Introduction to Geostatistics

Hans Wackernagel, Laurent Bertino

Bjerknes Centre for Climate Research

18–20 October 2005, Bergen

Centre de Géostatistique - EMP

Founded 1968 by: Georges MATHERON

Director: Jean-Paul CHILÈS

Permanent staff: 14 research scientists

Funding (salaries): 60% by contract research,

$\hookrightarrow$ this conditions an application driven research

CG: Main application fields

- Petroleum exploration, mining
- Environmental sciences, climatology
- Health: epidemiology
- Fisheries, demography . . .

Software products: Isatis, Heresim. . .

sold by Geovariances International (www.geovariances.fr)

Also:

- Bioinformatics group (SVM, kernel methods)

Geostatistics worldwide

Other groups:

- Stanford (petroleum), Trondheim (petroleum), Calgary (mining, petroleum), Brisbane (mining), Johannesburg (mining), Valencia (hydrogeology),. . .

Main meetings:

- International Geostatistics Conference:
  1st in Rome (1975), . . . , 7th in Banff (2004)
  $\longrightarrow$ 2008: Santiago de Chile
- geoENV (european geostatistics conference for environmental applications):
  1st in Lisbon (1996),. . . , 5th in Neuchatel (2004)
  $\longrightarrow$ 2006: Greece

Software:

R (www.r-project.org), $\longrightarrow$ www.ai-geostats.org

Geostatistics

definition

Geostatistics

is an application of
the Theory of Regionalized Variables
(usually considered as realizations of Random Functions)

- to geology and mining (*fifties*)
- to natural phenomena in general (*seventies*)

- (re-)integrated mainstream statistics (*nineties*)

Concepts

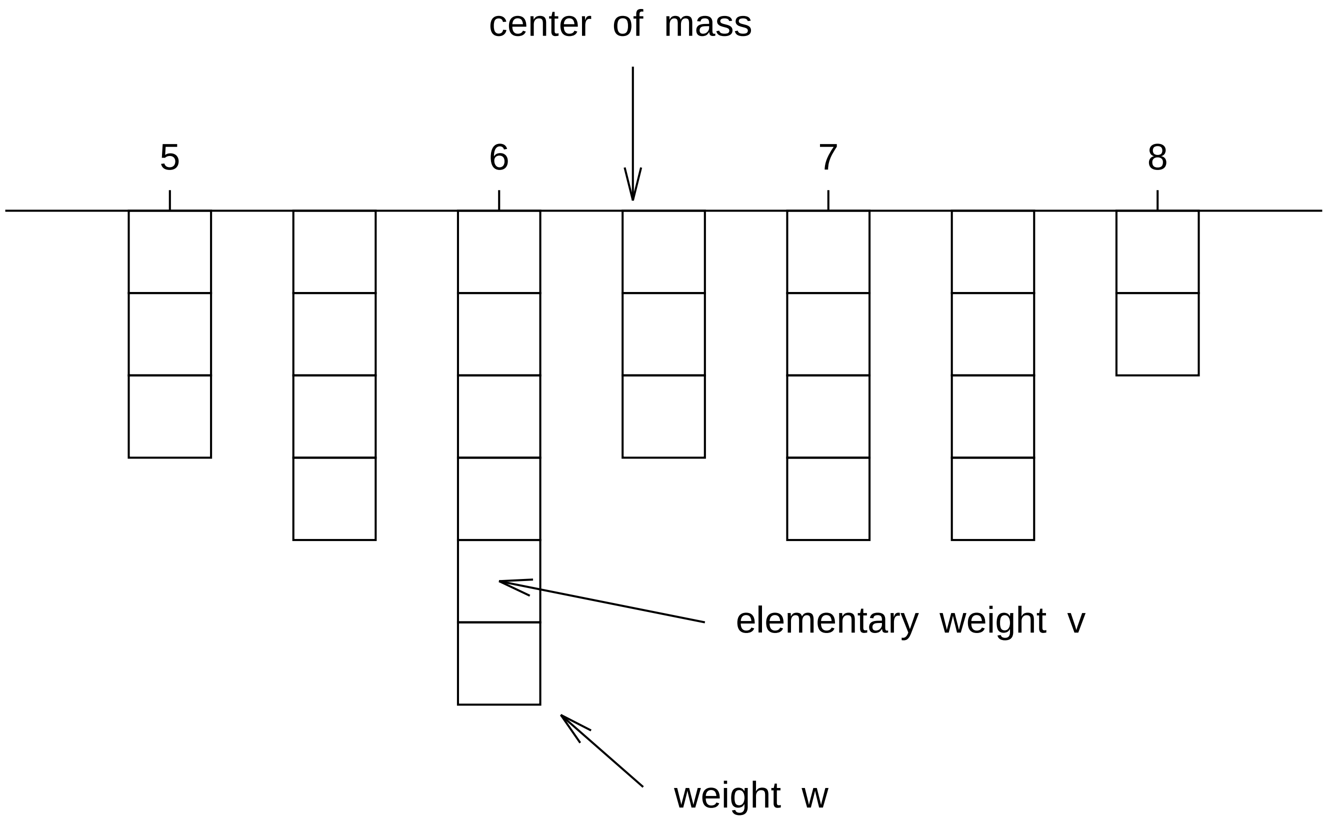
- Variogram: description of the spatial/temporal correlation of a phenomenon
- Kriging: optimal linear prediction method for estimating values of a phenomenon at any location of a region ($\longrightarrow$ D. G. KRIGE)
- Conditional Simulation: stochastic simulation of realizations, conditional upon the data.

Basic Statistics

concepts

Center of mass

Seven weights w are hanging on a bar whose own weight is negligible:



Center of mass

The weights w are suspended at points:

$$z \;=\; 5,\ 5.5,\ 6,\ 6.5,\ 7,\ 7.5,\ 8,$$

The mass $w(z)$ of the weights is

$$w(z) \;=\; 3,\ 4,\ 6,\ 3,\ 4,\ 4,\ 2.$$

The location $\overline{z}$ where the bar, when suspended, stays in equilibrium is:

$$\overline{z} \;=\; \frac{1}{\left(\sum_k w(z_k)\right)} \sum_{k=1}^{7} z_k\, w(z_k)$$

Center of mass

Defining normed weights:

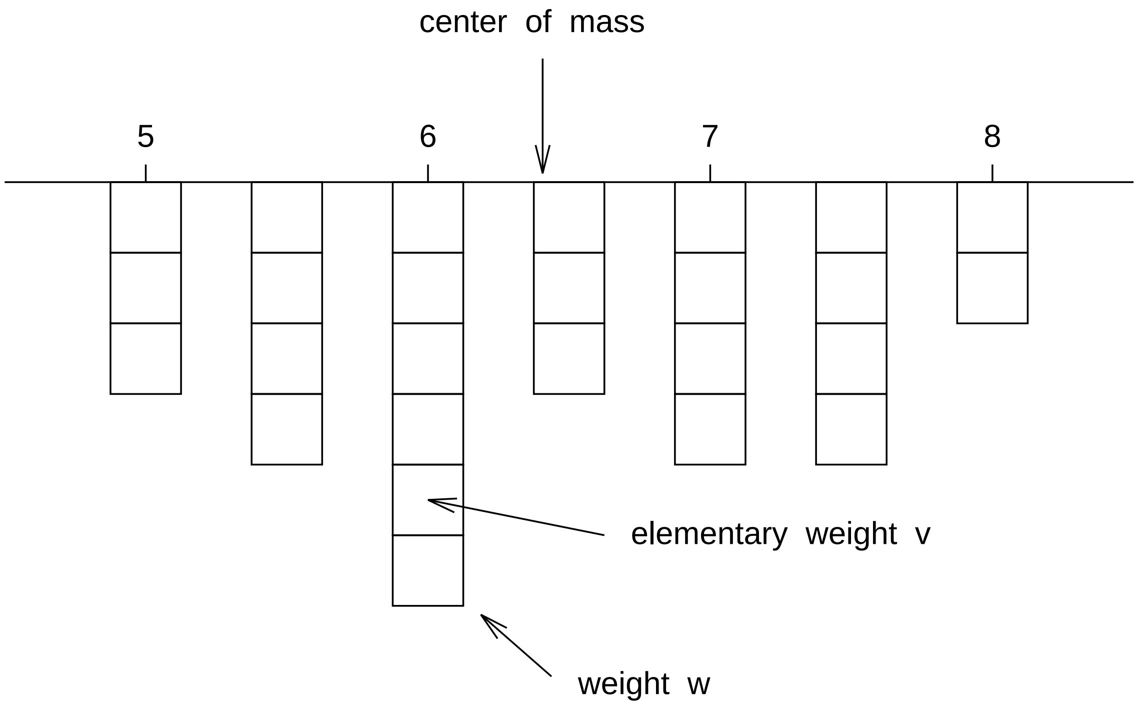
$$p(z_k) \quad = \quad \frac{w(z_k)}{\left(\sum_k w(z_k)\right)}$$

with $\sum_k p(z_k) = 1$, we can write:

$$\overline{z} \quad = \quad \sum_{k=1}^{7} z_k \, p(z_k)$$

Center of mass

The weights $w(z_k)$ are subdivided into n elementary weights v_α:



with corresponding normed weights $p_\alpha = 1/n$:

$$\bar{z} \;=\; \sum_{\alpha=1}^{n} z_\alpha \, p_\alpha = \frac{1}{26} \sum_{\alpha=1}^{26} z_\alpha = 6.4$$

