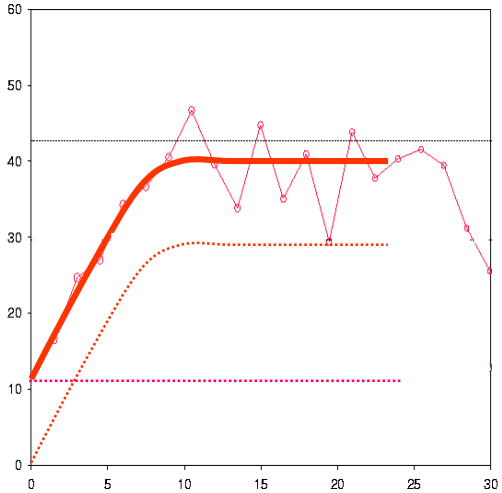
Spherical component
with sill = 43
and
range = 8

+



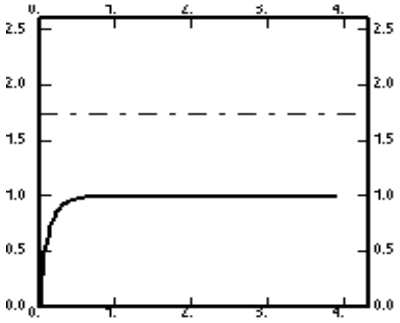
Nugget effect component
with sill = 10.5

=

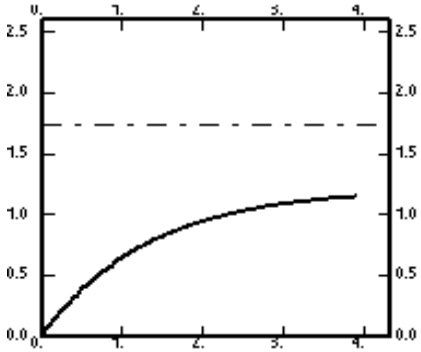


Three parameters
required here.

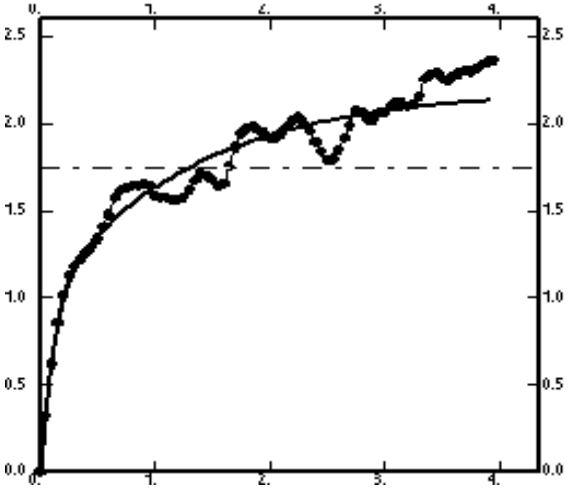
Nested structures (2)



Short range

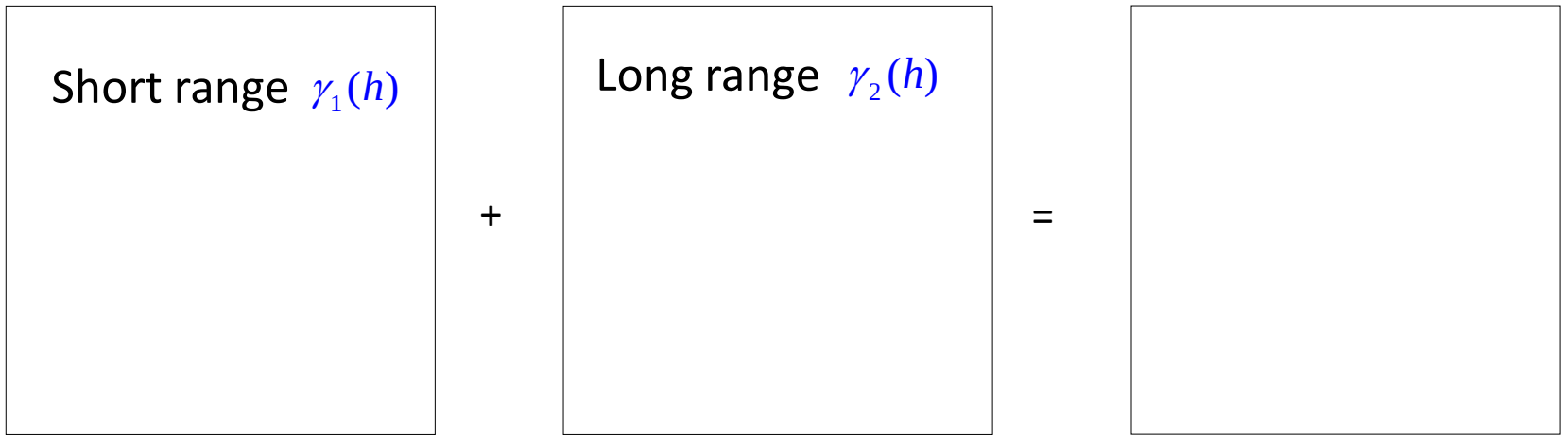


Long range



Nested structure

Nested Structures



$$\gamma(h) = \gamma_1(h) + \gamma_2(h)$$

Unsystematic measurement errors

Consider :

We want to study Y(x) with known variogram

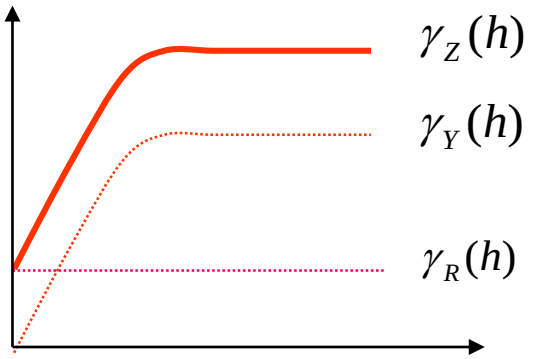
We record Z(x)=Y(x)+R(x) a measure of Y(x) with some unsystematic and random errors R(x). $\gamma_Y(h)$

If we can assume that this error is:

- on average 0
- with variance s^2
- without spatial correlation
- without correlation with Y(x)

Then $\gamma_R(h) = s^2.nugget(h)$

$\gamma_Z(h) = s^2.nugget(h) + \gamma_Y(h)$



Intrinsic random function: THE KEY equation

$$var\left(\sum_{i=0}^{N}\lambda_i Z(x_i)\right) = -\sum_{i=0}^{N}\sum_{j=0}^{N}\lambda_i\lambda_j\gamma\left(x_i - x_j\right)$$

provided that $\sum_{i=0}^{N}\lambda_i = 0$

$\Longrightarrow$ $-\gamma(h)$ conditionally positive definite

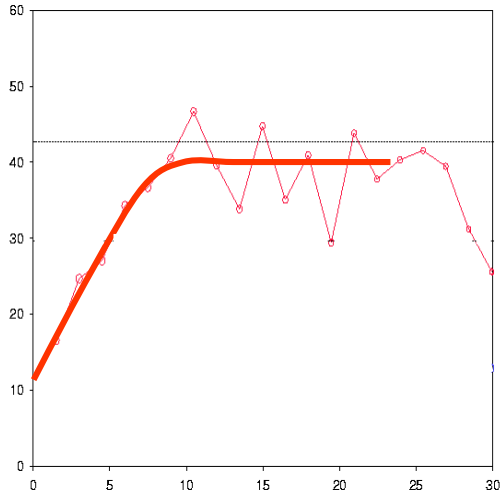
Modelisation

Allowed functions are defined by 1 or 2 parameters (C & a).

The definition of a variogram model :

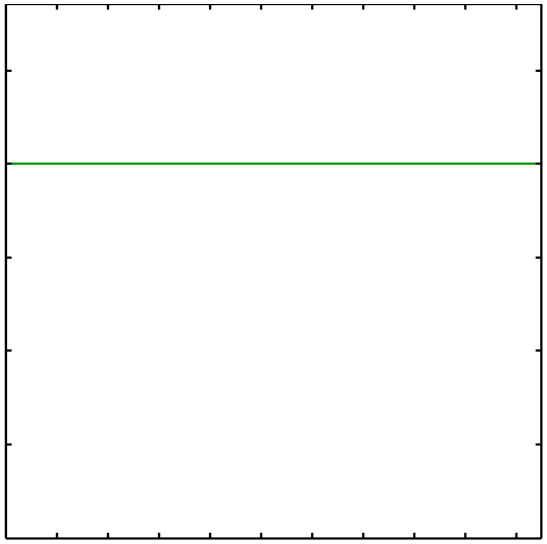
1. Choice of basic functions
2. Choice of parameters

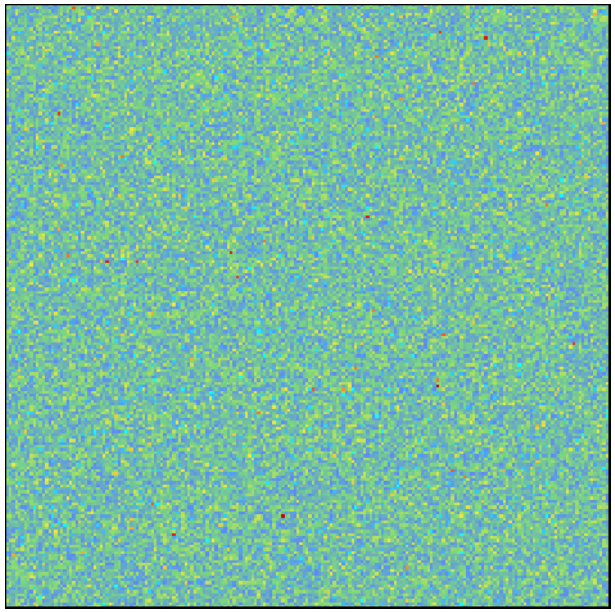
that allows to ‘best’ fit the experiemtal variogram



Isotropic spatial structure

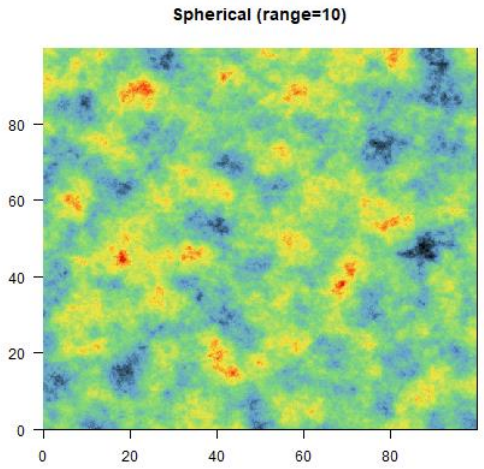
Nugget effect

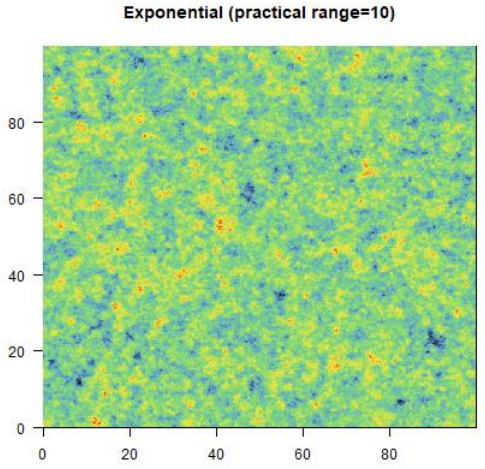


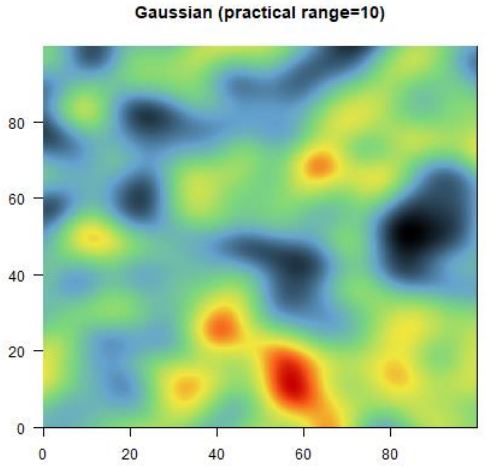


$$\gamma(h) = C \quad h \neq 0$$
$$\gamma(0) = 0$$

Model comparison







Spherical

$$\gamma(h) = \frac{C}{a}\left(\frac{3h}{a} - \frac{h^3}{a^3}\right) \; ; \; h < a$$

Exponential

$$\gamma(h) = C\left(1 - e^{-\frac{h}{a}}\right)$$

Practical range : $3a$

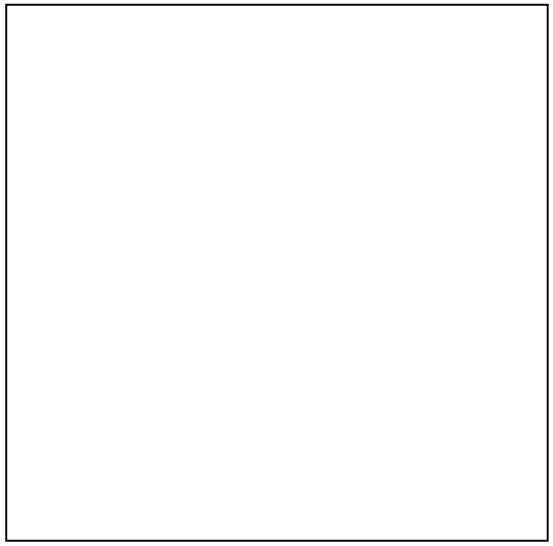
Gaussian

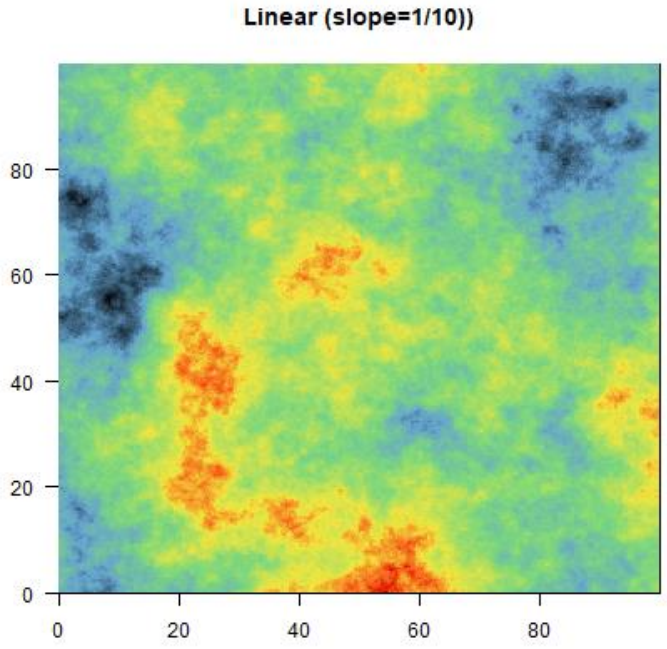
$$\gamma(h) = C\left(1 - e^{-\frac{h^2}{a^2}}\right)$$