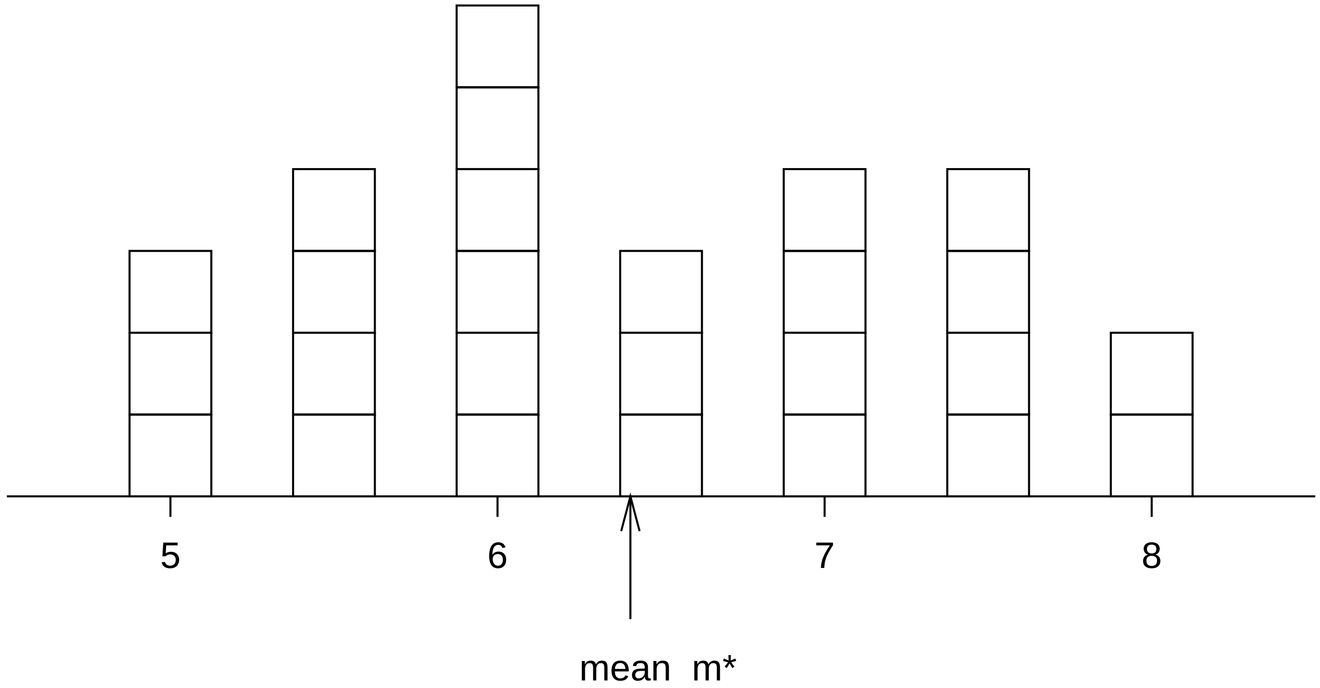
Center of mass

The average squared distance to the center of mass

$$\operatorname{dist}^2(\overline{z}) \quad = \quad \frac{1}{n}\sum_{\alpha=1}^{n}(z_\alpha - \overline{z})^2 = .83$$

gives an indication about the dispersion of the around the center of mass $\overline{z}$.

Histogram



The mean value $m^\star$ of data z_α is equivalently,

$$m^\star \;=\; \frac{1}{n}\sum_{\alpha=1}^{n} z_\alpha$$

Histogram

The average squared deviation from the mean is the *variance*

$$s^2 \quad = \quad \frac{1}{n} \sum_{\alpha=1}^{n} (z_\alpha - m^\star)^2$$

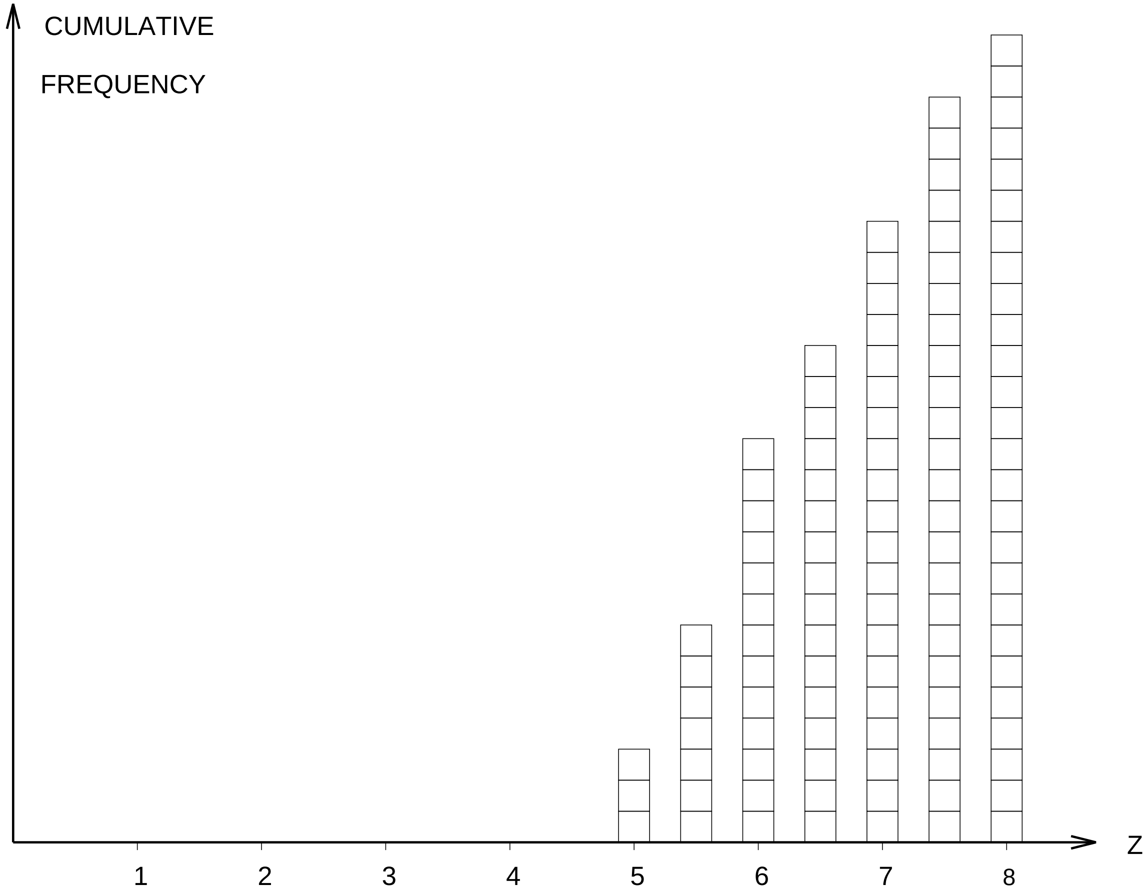
Its square-root is called the standard deviation.

The normalized weights $p(z_k)$ are the *frequencies* of the occurence of the values $z = 5,\ 5.5,\ 6,\ 6.5,\ 7,\ 7.5,\ 8$.

n is the number of samples.

Cumulative histogram

An alternate way to represent the frequencies of the values z is to cumulate them from left to right:



Probability distribution

Suppose we draw randomly values z from a set of values Z.

We call Z a random variable and z its realizations, $z \in \mathbb{R}$.

The mathematical idealization of the cumulative histogram is the *probability distribution function* $F(z)$ defined as:

$$F(z) \quad = \quad P(Z < z)$$

The probability $P(Z < z)$ indicates the theoretical frequency of drawing a realization lower than a given value z.

Probability density

We shall only consider differentiable distribution functions.

The derivative of the probability distribution function is the *probability density* $p(z)$:

$$F(dz) \quad = \quad p(z)\,dz$$

Properties:

$0 \leq p(z) \leq 1$

$\displaystyle\int p(z)\,dz = 1$

Expected value

The idealization of the concept of mean value is the *mathematical expectation*:

$$\mathrm{E}\left[\,Z\,\right] \;=\; \int_{z\,\in\,\mathbb{R}} z\,p(z)\,dz \;=m.$$

The expectation is a linear operator.
Let a, b be constants:

$$\mathrm{E}\left[\,a\,\right] \;=\; a, \qquad \mathrm{E}\left[\,b\,Z\,\right] \;=\; b\,\mathrm{E}\left[\,Z\,\right],$$

so that

$$\boxed{\mathrm{E}\left[\,a+b\,Z\,\right] \;=\; a+b\,\mathrm{E}\left[\,Z\,\right]}$$

Variance

The second moment of the random variable Z is:

$$\mathrm{E}[\,Z^2\,] \;=\; \int_{z\,\in\,\mathbb{R}} z^2\,p(z)\,dz$$

The variance σ^2 is defined as:

$$\mathrm{var}(Z) \;=\; \mathrm{E}\Big[\,(Z - \mathrm{E}[\,Z\,])^2\,\Big] \;=\; \mathrm{E}[\,(Z-m)^2\,] \;=\; \sigma^2$$

<u>Alternate expression:</u> multiplying out we get

$$\mathrm{var}(Z) \;=\; \mathrm{E}[\,Z^2 + m^2 - 2mZ\,]$$

and, as the expectation is a linear operator,

$$\boxed{\mathrm{var}(Z) \;=\; \mathrm{E}[\,Z^2\,] - \big(\mathrm{E}[\,Z\,]\big)^2}$$

Covariance

Covariance σ_{ij} between Z_i and Z_j:

$$\begin{aligned}
\operatorname{cov}(Z_i, Z_j) &= \mathrm{E}\Big[\, \big(Z_i - \mathrm{E}\big[\, Z_i \,\big]\big) \cdot \big(Z_j - \mathrm{E}\big[\, Z_j \,\big]\big) \,\Big] \\
&= \mathrm{E}\Big[\, (Z_i - m_i) \cdot (Z_j - m_j) \,\Big] \;=\; \sigma_{ij}
\end{aligned}$$

where m_i and m_j are the means of the random variables.