Practical range : $\sqrt{3}a$

Isotropic spatial structure

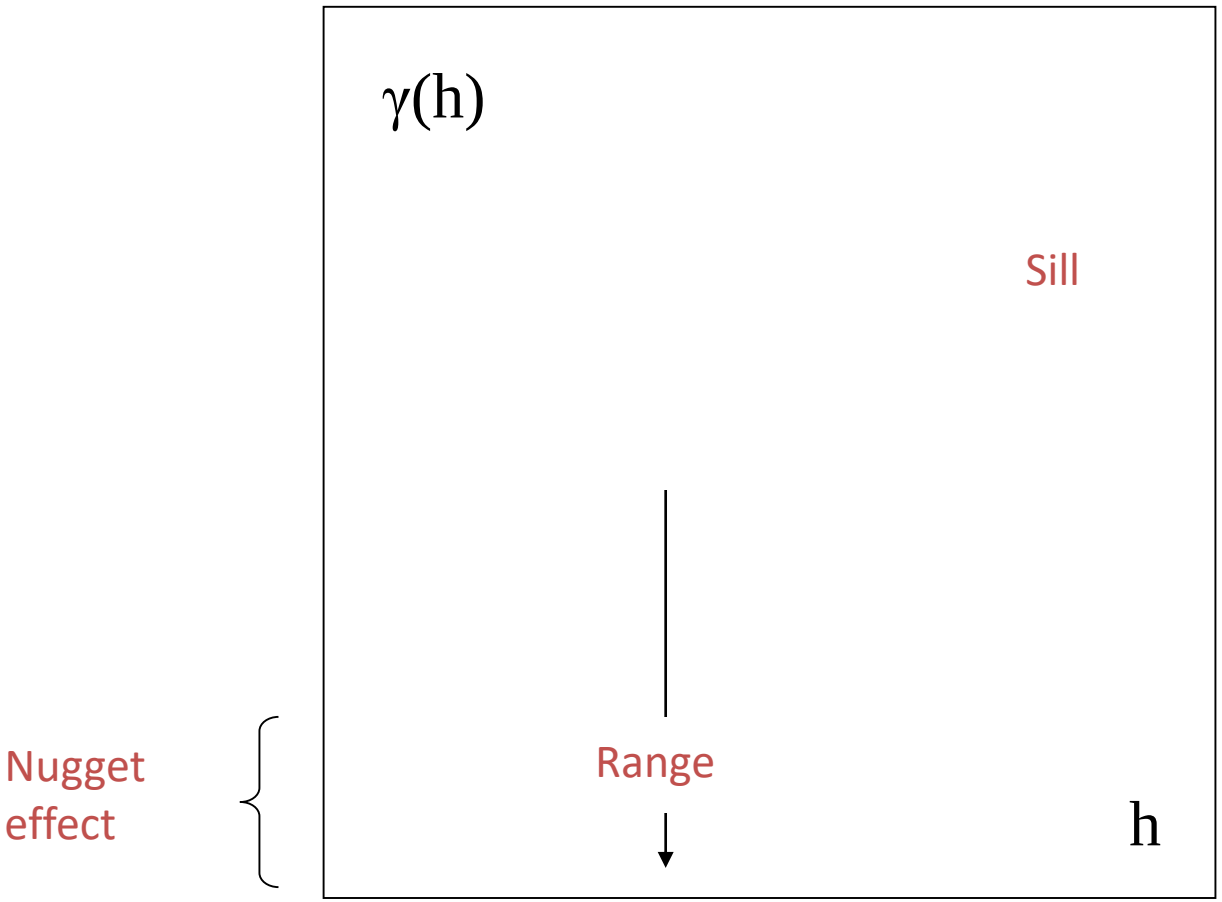
Linear





$$\gamma(h) = c\frac{|h|}{a}$$

Model characteristics



- Elements of a variogram model:

•Number of basic functions (typically 2 or 3) and their type

•Their parameter values :

- Sills for all functions
- Ranges for all but the nugget effect
- A third parameter e.g. for hole effect
- Anisotropy (direction, coefficient)

- Manual, semi-automatic, automatic fittings

Nota : ranges are non linear parameters ➔ iterative procedures

- Least squares criteria

- No statistical fitting test

non independence of variogram values, no parametric assumption of their statistical distribution with the exception of Gaussian geostatistics

- Goodness Of Fit

$$GOF = \frac{\sum_i (\gamma_{\mathrm{mod}}(h_i) - \gamma_{\exp}(h_i))^2}{\sum_i \gamma^2_{\exp}(h_i)}$$

Conclusion

Il existe des processus stochastiques pour lesquels l'inférence de certains paramètres du processus sont accessibles à partir d'une seule de ses réalisations.

Le prix à payer est double

- Hypothèse (de stationnarité)
- On ne cherchera à n'inférer que les moments d'ordre 1 et 2

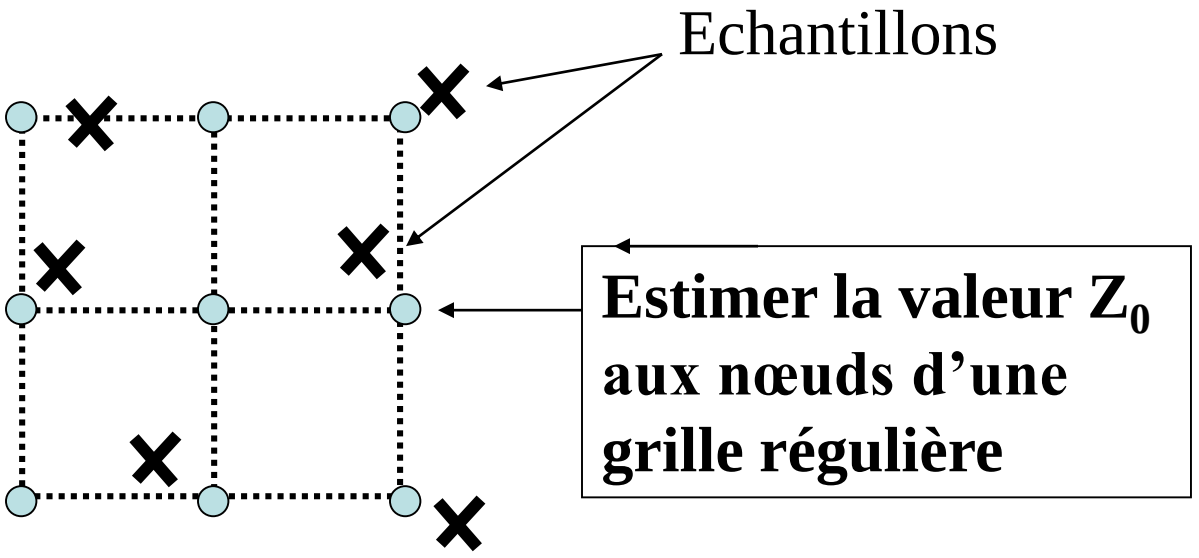
Il se trouve qu'ils sont suffisants pour faire de l'estimation

Interpolation spatiale et krigeage

Utilisation d’un modèle

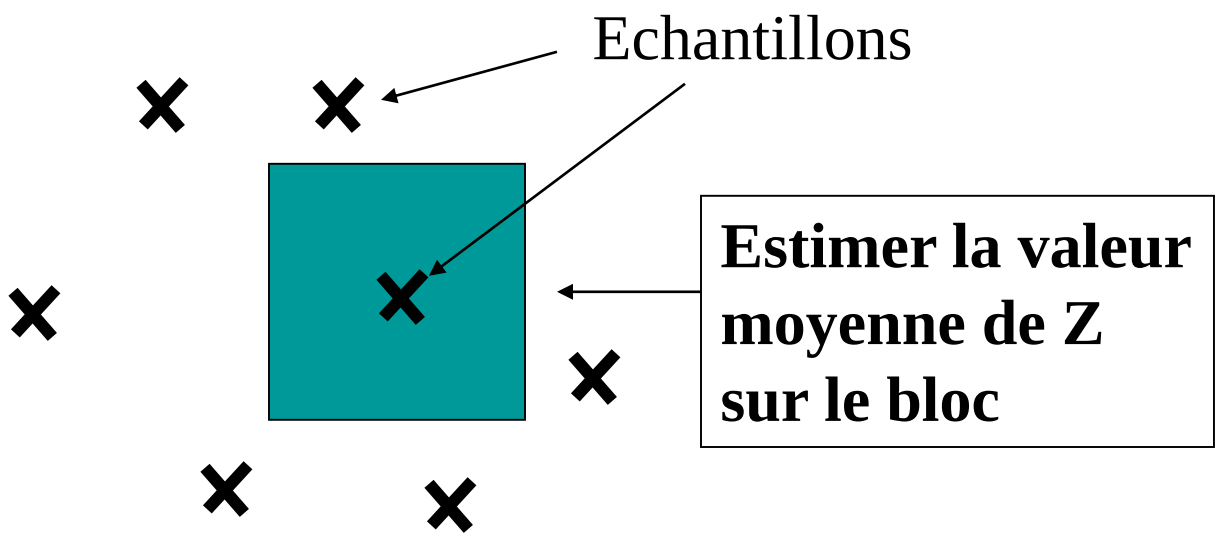
Que veut-on estimer ?

Estimation ponctuelle



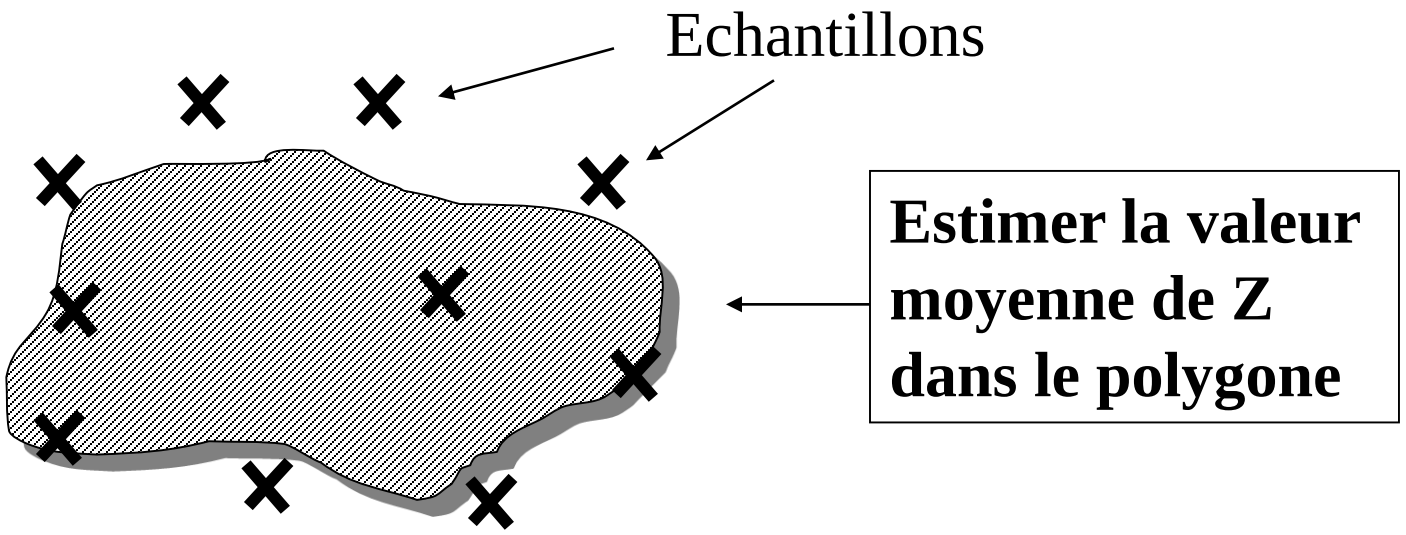
Que veut-on estimer ?

Estimation de bloc



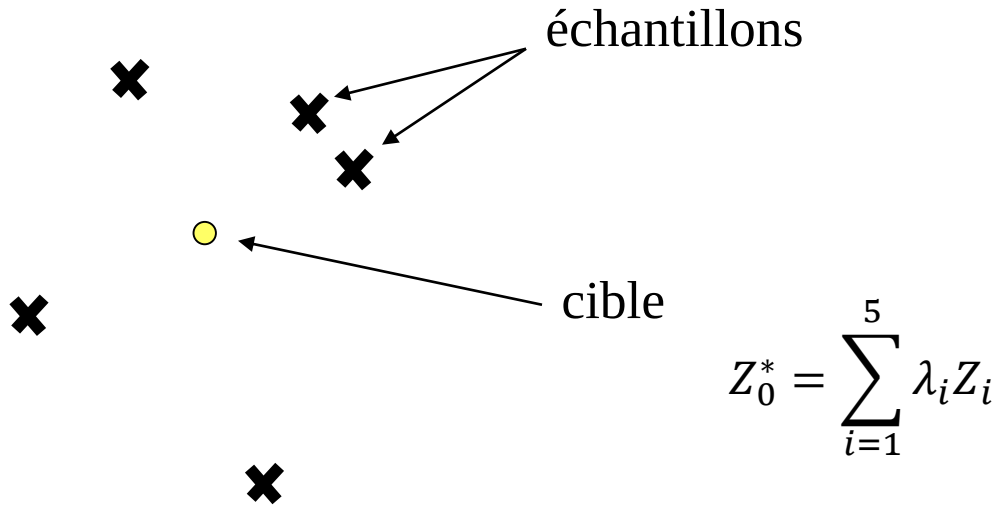
Que veut-on estimer ?

Estimation de blocs irréguliers



Estimateurs linéaires classiques

- Moyenne mobile
- Polygone d'influence
- Inverse des distances
- Ajustement polynomial par moindres carrés

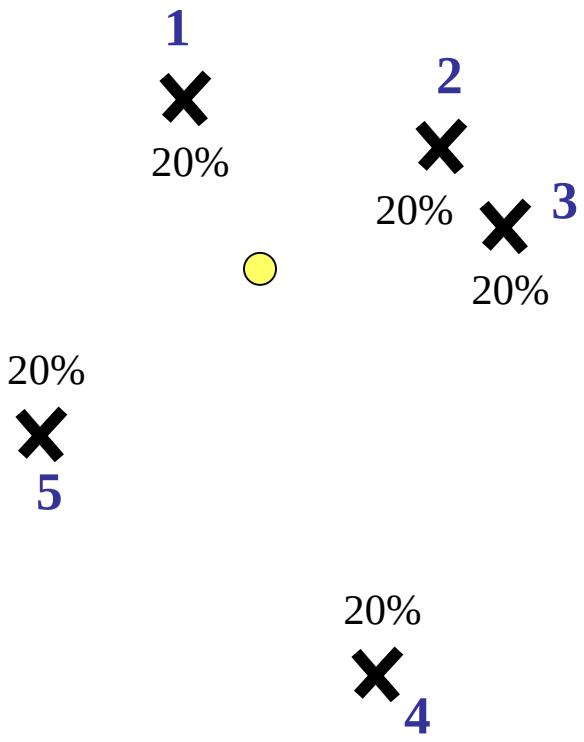


Illustration

Illustration

Moyenne mobile

Même pondération
pour chaque donnée



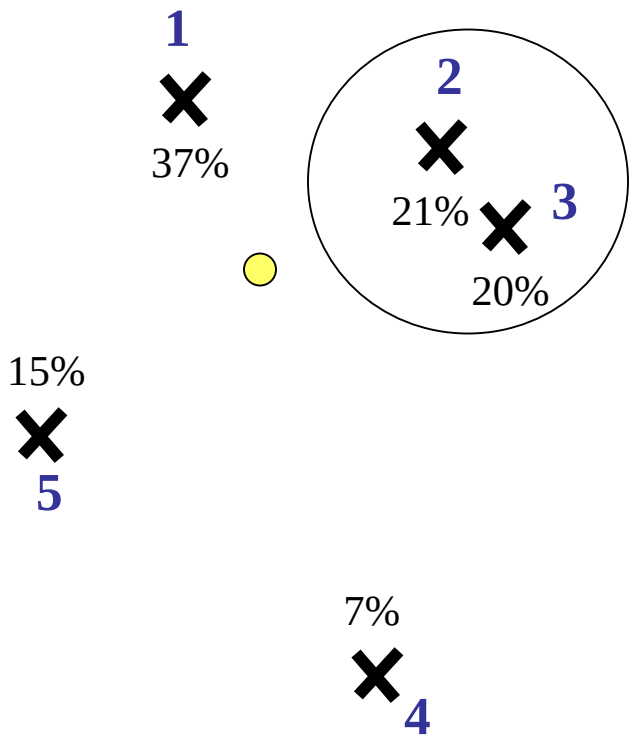
$$Z^* = \frac{\sum Z_i}{5}$$

Pas de discrimination:

- **en fonction de la distance à la cible**
- **des points redondants**

Moyenne mobile

Inverse des distances



La pondération dépend de la distance à la cible

$$Z^* = \frac{\sum \frac{Z_i}{\varphi(d_i)}}{\sum \frac{1}{\varphi(d_i)}}$$

$$\varphi(d) = d^2$$

- **Ne discrimine pas l'information redondante**
- **Choix du degré**

Inverse des distances

Inverse d

Inverse d^2

Moindres carrés

- La surface ne passe pas par les points de données
- Estimation par un polynôme d' ordre k

$$m(x) = \sum_{l=1}^{k} a_l f^l(x)$$

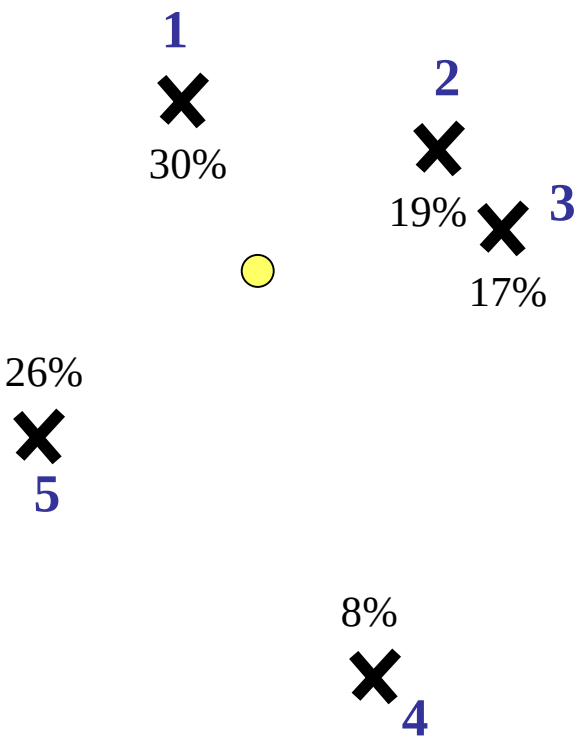
- Minimisation de la distance entre les données et l' estimation

$$\sum_{i=1}^{n} \left(\sum_{l=1}^{k} a_l f^l(x_i) - z(x_i) \right)^2$$

- Surface de « tendance »

Moindres carrés

Surface de “ tendance”
(poids pour k=1)



- **Ne restitue pas les valeurs aux points de données**
- **Choix du degré du polynôme**

Moindres carrés

Ordre 2

Ordre 6

Krigeage - Construction

Pour estimer au mieux la variable Z en un point inconnu, on construit un estimateur linéaire:

$$Z_0^* = \sum_{i=1}^{N} \lambda_i Z_i$$

conditionné par les N valeurs observées dans un certain voisinage, les pondérateurs λ_i étant les inconnues du problème (pondérateurs de krigeage).

Cet estimateur est optimal dans le cadre (multi)Gaussien. Dans les autres cas, on obtient le BLUE qui est sub-optimal.

BLUE: Best Linear Unbiased Estimator

Krigeage - Construction

- Estimation ponctuelle:

$$Z_0 = Z(x_0) \;\; ; \;\; \varepsilon_0 = \sum_i \lambda_i Z_i \; - \; Z_0$$

- Estimation non ponctuelle :

$$Z_0 = Z(V) = \frac{1}{V} \int_V Z(x)dx \; ; \;\; \varepsilon_0 = \sum_i \lambda_i Z_i \; - \; \frac{1}{V} \int_V Z(x)dx$$

Krigeage ordinaire ponctuel

Krigeage ordinaire ou krigeage à moyenne inconnue

- Non biais = Espérance nulle:

$$E(\varepsilon_0) = \sum_i \lambda_i E(Z_i) - E(Z_0) = m\sum_{i=1}^{N} \lambda_i - m = m(1 - \sum \lambda_i)$$

$$E(\varepsilon_0) = 0 \Rightarrow \left(1 - \sum \lambda_i\right) = 0 \qquad \sum_{i=1}^{N} \lambda_i = 1$$

- Variance d'estimation :

$$\sigma_E^2 = var(\varepsilon_0) = var\left(\sum_{i=1} \lambda_i Z_i - Z_0\right) = var\left(\sum_{i=0} \lambda_i Z_i\right)$$

avec $\lambda_0 = -1$

Krigeage ordinaire

Cas des F.A. avec covariances spatiales (F.A. Stationnaire)

$$Cov\big(Z(x_i), Z(x_j)\big) = C\big(x_i - x_j\big) = C_{i,j}$$

$$\sigma_E^2 = \sum_{i=0}^{N}\sum_{j=0}^{N} \lambda_i \lambda_j C_{i,j}$$

Cas des F.A. avec variogrammes (F.A. Intrinsèque)

$$\gamma\big(Z(x_i), Z(x_j)\big) = \gamma\big(x_i - x_j\big) = \gamma_{i,j}$$

$$\sum_{i=0}^{N} \lambda_i = 0 \quad \Rightarrow \quad \sigma_E^2 = -\sum_{i=0}^{N}\sum_{j=0}^{N} \lambda_i \lambda_j \gamma_{i,j}$$

$$\lambda_0 = -1$$

On prendre le cas le plus général, i.e. le cas des FAI.

Krigeage ordinaire

Intermediate step: partial derivatives with regards to the kriging weights λ_i:

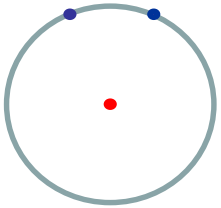
$$\frac{\partial \sigma_E^2}{\partial \lambda_i} = -2\sum_{j=0}^{N} \lambda_j \gamma_{i,j} = -2\sum_{j=1}^{N} \lambda_j \gamma_{i,j} + 2\gamma_{i,0}$$

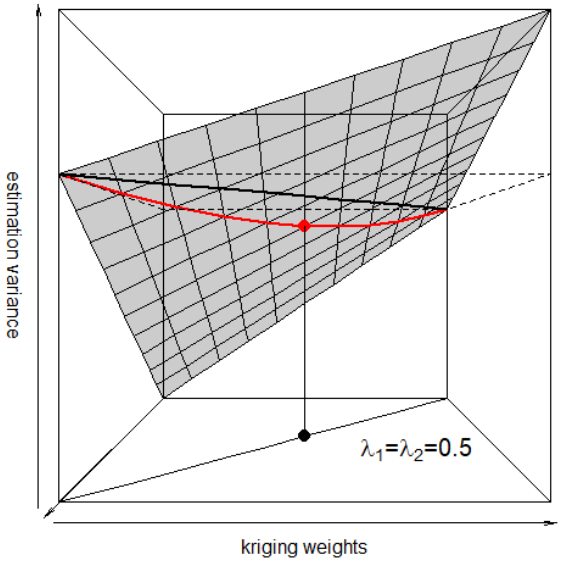
But we must consider also the constraint on the kriging weights (see further for the equations).

Illustration with two points with a linear variogram of unit slope.
The estimation variance is:

$$\sigma_E^2 = -2\gamma_{1,0} \cdot \lambda_1 - 2\gamma_{2,0} \cdot \lambda_2 + 2\gamma_{1,2} \cdot \lambda_1 \lambda_2$$

The red line corresponds to $\lambda_1 + \lambda_2 = 1$





Krigeage ordinaire

Revient à minimiser l'expression: $\phi = \sigma_E^2 - 2\mu\left(\sum_{i=1}^{N}\lambda_i - 1\right)$

en introduisant μ un paramètre de Lagrange, c.a.d. en annulant les dérivées partielles de ϕ par rapport à:

- chaque inconnue: $\frac{\partial\phi}{\partial\lambda_i} = 0 \Longrightarrow \sum_{j=1}^{N}\lambda_j\gamma_{i,j} + \mu = \gamma_{i,0}$
- au paramètre de Lagrange : $\frac{\partial\phi}{\partial\mu} = 0 \Longrightarrow \sum_{i=1}^{N}\lambda_i = 1$

Krigeage ordinaire

Le système de krigeage ordinaire:

n+1 inconnues ; n+1 équations

$$\sum_{j=1}^{n} \lambda_j \gamma_{1,j} + \mu = \gamma_{1,0}$$

$$\sum_{j=1}^{n} \lambda_j \gamma_{2,j} + \mu = \gamma_{2,0}$$

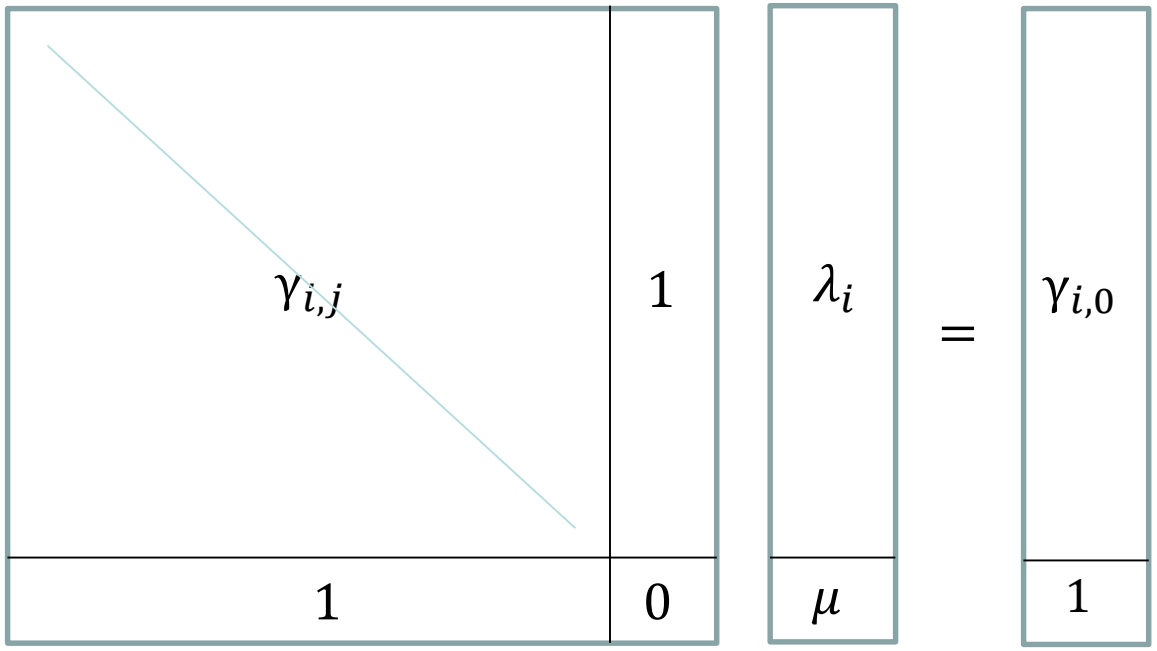
.....