Covariance of Z_i with itself:

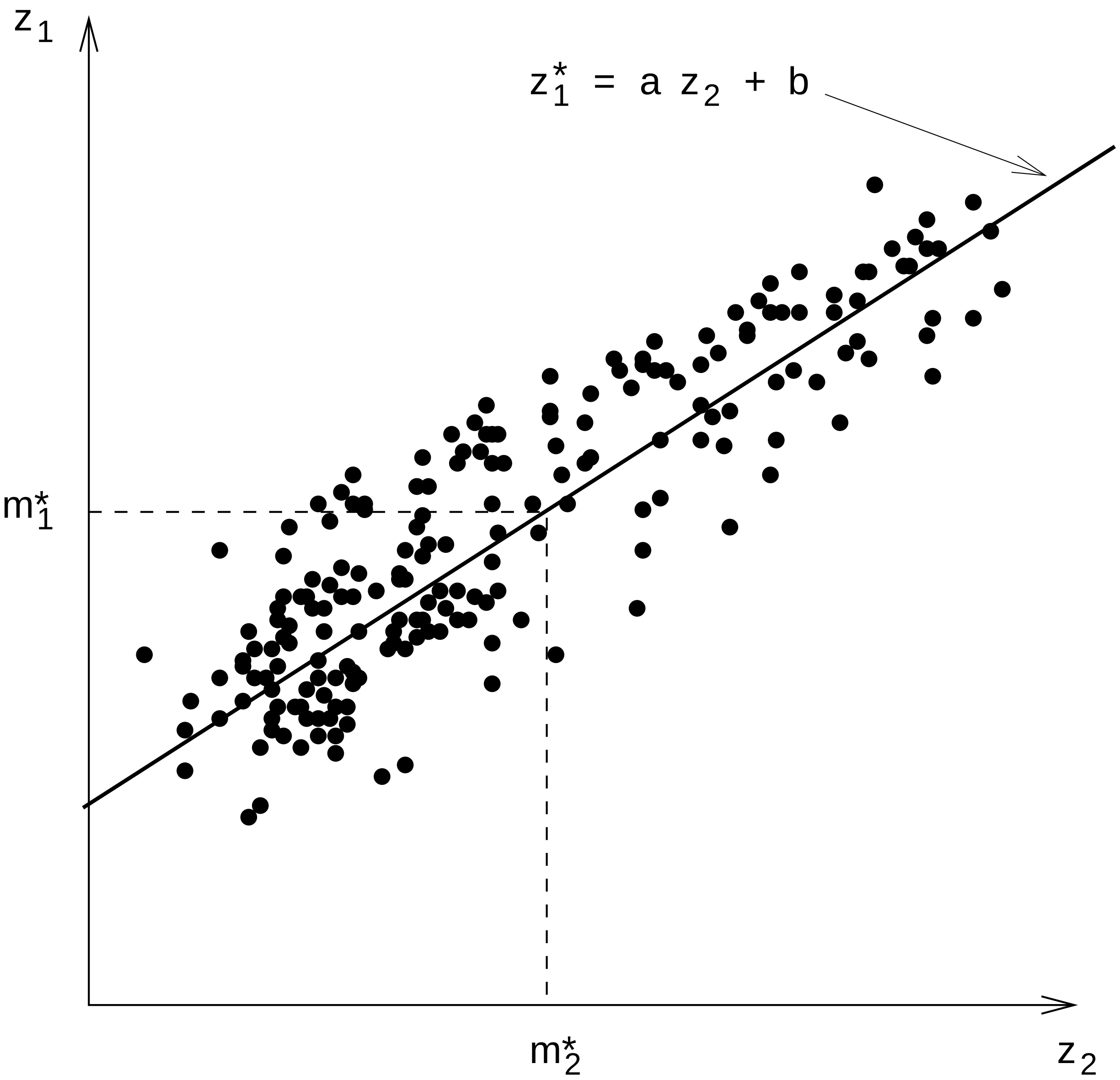
$$\sigma_{ii} \;=\; \mathrm{E}\Big[\, (Z_i - m_i)^2 \,\Big] \;=\; \sigma_i^2$$

Correlation coefficient:

$$\rho_{ij} \;=\; \frac{\sigma_{ij}}{\sqrt{\sigma_i^2\, \sigma_j^2}}$$

Linear regression

Regression line



Optimal regression line

Two variables with experimental covariance:

$$s_{12} = \frac{1}{n} \sum_{\alpha=1}^{n} (z_1^\alpha - m_1^\star) \cdot (z_2^\alpha - m_2^\star)$$

The regression line is:
with slope a and intercept b.

$$z_1^\star = a\, z_2 + b$$

Minimizing the quadratic distance:

$$\mathrm{dist}^2(a,b) = \frac{1}{n} \sum_{\alpha=1}^{n} (z_1^\alpha - a\, z_2^\alpha - b)^2$$

we get

$$\boxed{a = \frac{s_{12}}{s_2^2}} \qquad \boxed{b = m_1^\star - a\, m_2^\star}$$

Optimal regression line

$$\begin{aligned} z_1^\star &= \frac{s_{12}}{s_2^2}\,(z_2 - m_2^\star) + m_1^\star \\ &= m_1^\star + \frac{s_1}{s_2}\,r_{12}\,(z_2 - m_2^\star) \end{aligned}$$

At the minimum the squared distance is:

$$\boxed{\mathrm{dist}^2_{min}(a,b) \;=\; s_1^2\,\left(1 - (r_{12})^2\right)}$$

Multiple linear regression

Multivariate data set

The data matrix $\mathbf{Z}$ with n samples of N variables:

$$\begin{array}{cc} & Variables \\ Samples & \begin{pmatrix} z_{11} & \dots & z_{1i} & \dots & z_{1N} \\ \vdots & & \vdots & & \vdots \\ z_{\alpha 1} & \dots & z_{\alpha i} & \dots & z_{\alpha N} \\ \vdots & & \vdots & & \vdots \\ z_{n1} & \dots & z_{ni} & \dots & z_{nN} \end{pmatrix} \end{array}$$

Matrix of means

Define a matrix $\mathbf{M}$ with the same dimension $n \times N$ as $\mathbf{Z}$, replicating n times in its columns the mean value of each variable:

$$\mathbf{M} = \begin{pmatrix} m_1^\star & \dots & m_i^\star & \dots & m_N^\star \\ \vdots & & \vdots & & \vdots \\ m_1^\star & \dots & m_i^\star & \dots & m_N^\star \\ \vdots & & \vdots & & \vdots \\ m_1^\star & \dots & m_i^\star & \dots & m_N^\star \end{pmatrix}$$

Centered variables

A matrix $\mathbf{Z}_c$ of centered variables is obtained by subtracting $\mathbf{M}$ from the raw data matrix:

$$\mathbf{Z}_c \quad = \quad \mathbf{Z} \quad - \quad \mathbf{M}$$

Variance-covariance matrix

The matrix $\mathbf{V}$ of experimental variances and covariances is:

$$\mathbf{V} = \frac{1}{n}\mathbf{Z}_c^\top \mathbf{Z}_c = \begin{pmatrix} \mathrm{var}(\mathbf{z}_1) & \dots & \mathrm{cov}(\mathbf{z}_1, \mathbf{z}_j) & \dots & \mathrm{cov}(\mathbf{z}_1, \mathbf{z}_N) \\ \vdots & \ddots & & & \vdots \\ \mathrm{cov}(\mathbf{z}_i, \mathbf{z}_1) & \dots & \mathrm{var}(\mathbf{z}_i) & \dots & \mathrm{cov}(\mathbf{z}_i, \mathbf{z}_N) \\ \vdots & & & \ddots & \vdots \\ \mathrm{cov}(\mathbf{z}_N, \mathbf{z}_1) & \dots & \mathrm{cov}(\mathbf{z}_N, \mathbf{z}_j) & \dots & \mathrm{var}(\mathbf{z}_N) \end{pmatrix}$$

$$= \begin{pmatrix} s_{11} & \dots & s_{1j} & \dots & s_{1N} \\ \vdots & \ddots & & & \vdots \\ s_{i1} & \dots & s_{ii} & \dots & s_{iN} \\ \vdots & & & \ddots & \vdots \\ s_{N1} & \dots & s_{Nj} & \dots & s_{NN} \end{pmatrix}$$

Multiple linear regression

For a regression of $\mathbf{z}_0$ on the N variables from n samples we have the matrix equation

$$\mathbf{z}_0^\star = \mathbf{m}_0 + (\mathbf{Z} - \mathbf{M})\,\mathbf{a}$$

The squared distance between $\mathbf{z}_0$ and the hyperplane is:

$$\begin{aligned}\mathrm{dist}^2(\mathbf{a}) &= \frac{1}{n}(\mathbf{z}_0 - \mathbf{z}_0^\star)^\top(\mathbf{z}_0 - \mathbf{z}_0^\star) \\ &= \mathrm{var}(\mathbf{z}_0) + \mathbf{a}^\top\mathbf{V}\mathbf{a} - 2\,\mathbf{a}^\top\mathbf{v}_0,\end{aligned}$$

where $\mathbf{v}_0$ is the vector of covariances between $\mathbf{z}_0$ and $\mathbf{z}_i,\ i = 1, \ldots, N$.

Minimizing the squared distance

The minimum is found for:

$$\frac{\partial \, \mathrm{dist}^2(\mathbf{a})}{\partial \mathbf{a}} = 0 \quad \Longleftrightarrow \quad 2\,\mathbf{V}\mathbf{a} - 2\,\mathbf{v}_0 = 0 \quad \Longleftrightarrow \quad \mathbf{V}\mathbf{a} = \mathbf{v}_0$$

This system of linear equations:

$$\begin{pmatrix} \mathrm{var}(\mathbf{z}_1) & \ldots & \mathrm{cov}(\mathbf{z}_1, \mathbf{z}_N) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathbf{z}_N, \mathbf{z}_1) & \ldots & \mathrm{var}(\mathbf{z}_N) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} \mathrm{cov}(\mathbf{z}_0, \mathbf{z}_1) \\ \vdots \\ \mathrm{cov}(\mathbf{z}_0, \mathbf{z}_N) \end{pmatrix}$$

has exactly one solution,
if the determinant of $\mathbf{V}$ is different from zero.