$$\sum_{j=1}^{n} \lambda_j \gamma_{n,j} + \mu = \gamma_{n,0}$$

$$\sum_{i=1}^{n} \lambda_i = 1$$

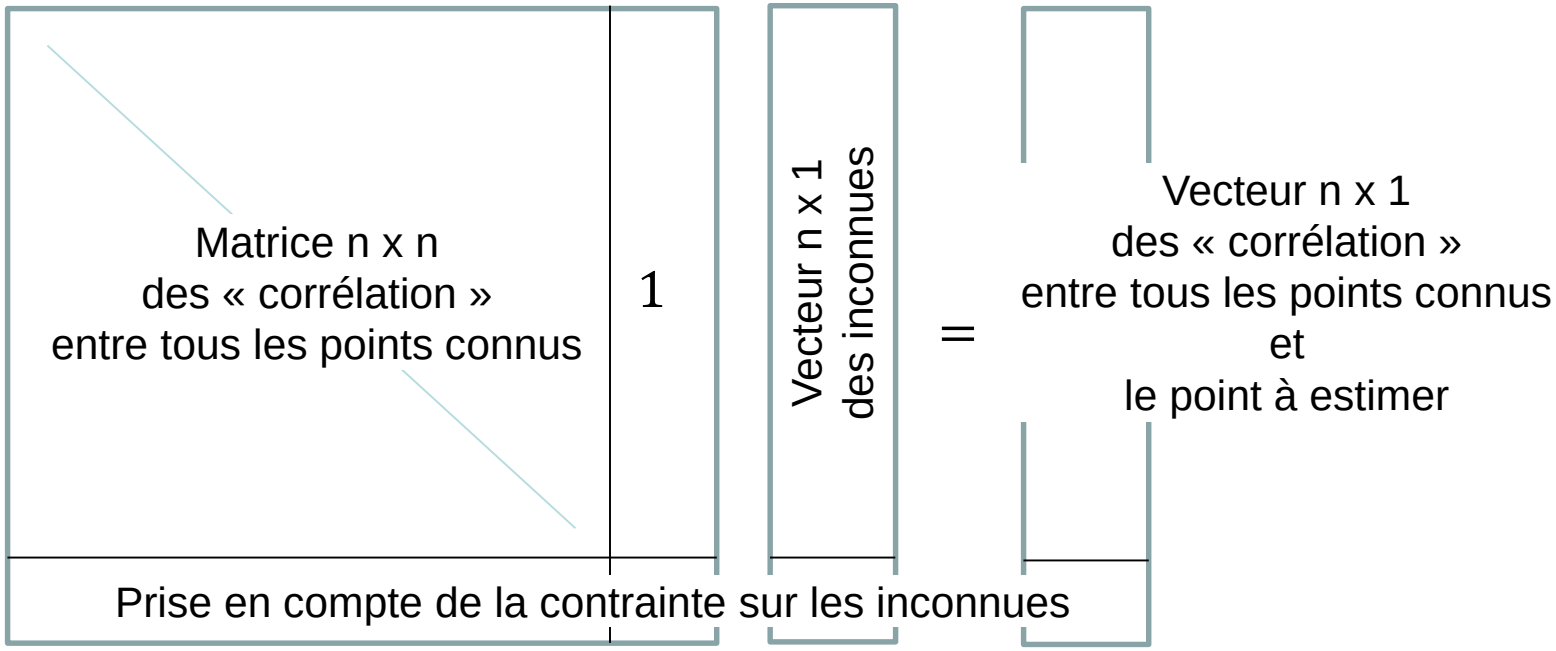
Krigeage ordinaire

Le système de krigeage ordinaire:



Krigeage ordinaire

Le système de krigeage ordinaire:



Le modèle de variogramme suffit à remplir le système

Krigeage ordinaire

$$\begin{bmatrix} \gamma_{i,j} & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_i \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{i,0} \\ 1 \end{bmatrix}$$

(n+1) x (n+1) *(n+1) x 1* *(n+1) x 1*

$$\begin{bmatrix} \lambda_i^* \\ \mu^* \end{bmatrix} = \begin{bmatrix} \gamma_{i,j} & 1 \\ 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \gamma_{i,0} \\ 1 \end{bmatrix}$$

$$[\,L\,] \cdot [\Lambda] = [R]$$

Left *Right*

$$[\Lambda^*] = [\,L\,]^{-1} \cdot [R]$$

Krigeage ordinaire

Inversion de la matrice (n+1) x (n+1) ➔ $\lambda_i^*, i = 1, \dots, n$ et μ^*

D'où l'estimateur par krigeage et la variance de l'erreur d'estimation:

$$Z_0^* = \sum_{i=1}^{n} \lambda_i^* Z_i$$

$$\sigma_E^2 = \sum_{i=1}^{n} \lambda_i^* \gamma_{i,0} + \mu^* = [\Lambda^*] \cdot [R]$$

Krigeage

Variogramme linéaire - voisinage unique

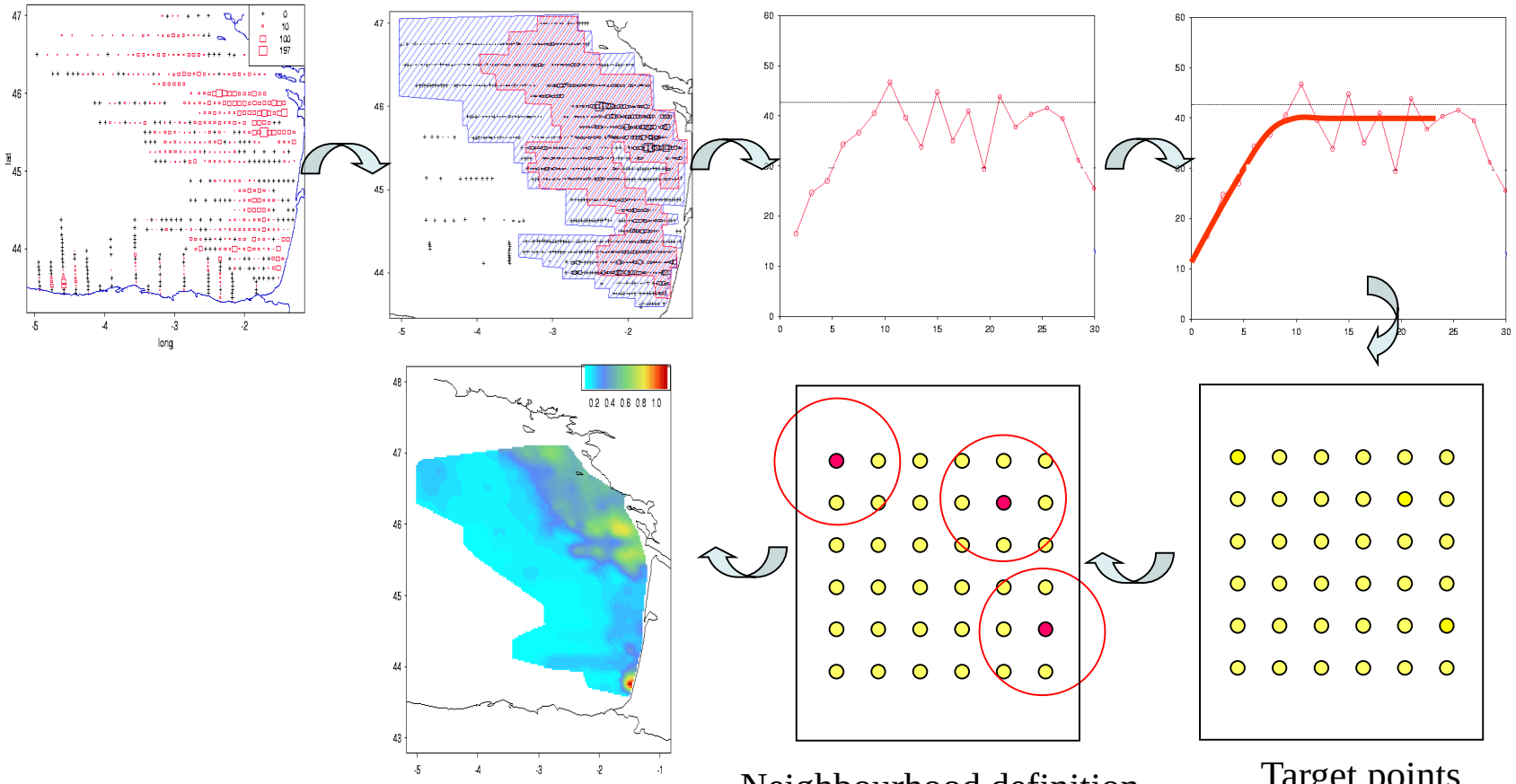
Main steps of a kriging

Data representation

Field delineation

Experimental variogram

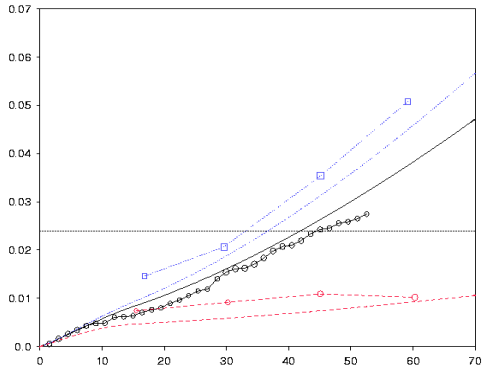
Model definition



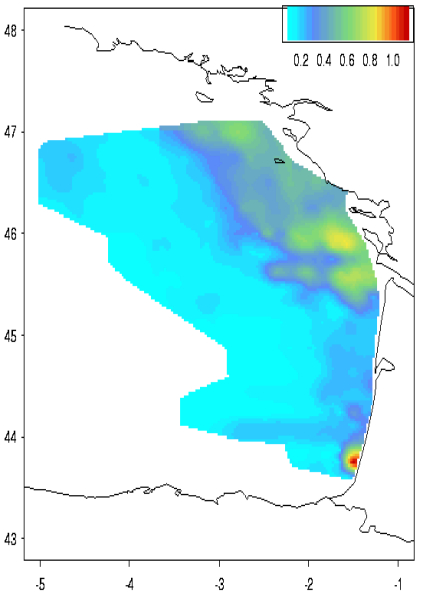
Neighbourhood definition

Target points

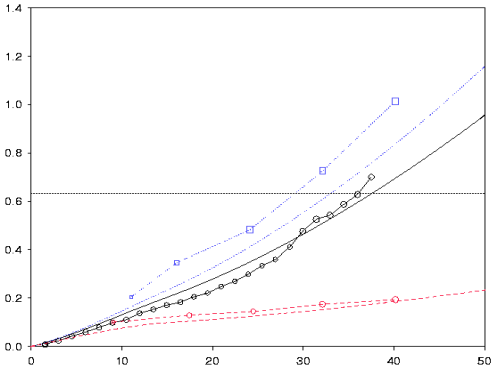
Chlorophyll (µmole.l^{-1})



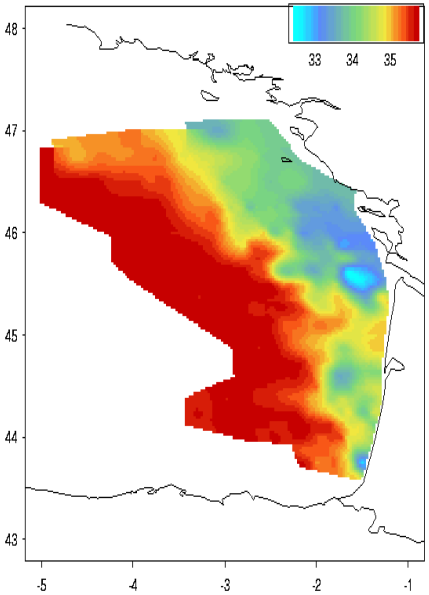
Distances (n. mi.)



Salinity (‰)



Distances (n. mi.)



Krigeage - Propriétés

Le système de krigeage fait intervenir:

au travers de la structure de la variable C ou γ

- les distances entre points de données $C_{\alpha\beta}$ ou $\gamma_{\alpha\beta}$
- les distances entre les données et la cible $C_{\alpha 0}$ ou $\gamma_{\alpha 0}$
- la géométrie de la cible C_{vv} ou γ_{vv}

Krigeage - Propriétés

Ni le système de krigeage, ni la variance de l'erreur d'estimation ne font intervenir les valeurs des données.

Les pondérateurs restent inchangés lorsqu'on multiplie le palier du variogramme par une constante. Le paramètre de Lagrange est multiplié par cette constante.

La variance de l'erreur d'estimation est directement proportionnelle au palier du variogramme.

Krigeage - Propriétés

- Le krigeage consiste à minimiser une forme quadratique sous une éventuelle contrainte linéaire
- Le système de krigeage est régulier si:
  - la structure est autorisée
  - il n'y a pas d'information redondante (ex. points doubles)

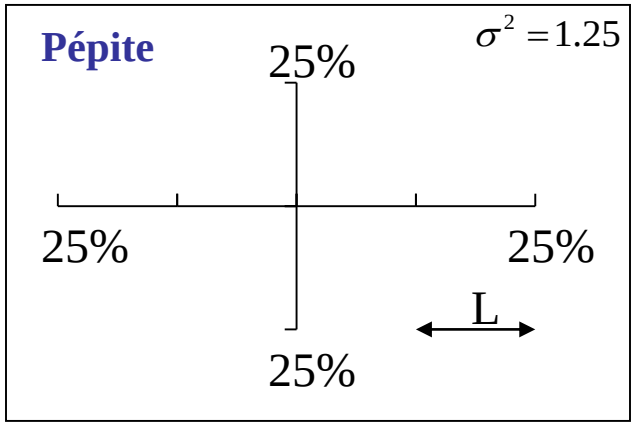
  La solution est unique
- Le krigeage est un interpolateur exact

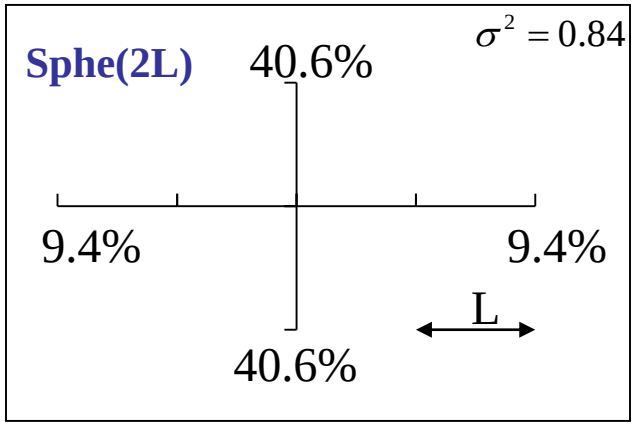
$$Z^*(x_\alpha) = Z_\alpha \quad \text{et} \quad Var\left(\varepsilon_\alpha\right) = 0$$

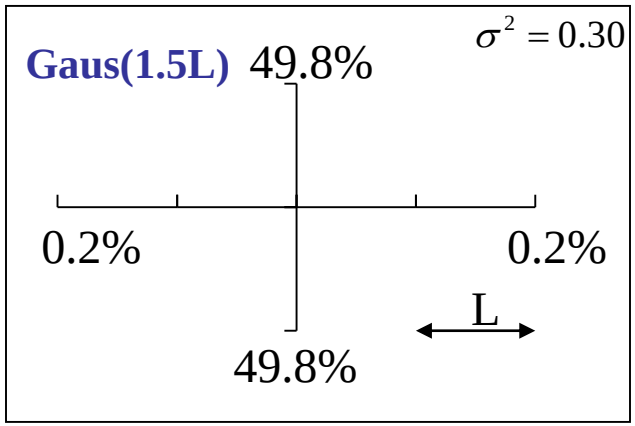
- Le krigeage permet d'estimer toute fonction linéaire de la variable

Influence du modèle sur les poids

Influence du choix du modèle dans le krigeage ordinaire:

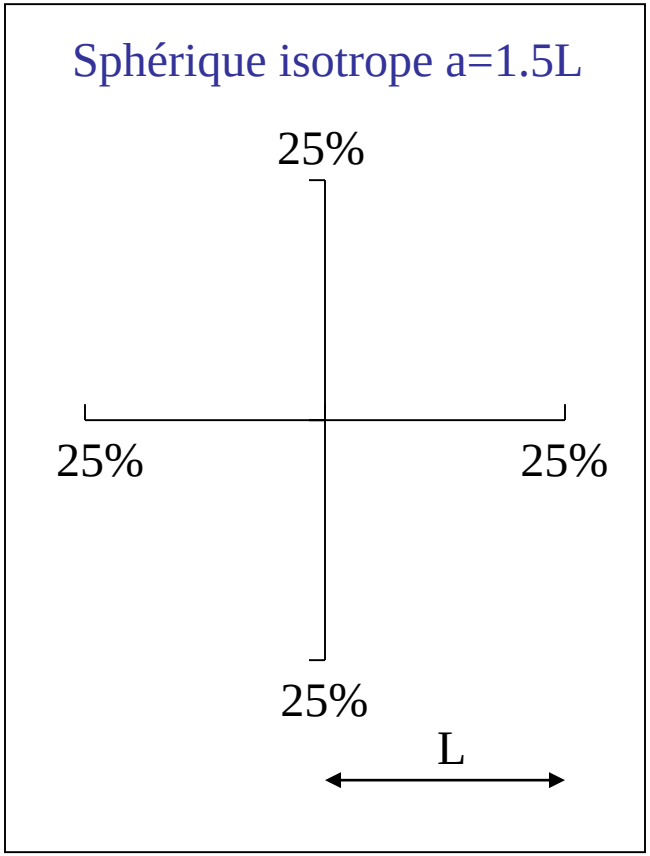






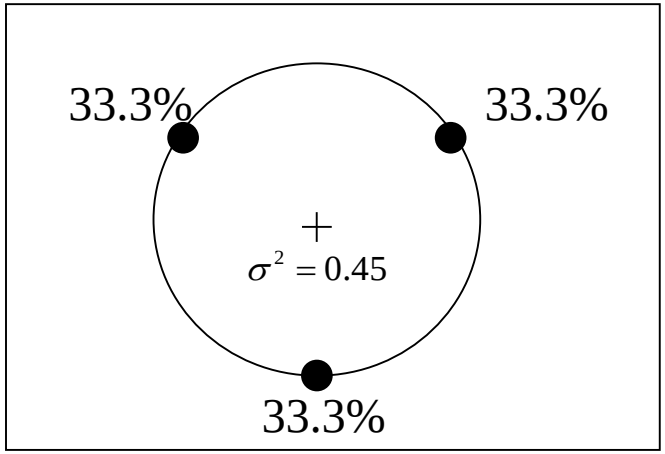
Influence de l'anisotropie sur les poids

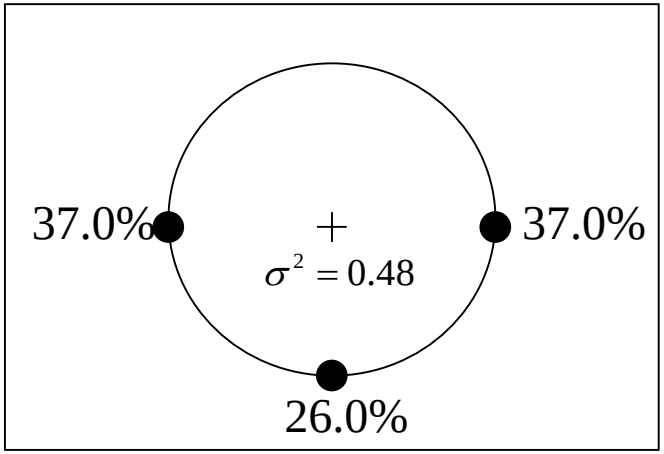
Influence de l'anisotropie dans le krigeage ordinaire:

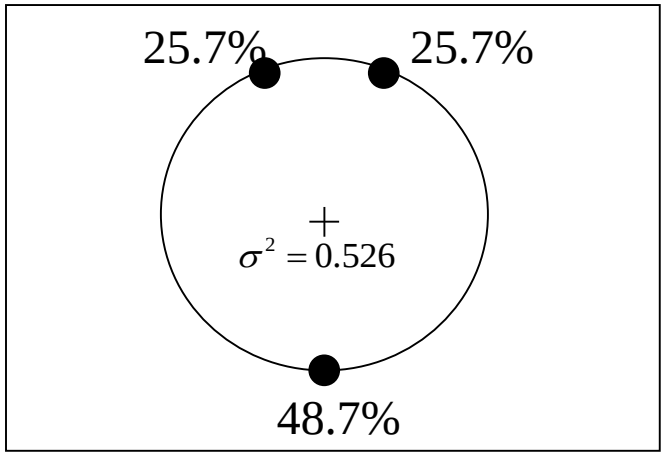


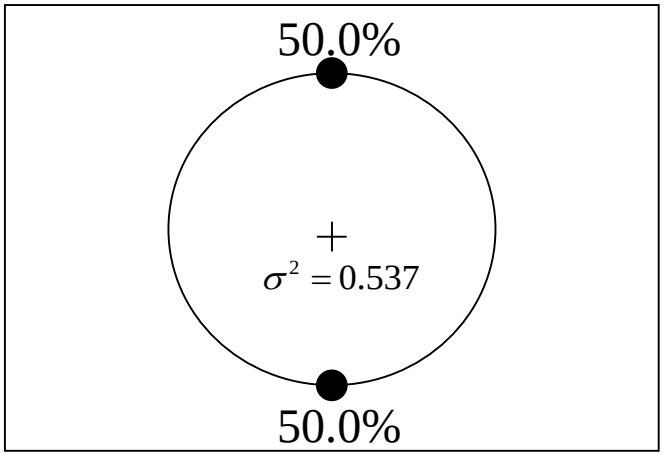
Disposition des points

Sphérique, portée = 3 * rayon



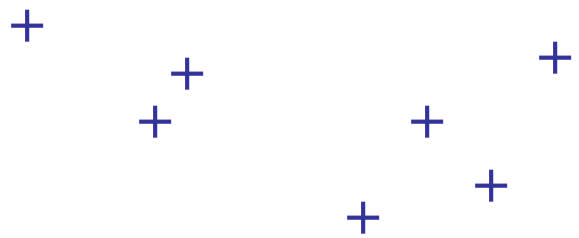






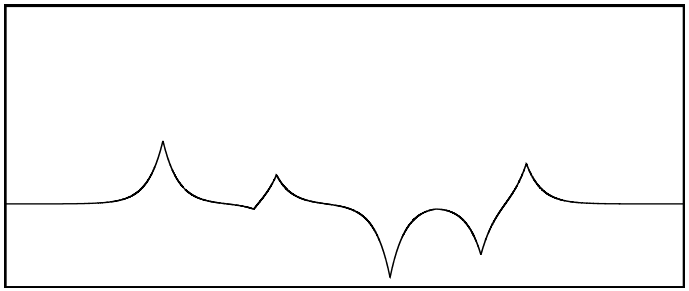
Influence de la portée

Krigeage ordinaire à 1-D avec 7 points de données
Variogramme sphérique de palier=100
Portées 5, 10, 15, 20, 25, 30

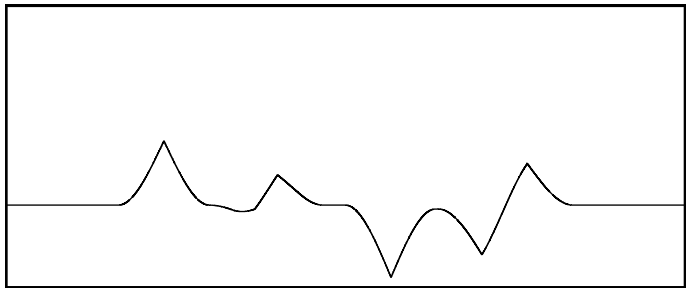


Influence du type de modèle

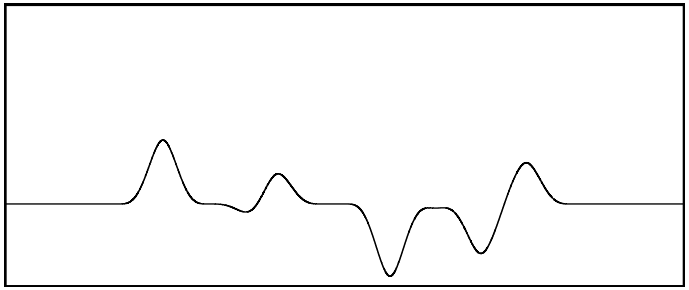
Krigeage ordinaire: palier 100 et portée 10



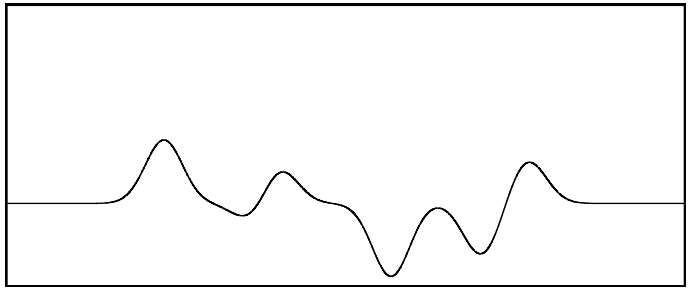
Exponentiel



Sphérique



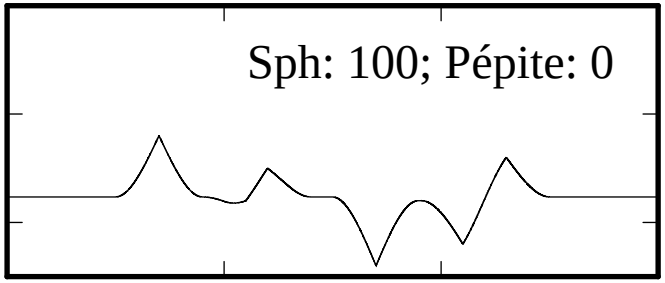
Cubique

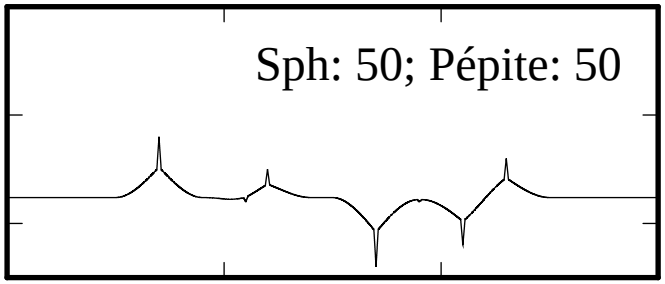


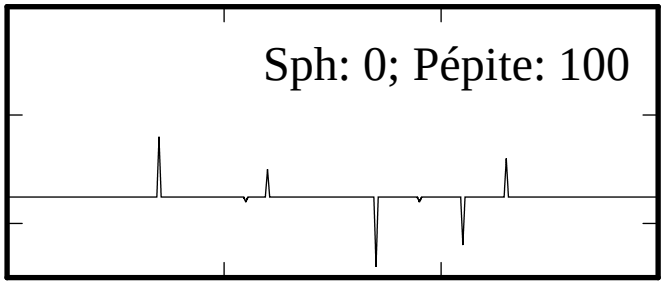
Gaussien

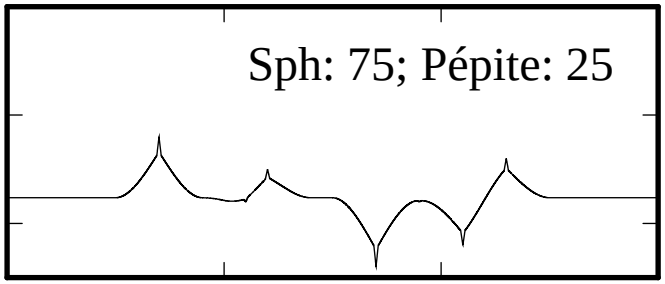
Influence de l'effet de pépite

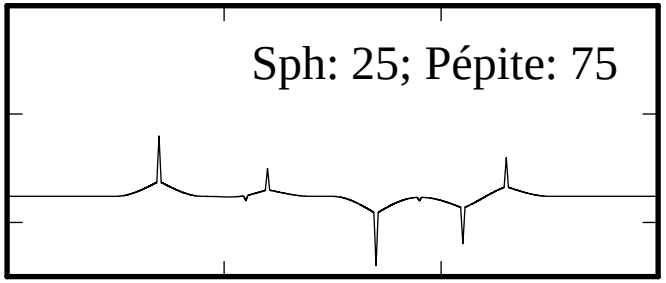
Modèle sphérique (portée=10)
+ effet de pépite