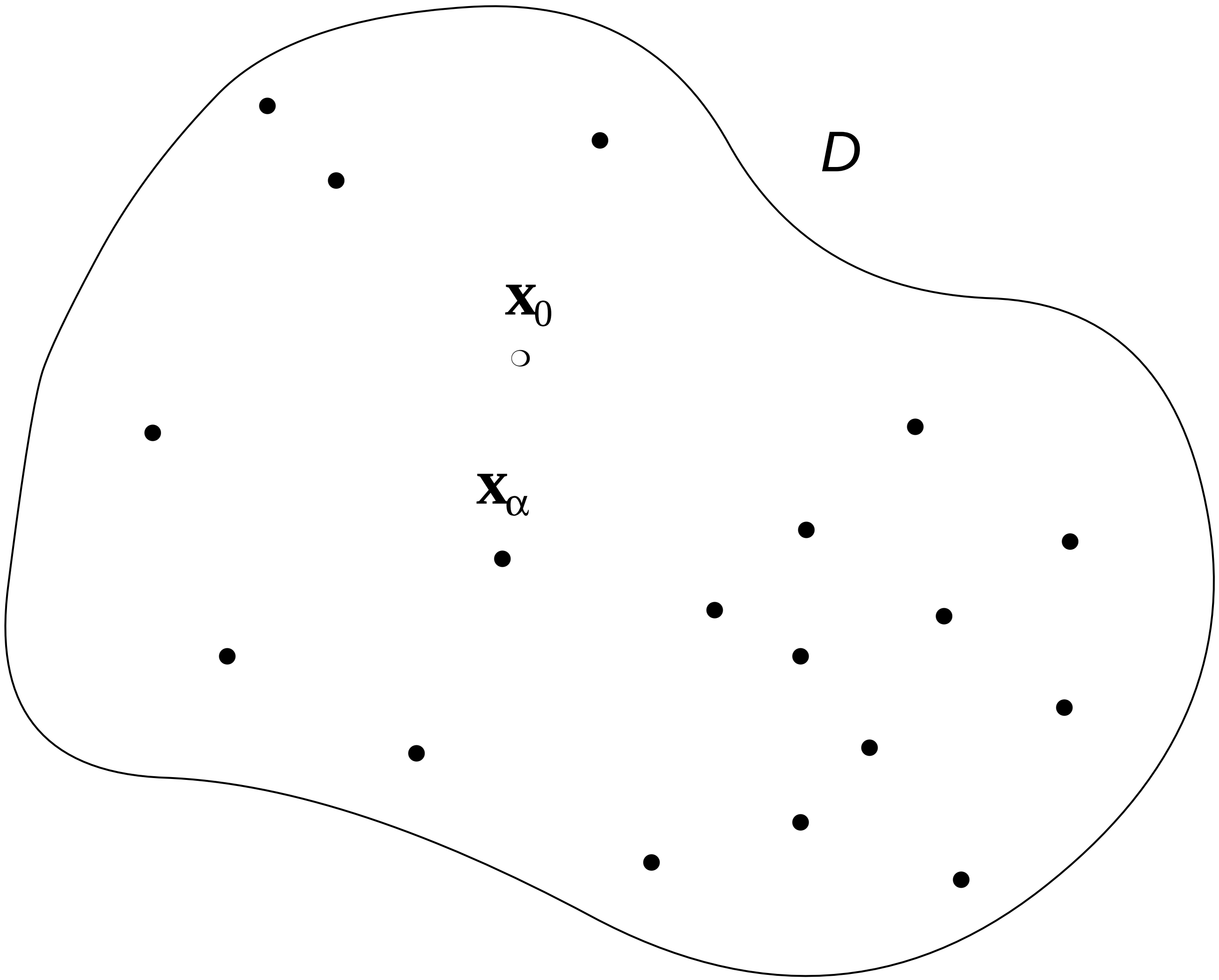
The squared distance at the minimum is:

$$\mathrm{dist}^2_{min}(\mathbf{a}) \;=\; \mathrm{var}(\mathbf{z}_0) - \mathbf{a}^\top \mathbf{v}_0$$

Simple kriging

Spatial data

Data points $\mathbf{x}_\alpha$ and the estimation point $\mathbf{x}_0$ in a spatial domain $\mathcal{D}$



Translation invariance

The expectation and the covariance are both assumed *translation invariant* over the domain,
i.e. for any vector $\mathbf{h}$ between points $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$:

$$\begin{aligned} \mathrm{E}\big[\,Z(\mathbf{x}+\mathbf{h})\,\big] = \mathrm{E}\big[\,Z(\mathbf{x})\,\big] &= m \\ \mathrm{cov}\Big(Z(\mathbf{x}+\mathbf{h}), Z(\mathbf{x})\Big) &= C(\mathbf{h}) \end{aligned}$$

- The expectation $\mathrm{E}\big[\,Z(\mathbf{x})\,\big]$ has the same value m at any point $\mathbf{x}$ of the domain $\mathcal{D}$.
- The covariance between any pair of locations depends only on the vector $\mathbf{h}$.

Known mean

We assume the mean m is known
and build the estimator:

$$
\begin{aligned}
Z^\star(\mathbf{x}_0) &= m + \sum_{\alpha=1}^{n} w_\alpha \Big(Z(\mathbf{x}_\alpha) - m \Big) \\
\text{i.e.} \qquad Z^\star(\mathbf{x}_0) - m &= \sum_{\alpha=1}^{n} w_\alpha \Big(Z(\mathbf{x}_\alpha) - m \Big)
\end{aligned}
$$

which is implicitly without bias:

$$
\mathrm{E}\Big[\, Z^\star(\mathbf{x}_0) - m \,\Big] = \sum_{\alpha=1}^{n} w_\alpha \, \mathrm{E}\Big[\, Z(\mathbf{x}_\alpha) - m \,\Big] = 0
$$

Simple kriging equations

The kriging equations with known mean are simple:

$$\sum_{\beta=1}^{n} w_{\beta}^{\mathrm{SK}}\, C(\mathbf{x}_\alpha - \mathbf{x}_\beta) = C(\mathbf{x}_\alpha - \mathbf{x}_0) \qquad \text{for} \quad \alpha = 1, \ldots, n$$

i.e.
the linear combination of weights with
the covariances between a data point
and the other data points
=
the covariance between that data point
and the point to estimate.

The variance of the Simple Kriging estimate is:

$$\sigma_{\mathrm{SK}}^2 \quad = \quad \sigma^2 - \sum_{\alpha=1}^{n} w_{\alpha}^{\mathrm{SK}}\, C(\mathbf{x}_\alpha - \mathbf{x}_0)$$

Simple kriging: a multiple linear regression

Simple kriging is a multiple linear regression between spatial random variables.

Like: $\mathbf{V}\mathbf{a} = \mathbf{v}_0,$ we have: $\mathbf{C}\mathbf{w} = \mathbf{c}_0$

Writing out the equation system:

$$\begin{pmatrix} \mathrm{var}(Z(\mathbf{x}_1)) & \dots & \mathrm{cov}(Z(\mathbf{x}_1), Z(\mathbf{x}_N)) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(Z(\mathbf{x}_N), Z(\mathbf{x}_1)) & \dots & \mathrm{var}(Z(\mathbf{x}_N)) \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix}$$

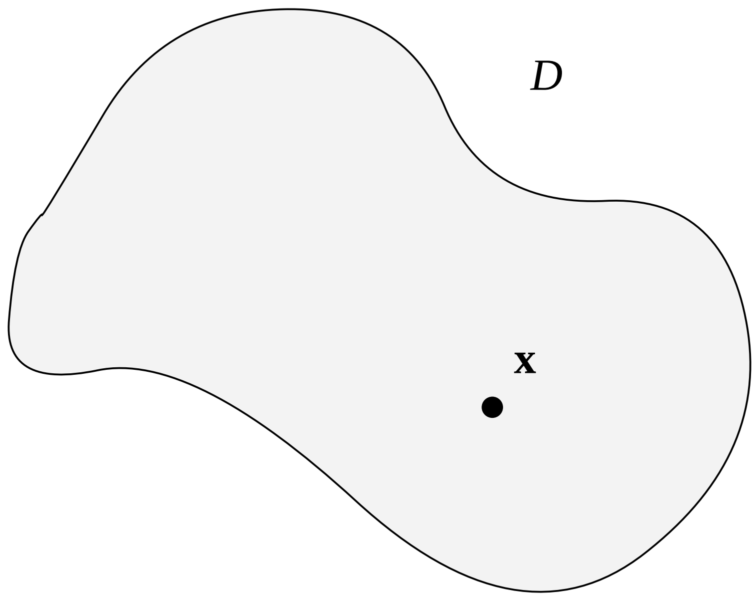
$$= \begin{pmatrix} \mathrm{cov}(Z(\mathbf{x}_0), \mathrm{cov}(Z(\mathbf{x}_1)) \\ \vdots \\ \mathrm{cov}(Z(\mathbf{x}_0), Z(\mathbf{x}_N)) \end{pmatrix}$$

Regionalized variables

and random function

The concept of a Random Function

Consider a domain $\mathcal{D}$ with points $\mathbf{x}$:



Let $Z(\mathbf{x})$ be a random variable at a location $\mathbf{x} \in \mathcal{D}$.
The family of random variables

$$\left\{ Z(\mathbf{x});\ \mathbf{x} \in \mathcal{D} \right\}$$

is called a *Random Function*.

Regionalized Variable

The *regionalized variable* $z(\mathbf{x})$ is the spatial variable of interest ("reality").

Data does not generally allow a deterministic reconstruction of the regionalized variable.

Probabilistic approach:

The regionalized variable $z(\mathbf{x})$ is considered as a realization (draw) of a random function $Z(\mathbf{x})$.