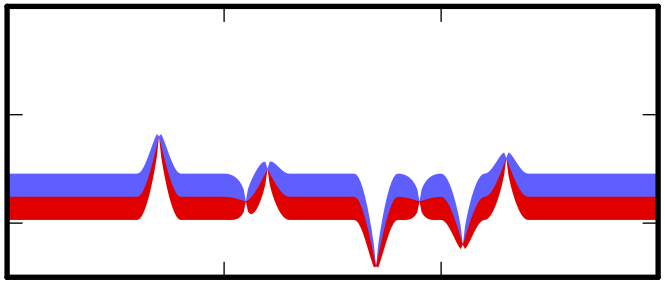




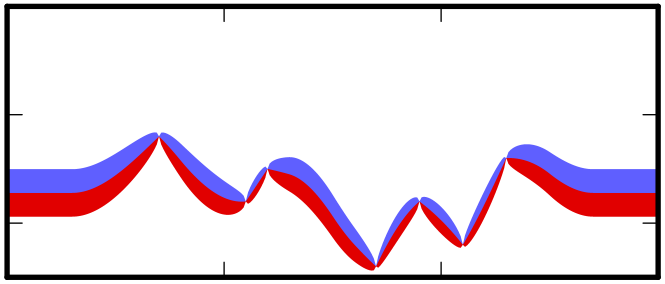
Influence de la portée

Sphérique de palier 100

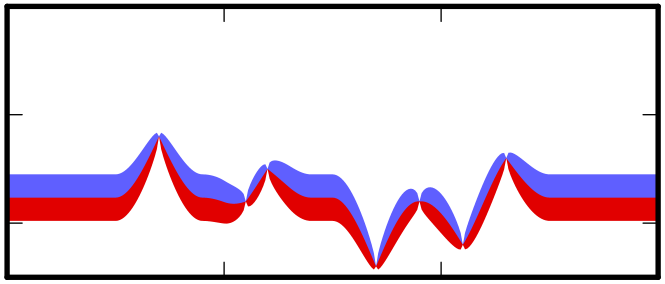
(rouge et bleu = ⟨1 écart-type de krigeage)



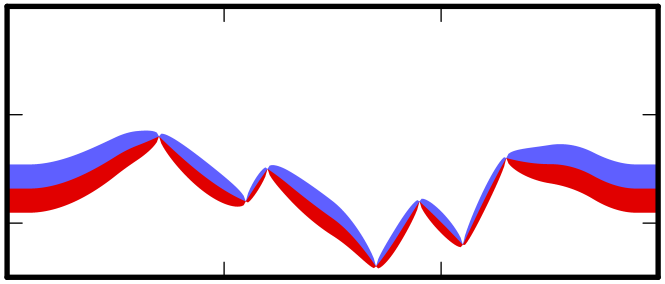
Portée 5



Portée 20



Portée 10



Portée 30

The intrinsic approach of geostatistics

- Due to its random character, the variable is represented by a Random Function (RF)model
- A stationarity hypothesis (ie invariance under translation), eg on the variable or on its increments, is made for the inference of the model
- Order-2 stationarity: we only pay attention to the first two moments of the RF, i.e. (1) mean and (2) variance or covariance

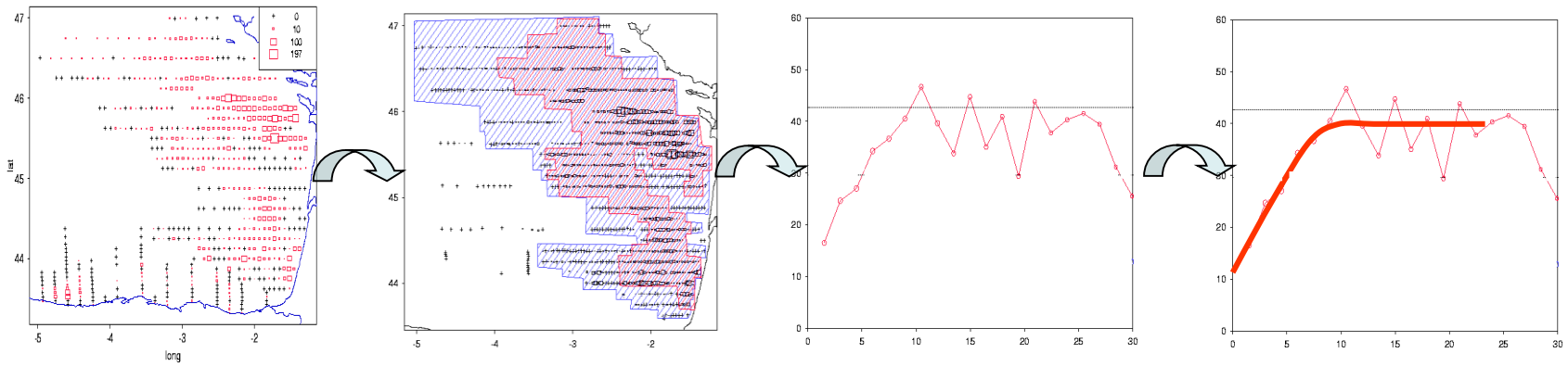
Main steps of a geostatistical analysis

Data representation

Field delineation

Experimental variogram

Model definition

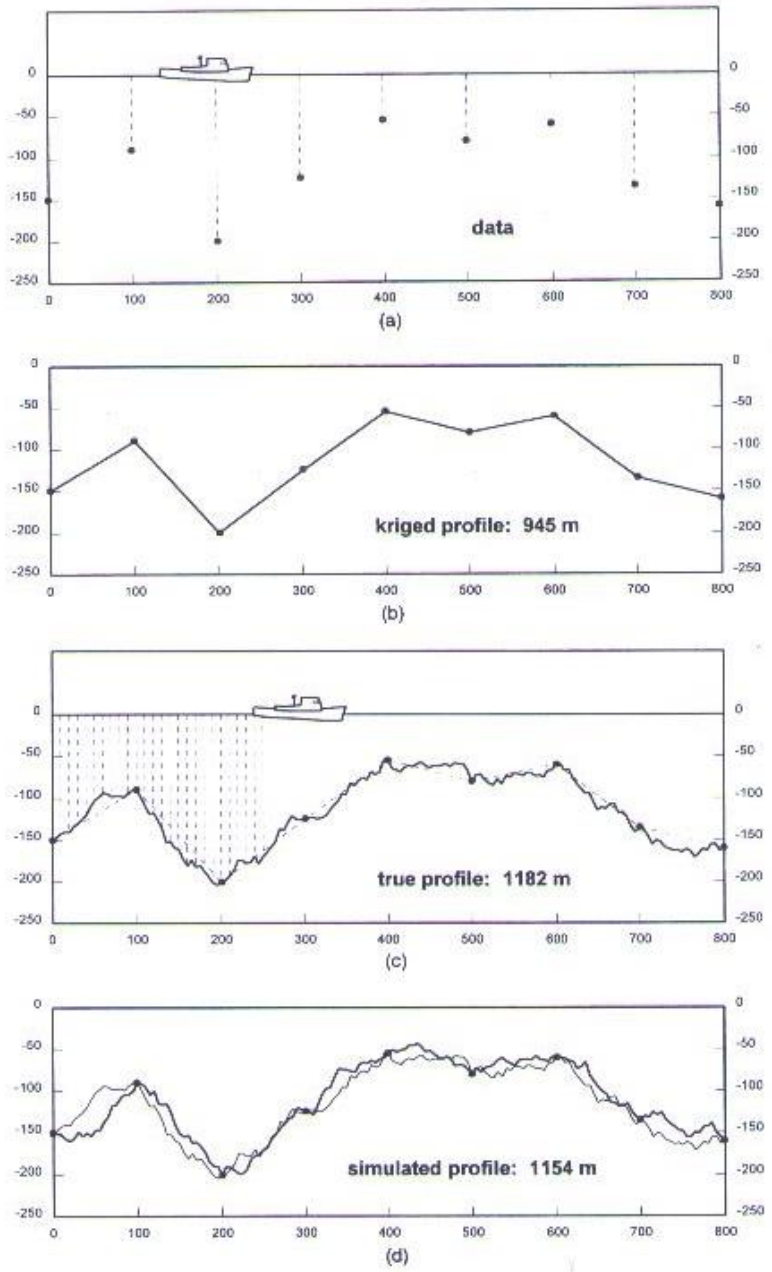


Use of the model

- Interpolations (local = **kriging** & global)
- Estimation Variance
- Scale, support
- Simulations
- etc

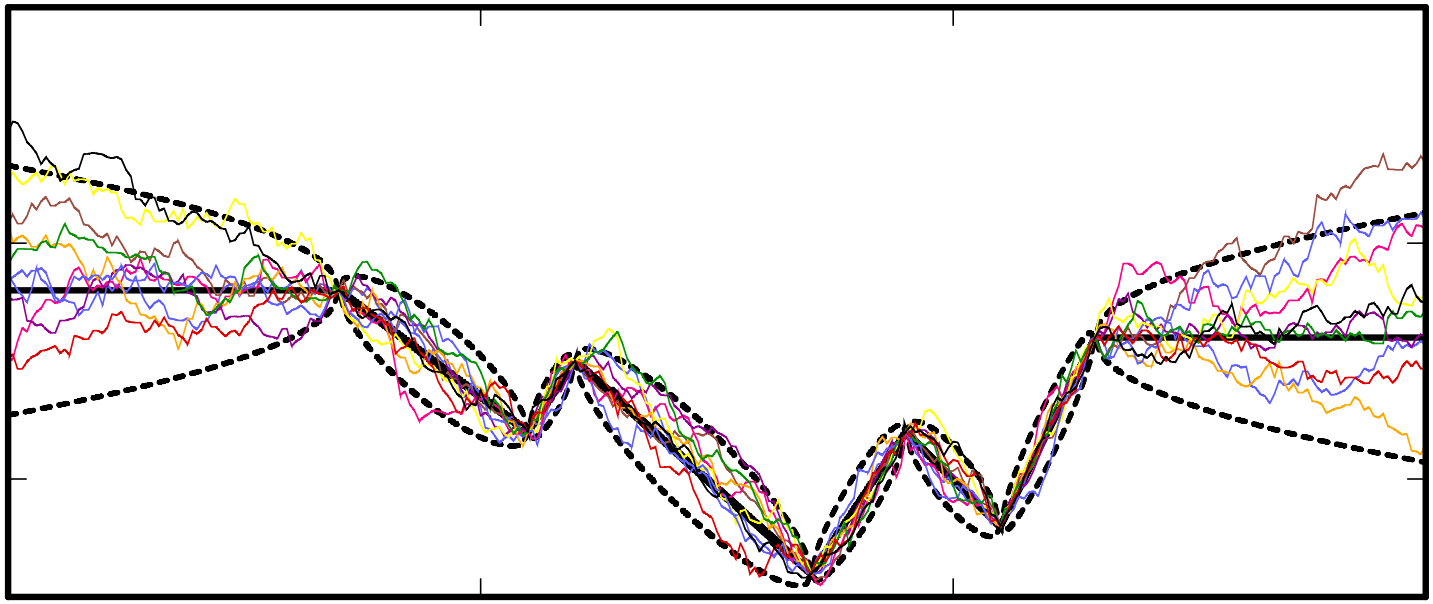
Estimation de la longueur d'un cable sous marin

D'après Chilès et Delfiner, 1999.



Simulations

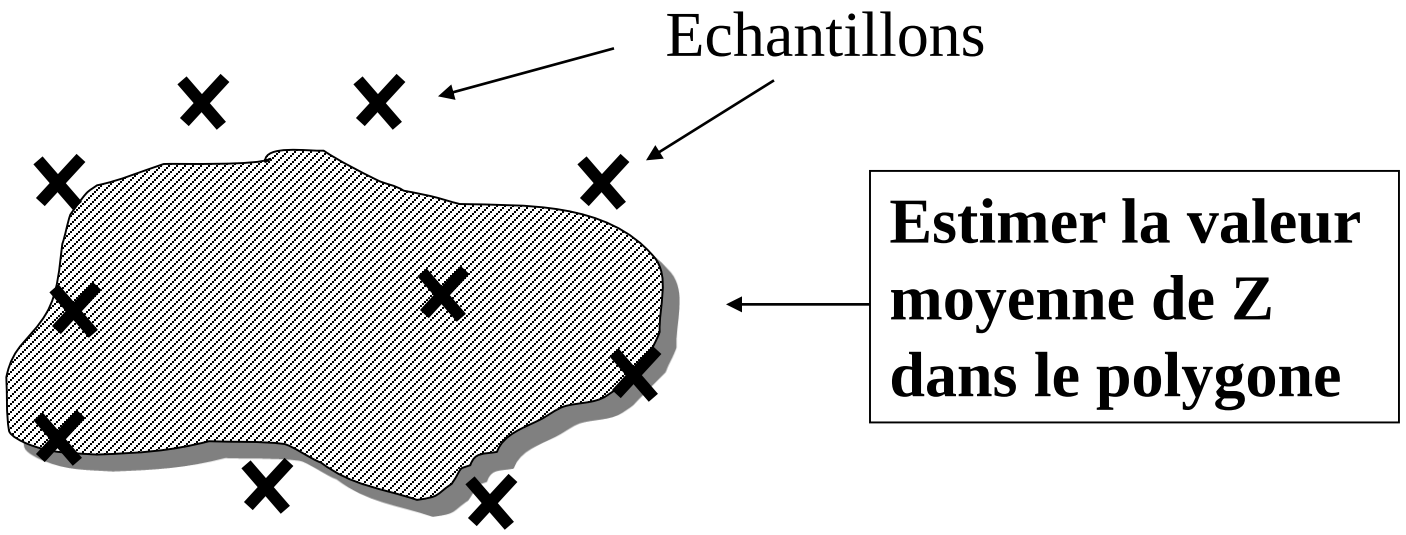
Modèle intrinsèque avec variogramme linéaire:
10 simulations, le krigeage ±un écart-type



Estimation globale (de biomasse)

Que veut-on estimer ?

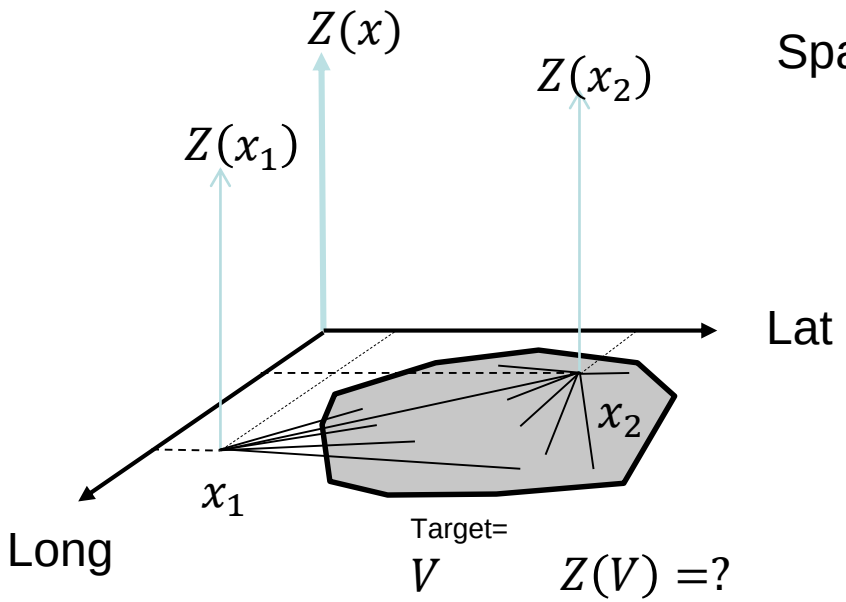
Estimation de blocs irréguliers



Global or block kriging

$$Z(V) = \frac{1}{V}\int_V Z(x)\,dx$$

Space integral of the local density over V.
Mean density over V.



To perform global kriging,
we need to know all the
spatial correlations
between samples points
and target area

$cov(Z_i, Z_V)$

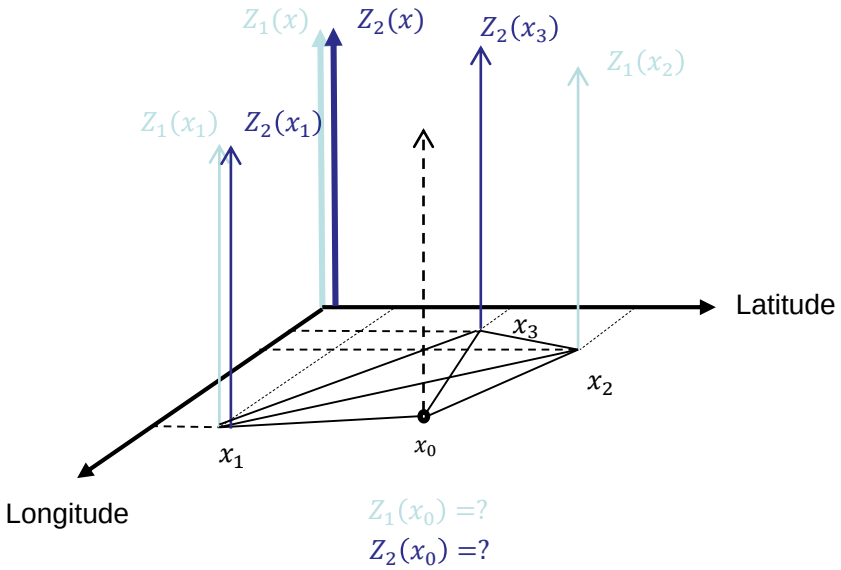
$$Z_V^{\boldsymbol{KRIGING}} = E[Z_V|Z_1, \ldots, Z_N] \approx \sum_{\alpha=1}^{N} \lambda_\alpha Z_\alpha + \lambda_0$$

Cela permet d’avoir une variance d’estimation

En effet, la moyenne des variance d’estimation de la carte n’est pas la variance d’estimation de la moyenne de la carte.
L’estimation de biomasse doit donc reposer sur une estimation globale.

La géostatistique multivariée en une seule diapo

Ponctual Co-Kriging



The requirement is now to know:

$$\begin{bmatrix} C_{1,1}(h) & C_{1,2}(h) \\ C_{2,1}(h) & C_{2,2}(h) \end{bmatrix}$$

i.e. the monovariate covariances and the cross-covariances (co-regionalization)

Co-Kriging configurations

Z_1 and Z_2 can be observed:

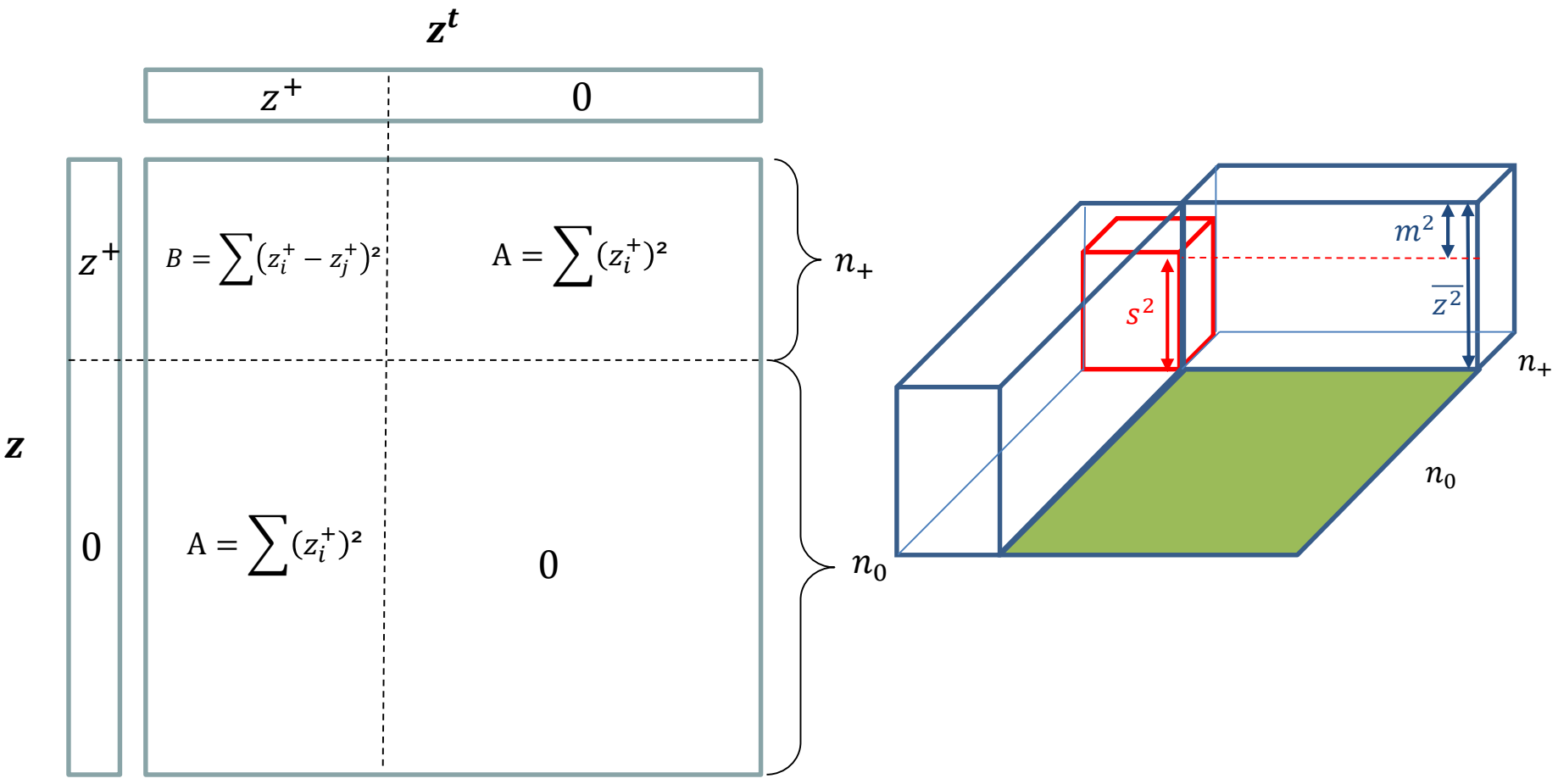
- always at the same points (homotopy),
- never at different points (heterotopy),
- at some similar points.

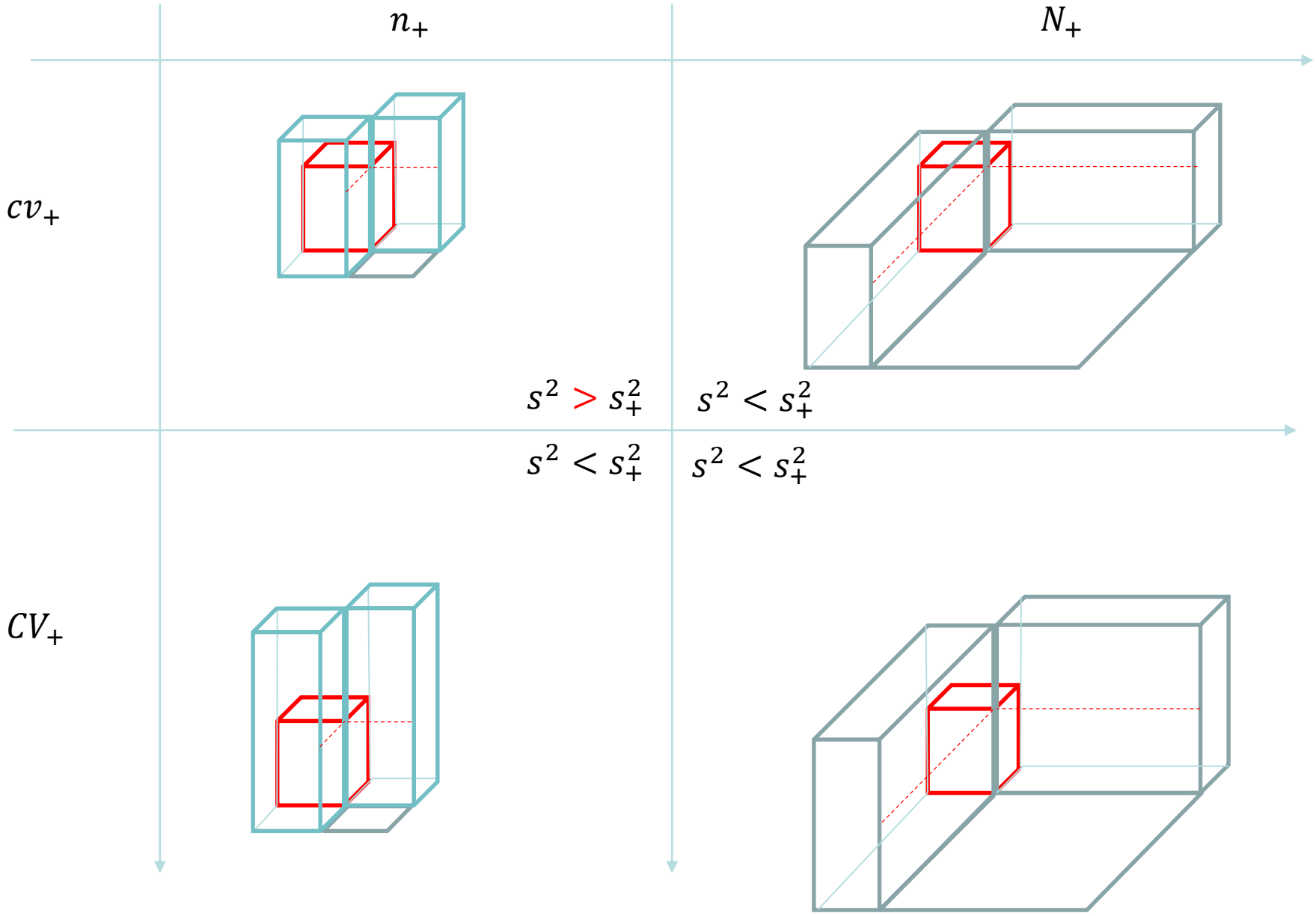
When estimating $z_{1,0}, z_{2,0}$ may or not be available.