For a given data set, different realizations containing the data are equally plausible to represent the regionalized variable.

Epistemological Problem

We possess data about only one realization:
how can we specify the random function?

Objective quantities that describe the regionalized variable and conventional parameters that are constitutive of the model have to be distinguished.

The quantities are estimated from data,

but the parameters are chosen.

$\longrightarrow$ G Matheron "Estimating and Choosing" (1989)

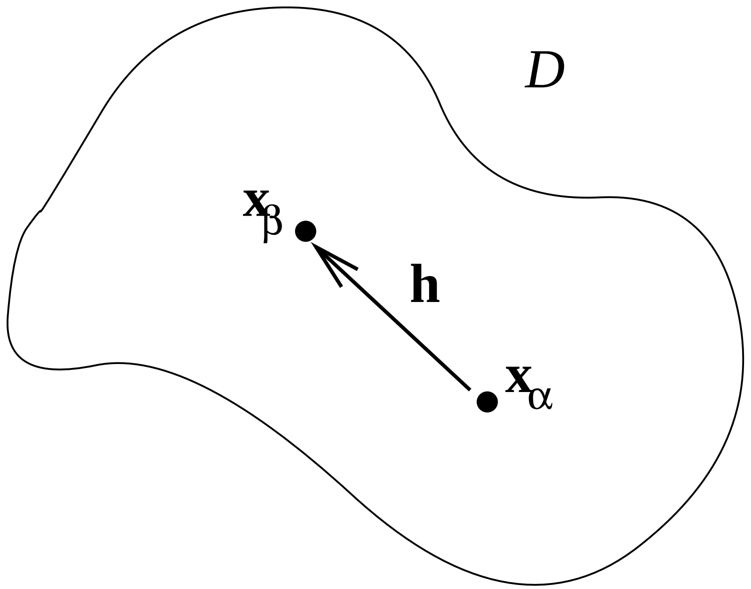
Variogram

definition

The Variogram

The vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$: coordinates of a point in 2D.

Let $\mathbf{h}$ be the vector separating two points:

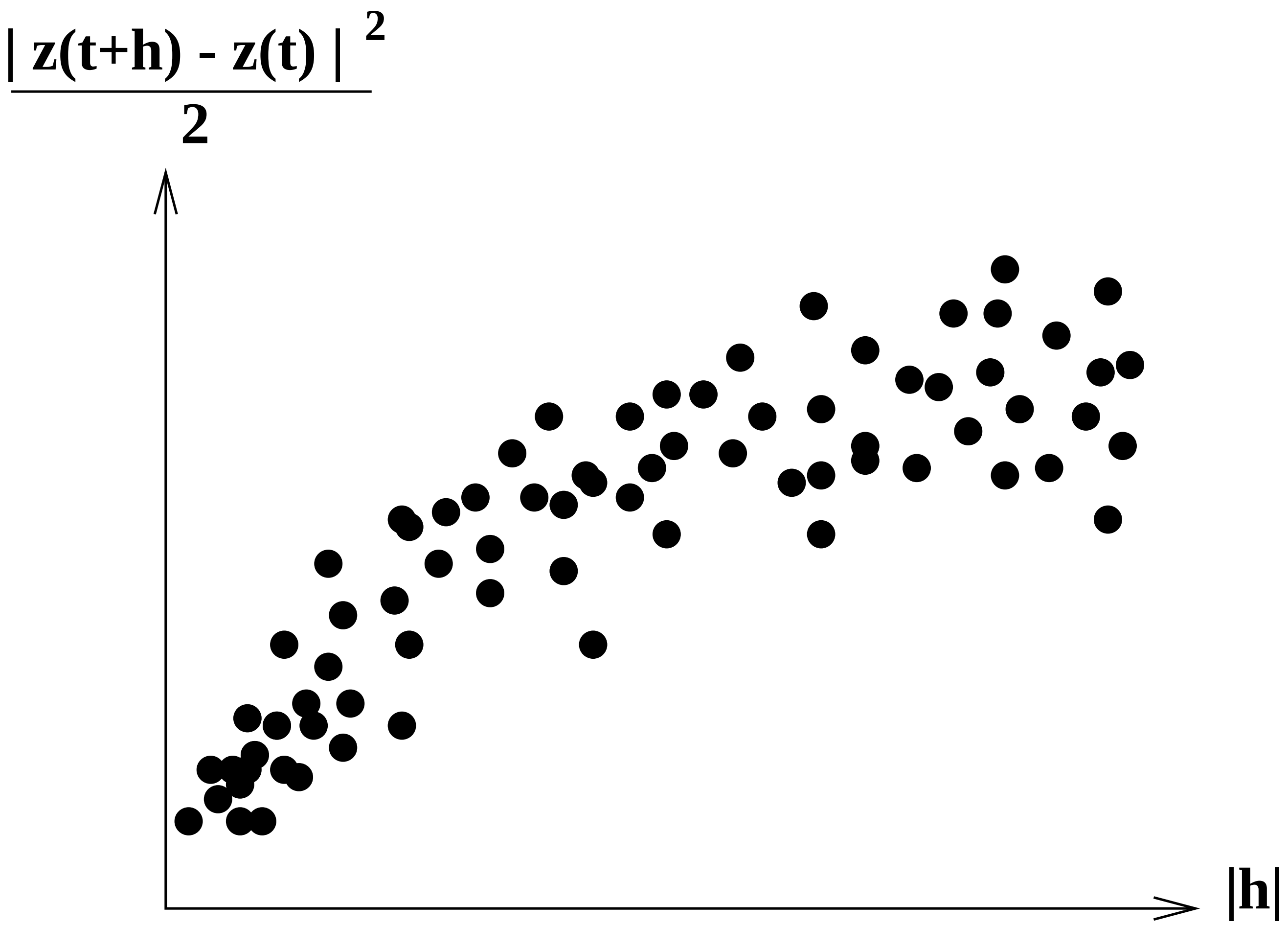


We compare sample values z at a pair of points with:

$$\frac{\Big(z(\mathbf{x}+\mathbf{h}) - z(\mathbf{x}) \Big)^2}{2}$$

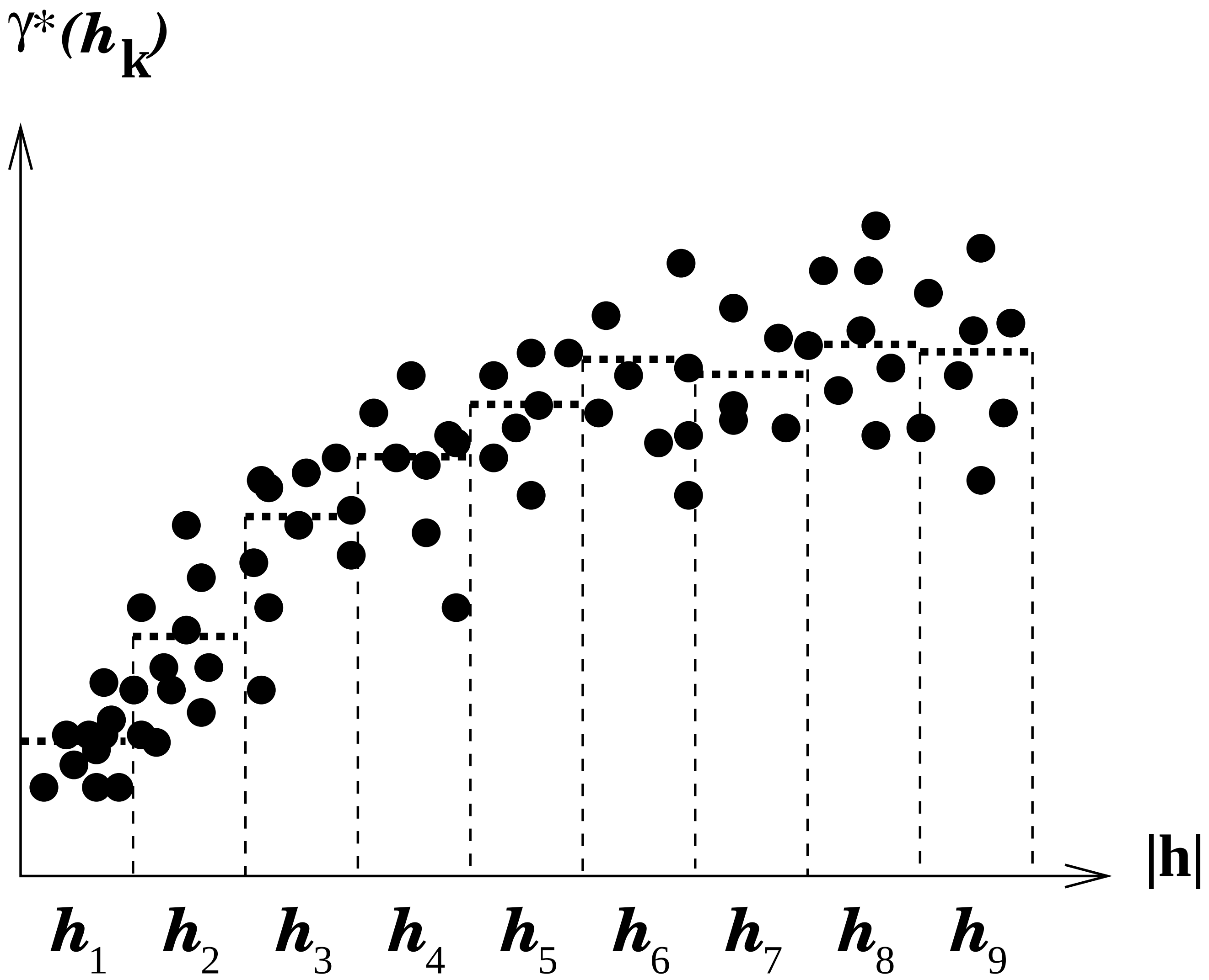
The Variogram Cloud

Variogram values are plotted against distance in space:



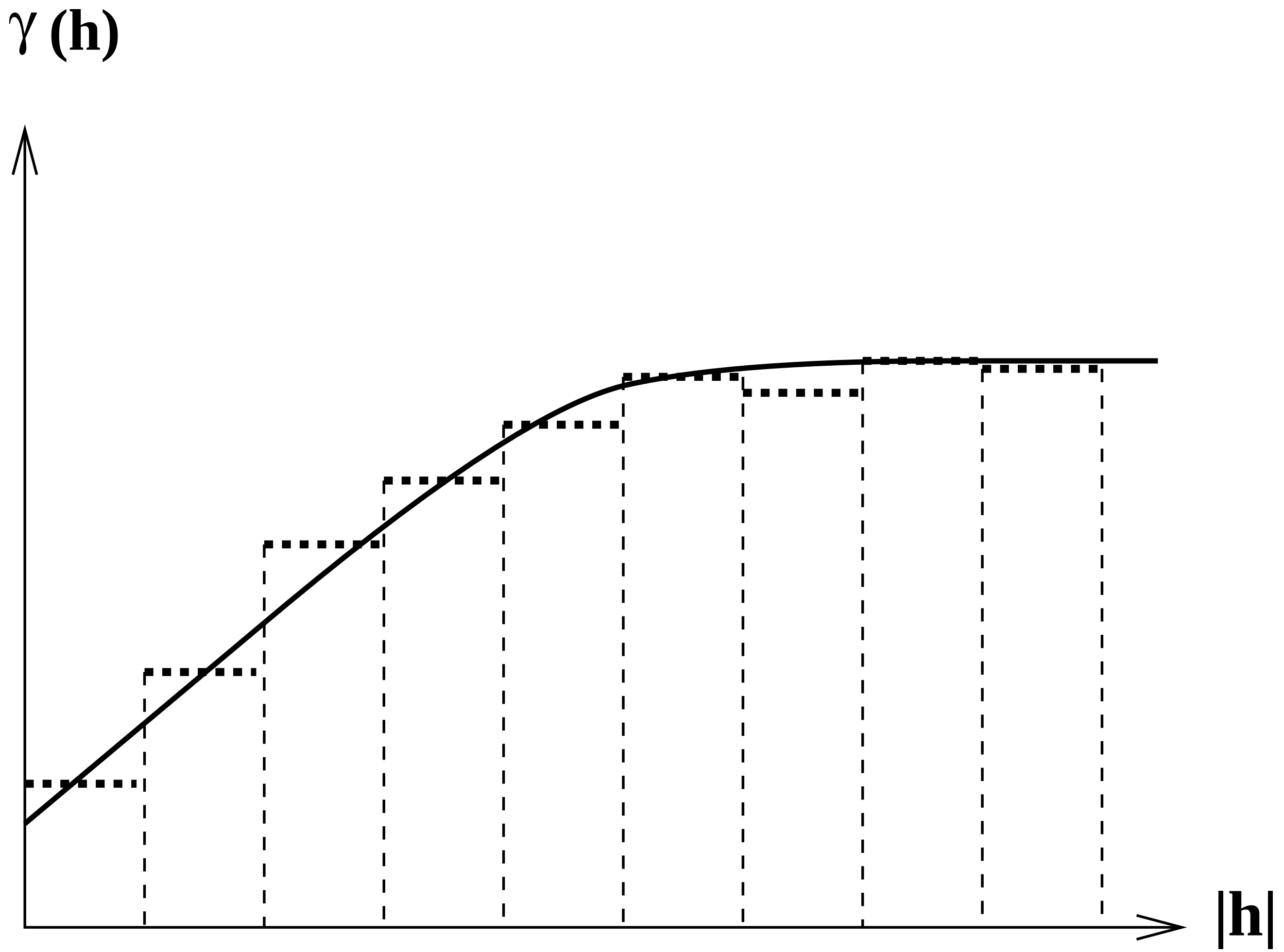
The Experimental Variogram

Averages within distance (and angle) classes h_k are computed:



The Theoretical Variogram

A theoretical model is fitted:



Intrinsic Hypothesis

The first two moments of the increments are assumed stationary (translation-invariant):

- the expectation does not depend on $\mathbf{x}$

$$\mathrm{E}\Big[\,Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\,\Big] = 0$$

- the variance depends only on $\mathbf{h}$

$$\mathrm{var}\Big[\,Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\,\Big] = 2\,\gamma(\mathbf{h})$$

This type of stationarity is called intrinsic.

$\hookrightarrow$ The stationarity of the increments does not imply the stationarity of Z.

Definition of the Variogram

By the intrinsic hypothesis:

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

Properties

- zero at the origin $\qquad \gamma(0) \;=\; 0$
- positive values $\qquad \gamma(\mathbf{h}) \;\geq\; 0$
- even function $\qquad \gamma(\mathbf{h}) \;=\; \gamma(-\mathbf{h})$

Regionalized variable		Behavior at the origin
smooth	$\longleftrightarrow$	continuous and differentiable
rough	$\longleftrightarrow$	not differentiable
speckled	$\longleftrightarrow$	discontinuous

Variogram and Covariance Function

The covariance function is defined as:

$$C(\mathbf{h}) \;=\; \mathrm{E}\Big[\Big(Z(\mathbf{x})-m\Big)\cdot\Big(Z(\mathbf{x}+\mathbf{h})-m\Big)\Big]$$

where stationarity of the first two moments of Z is assumed.

A variogram can be constructed from any covariance function:

$$\gamma(\mathbf{h}) \;=\; C(0)-C(\mathbf{h})$$

Conversely, however, only if the variogram is bounded does a corresponding covariance function $C(\mathbf{h})$ exist.

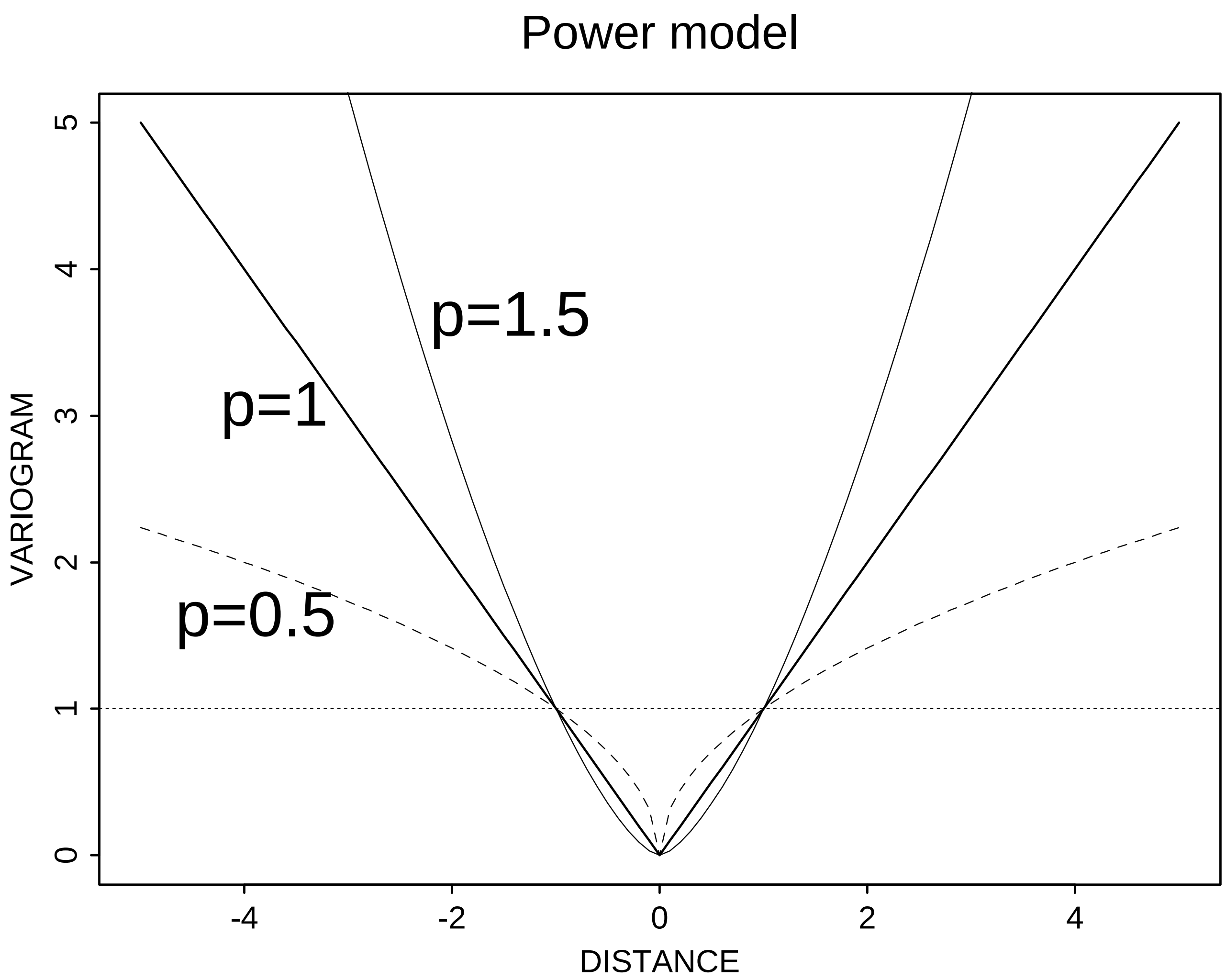
The variogram characterizes a larger class of random functions. This is why it is preferred in geostatistics.