$$Z_0^{CoK} = E\left[Z_0 \middle| Z_{1,1}, \dots, Z_{1,N_1}, Z_{2,1}, \dots, Z_{2,N_2}\right] \approx \sum_{\alpha=1}^{N_1} \lambda_{1,\alpha} Z_{1,\alpha} + \sum_{\alpha=1}^{N_2} \lambda_{2,\alpha} Z_{2,\alpha} + \lambda_0$$

Matériel Supplémentaire

Impact des 0


z^t
z^+
0
z^+
$B = \sum (z_i^+ - z_j^+)^2$
$A = \sum (z_i^+)^2$
n_+
z
0
$A = \sum (z_i^+)^2$
0
n_0
m^2
$\overline{z^2}$
s^2
n_+
n_0

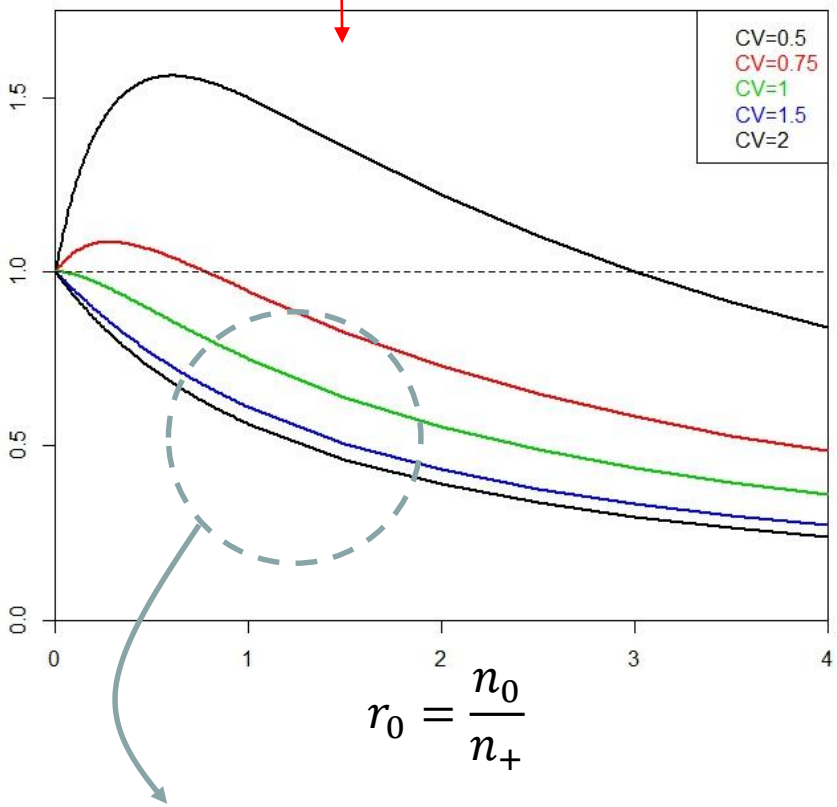

n_+
N_+
cv_+
$s^2 > s_+^2$
$s^2 < s_+^2$
$s^2 < s_+^2$
$s^2 < s_+^2$
CV_+

Impact des 0

$$s^2 = s_+^2 \frac{r_0 + (1 + r_0)CV_+^2}{CV_+^2(1 + r_0)^2}$$

$$s^2 = A \cdot s_+^2$$

z \ z^t	0	z^+
0	0	$A = \sum (z^+)^2$
z^+	$A = \sum (z^+)^2$	$B = \sum (z_i - z_j)^2$



$$r_0 = \frac{n_0}{n_+}$$

➔ La variance des données diminue souvent avec les 0

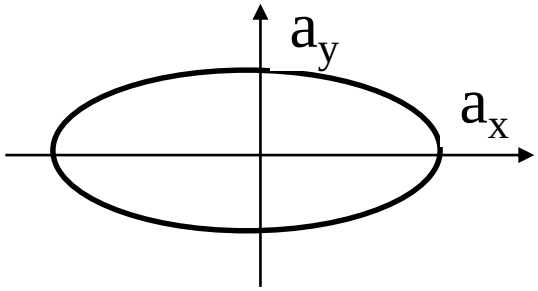
Anisotropie géométrique

γ(h)

palier

a_y a_x h

Ellipse d'anisotropie



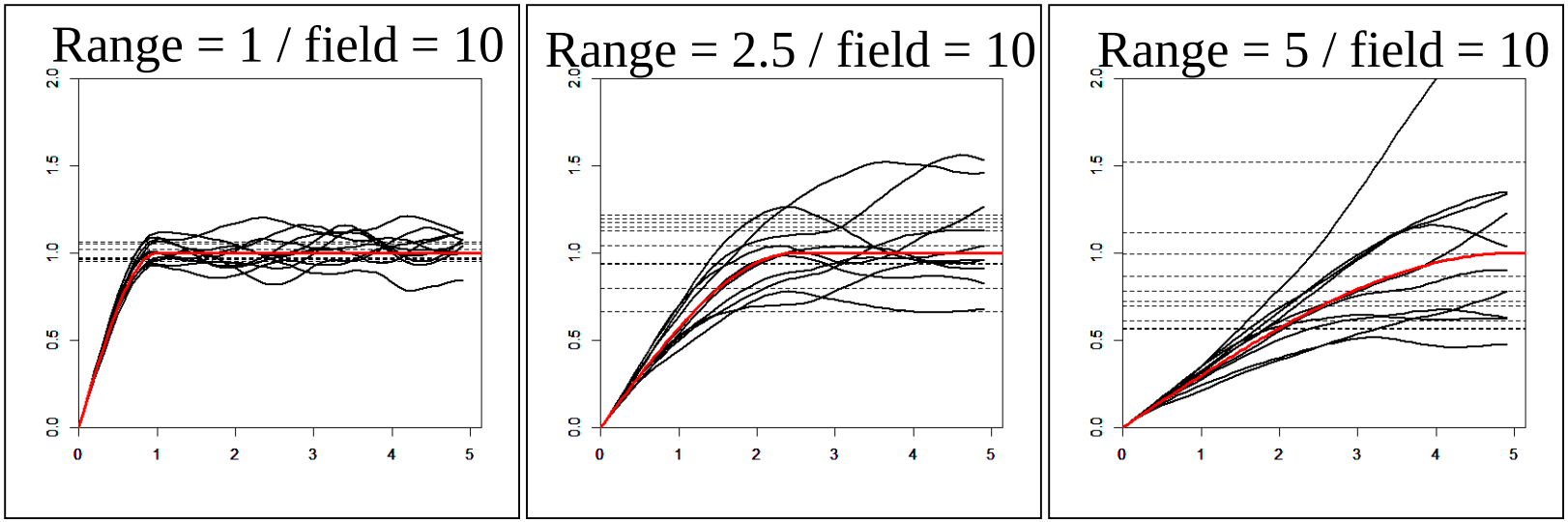
Variogrammes suivant les 2 directions principales

$$\gamma(h) = \gamma_0 \left(\sqrt{\frac{h_x^2}{a_x^2} + \frac{h_y^2}{a_y^2}} \right)$$

Lentilles minéralisées

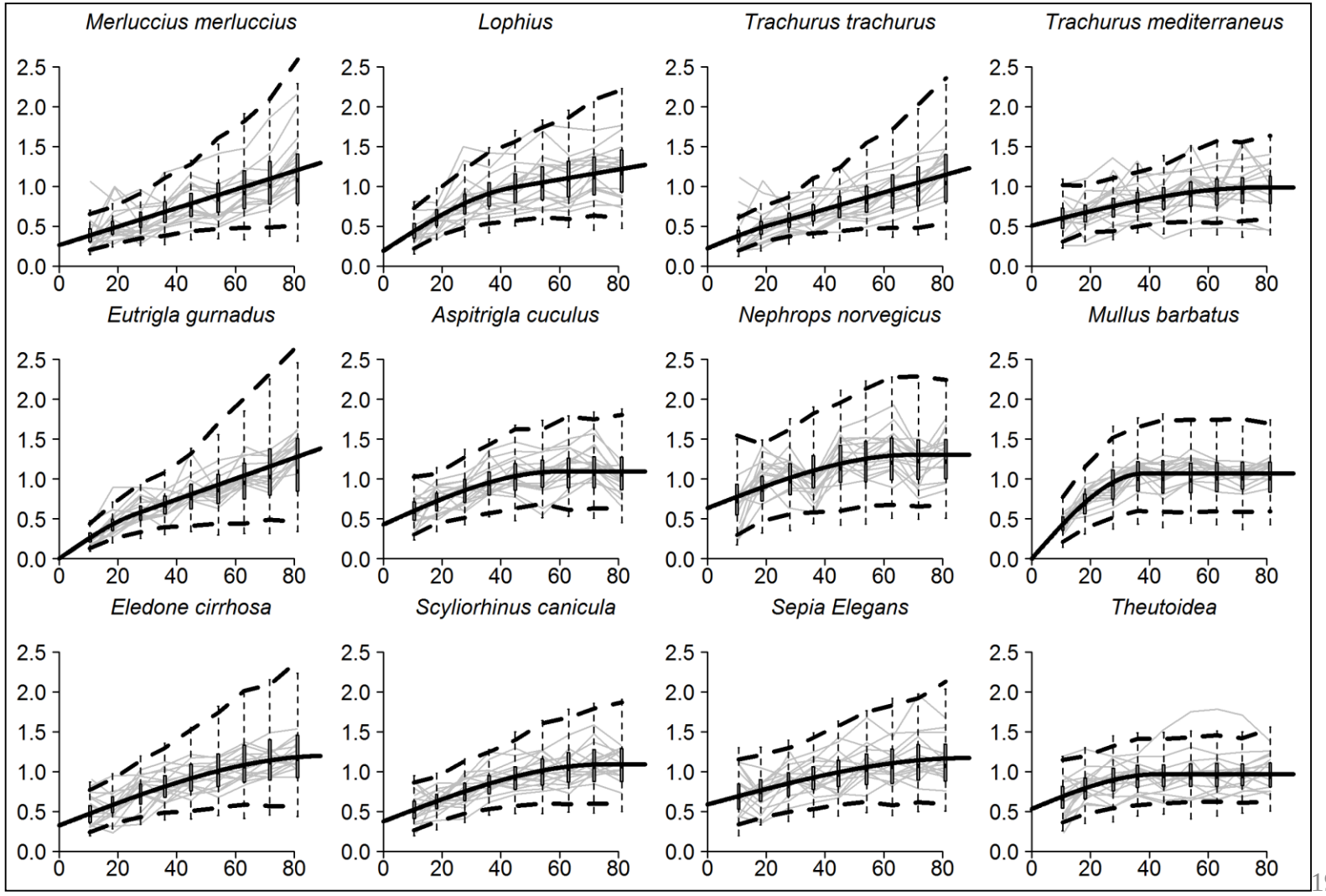
Statistical fluctuations

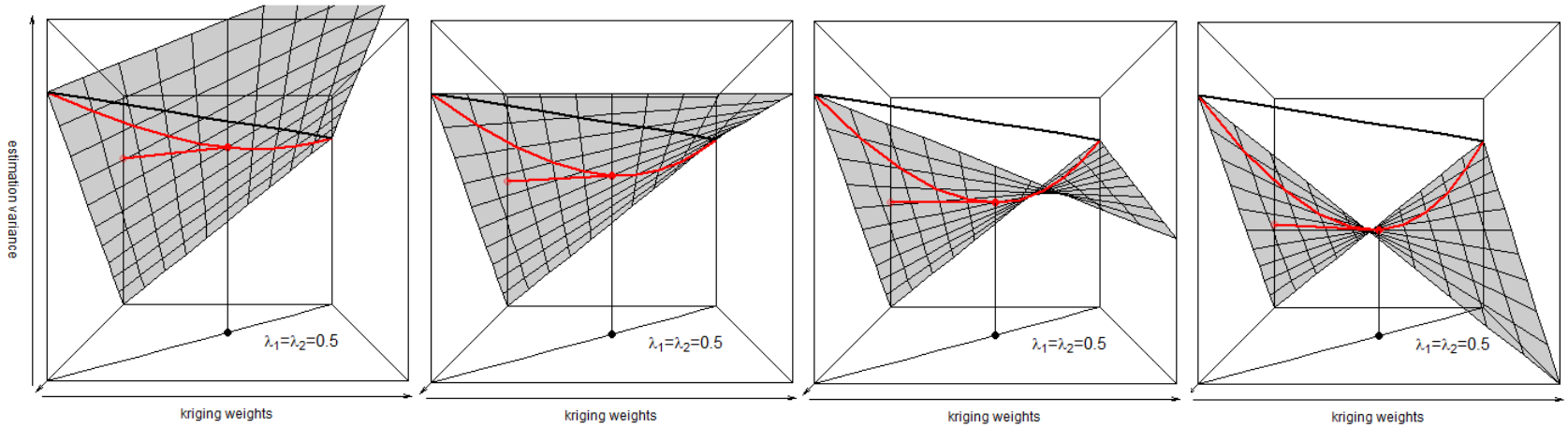
Natural statistical fluctuations of the variograms of different realisations of a given RF.



Statistical fluctuations decrease when the field increases wrt the range.

Comparison between actual and expected fluctuations




estimation variance
kriging weights
$\lambda_1=\lambda_2=0.5$
kriging weights
$\lambda_1=\lambda_2=0.5$
kriging weights
$\lambda_1=\lambda_2=0.5$
kriging weights
$\lambda_1=\lambda_2=0.5$