Variogram

examples

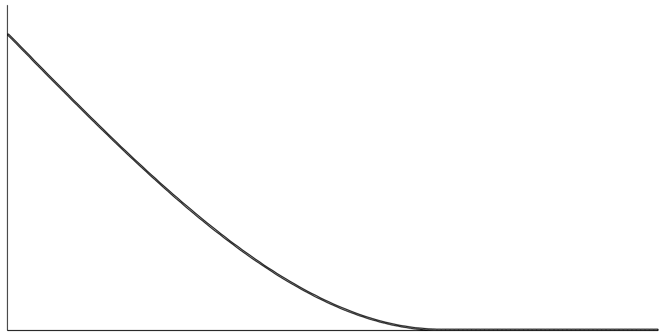
Power variogram

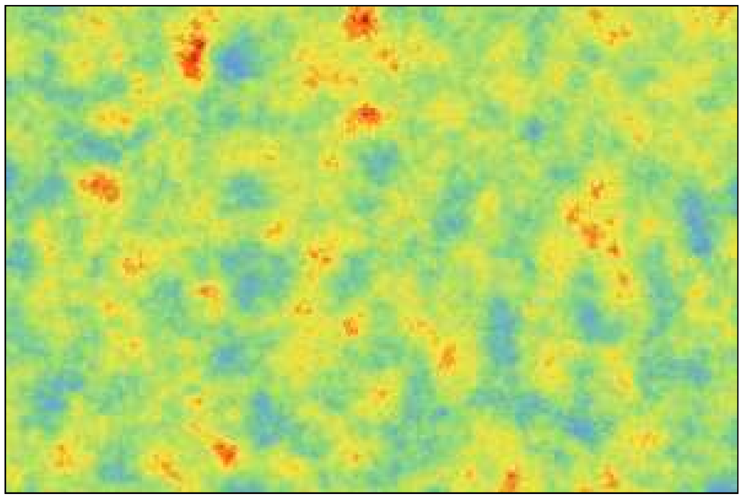
$$\gamma(\mathbf{h}) = |\mathbf{h}|^p, \qquad 0 < p \leq 2$$

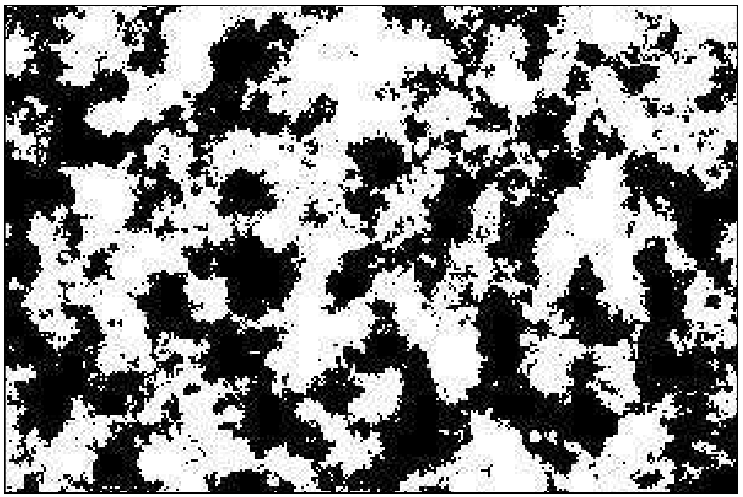


Spherical covariance function

$$C(\mathbf{h}) = \left(\frac{3}{2} \frac{|\mathbf{h}|}{a} - \frac{1}{2} \frac{|\mathbf{h}|^3}{a^3} \right) \mathbf{1}_{|\mathbf{h}| \leq a}$$

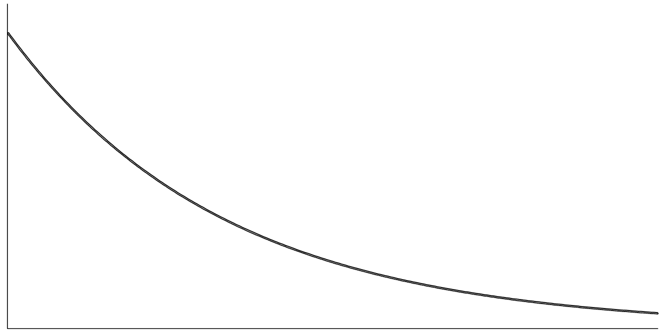


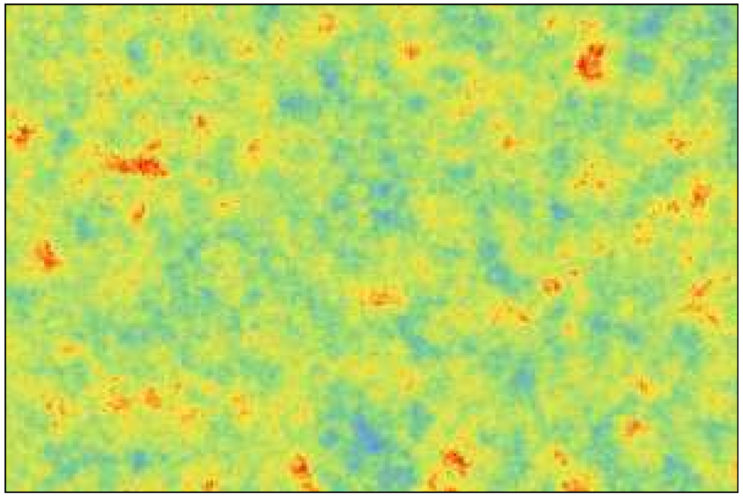


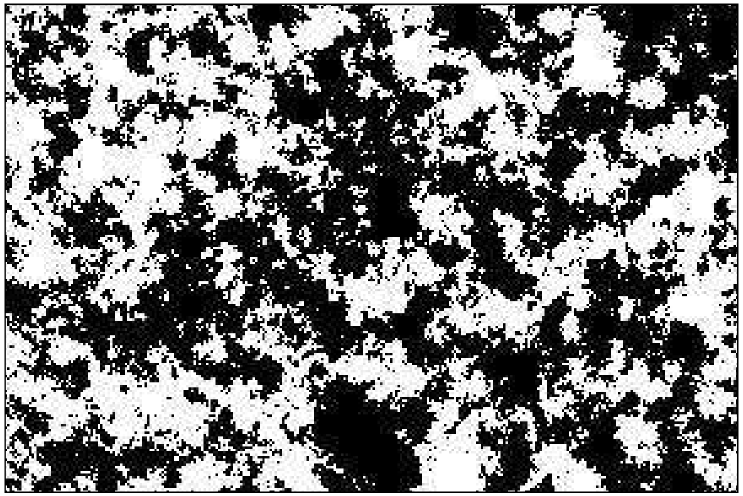


Exponential covariance function

$$C(\mathbf{h}) = \exp\left(-\frac{|\mathbf{h}|}{a}\right)$$

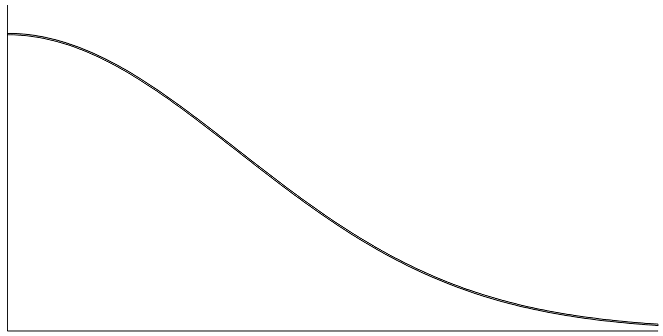


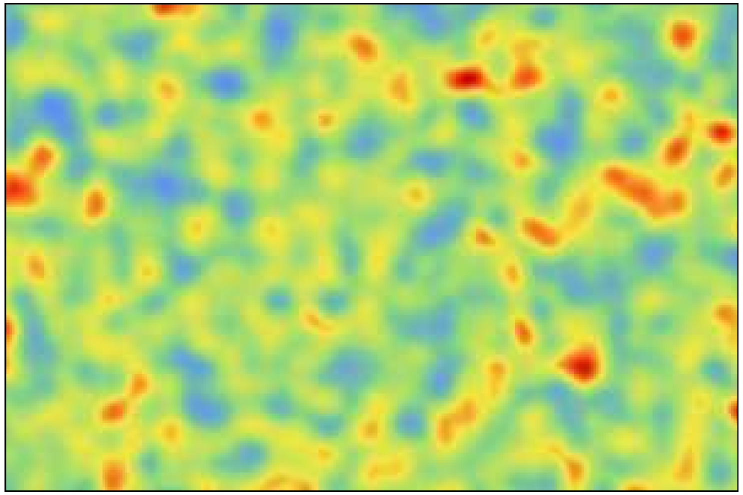


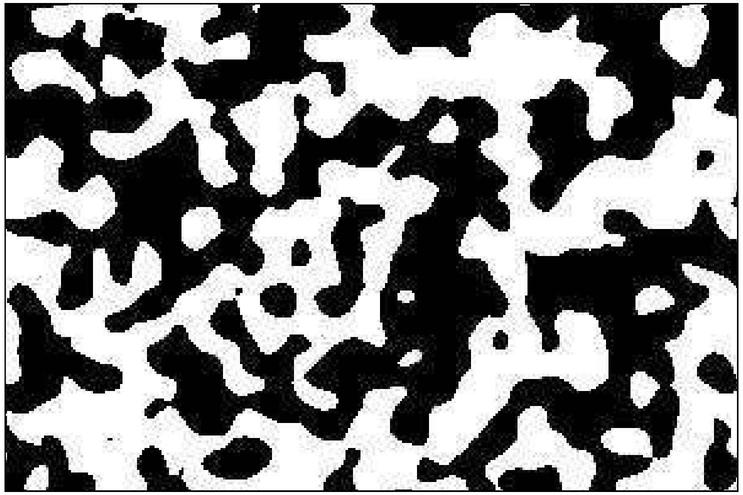


Gaussian covariance function

$$C(\mathbf{h}) = \exp\left(-\frac{|\mathbf{h}|^2}{a^2}\right)$$

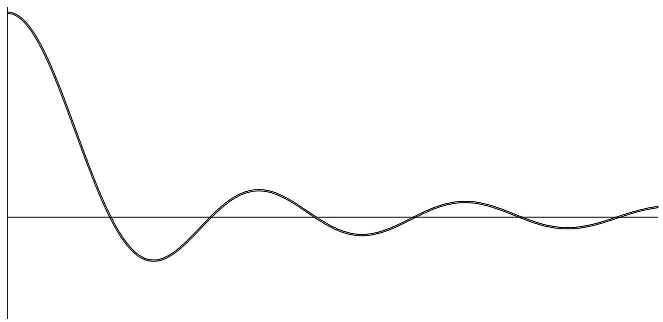


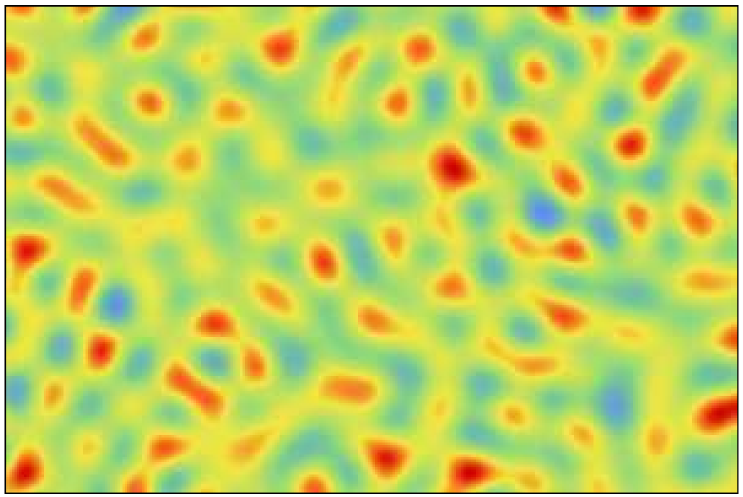


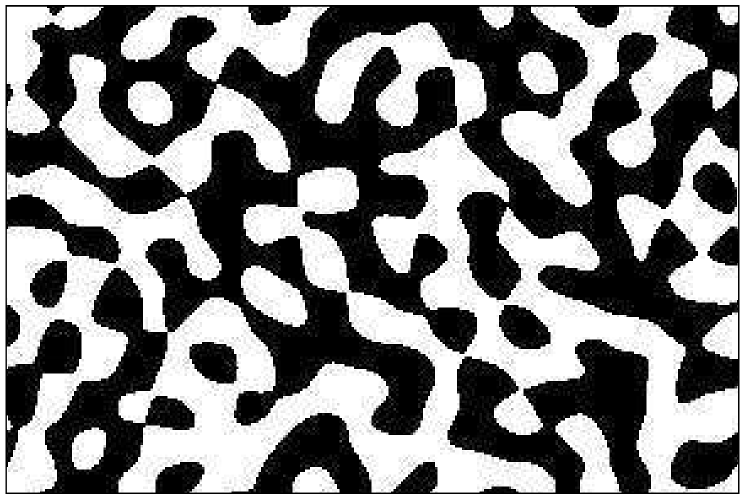


Cardinal sine covariance function

$$C(\mathbf{h}) = \frac{\sin\left(\frac{|\mathbf{h}|}{a}\right)}{\frac{|\mathbf{h}|}{a}}$$





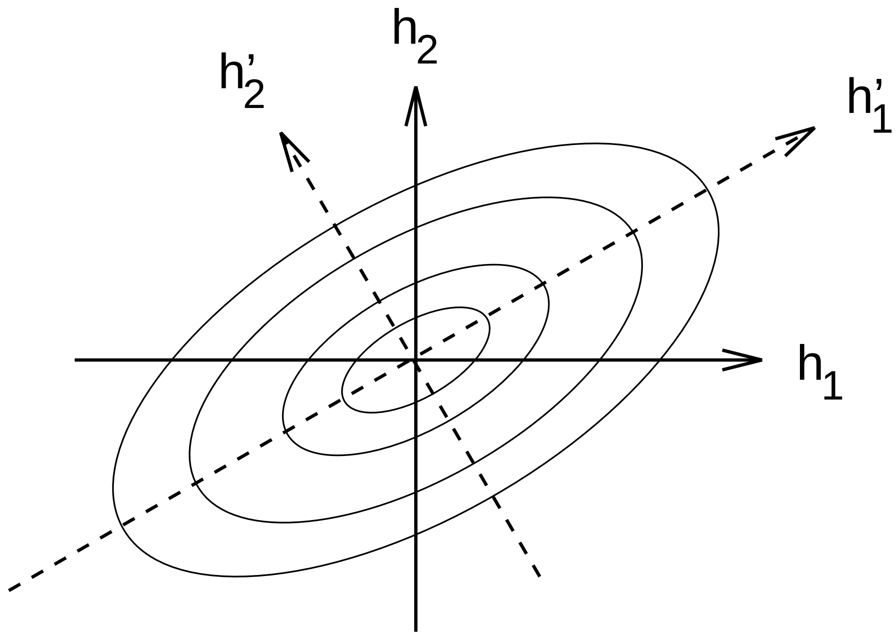


Geometric anisotropy

of the variogram

Geometric anisotropy

In practice the *range* of the variogram may change depending on the direction:



Correction:

- rotation $\mathbf{h}' = \mathbf{Q}\mathbf{h}$ of angle θ where $\mathbf{Q} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$
- linear transformation of the coordinates $\mathbf{h}' = (h'_1, h'_2)$

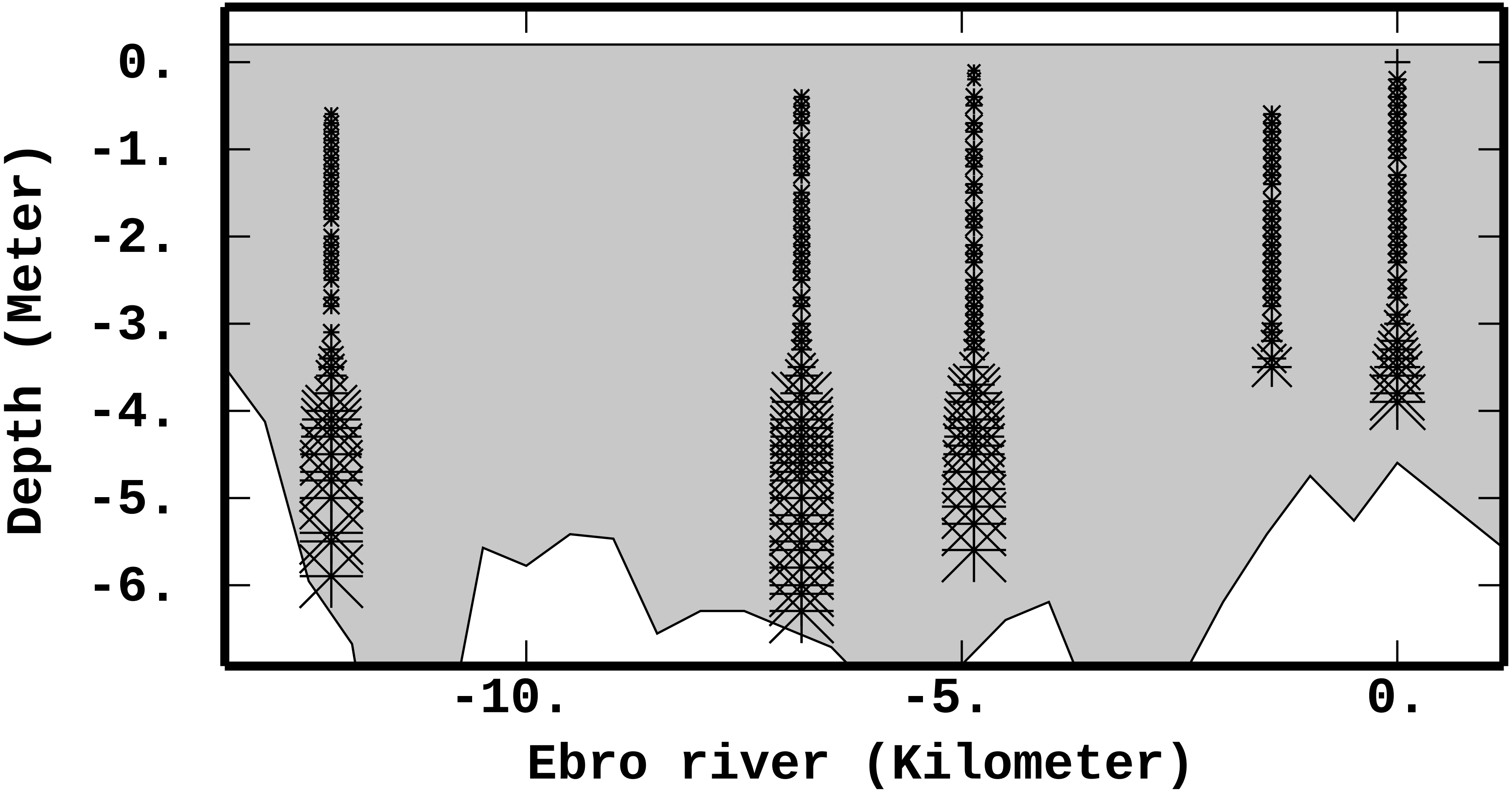
Rotation in 3D

In 3D the rotation is obtained by a composition of elementary rotations:

$$\mathbf{Q} = \begin{pmatrix} \cos\theta_3 & \sin\theta_3 & 0 \\ -\sin\theta_3 & \cos\theta_3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ \cos\theta_2 & \sin\theta_2 & 0 \\ -\sin\theta_2 & \cos\theta_2 & 0 \end{pmatrix} \begin{pmatrix} \cos\theta_1 & \sin\theta_1 & 0 \\ -\sin\theta_1 & \cos\theta_1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

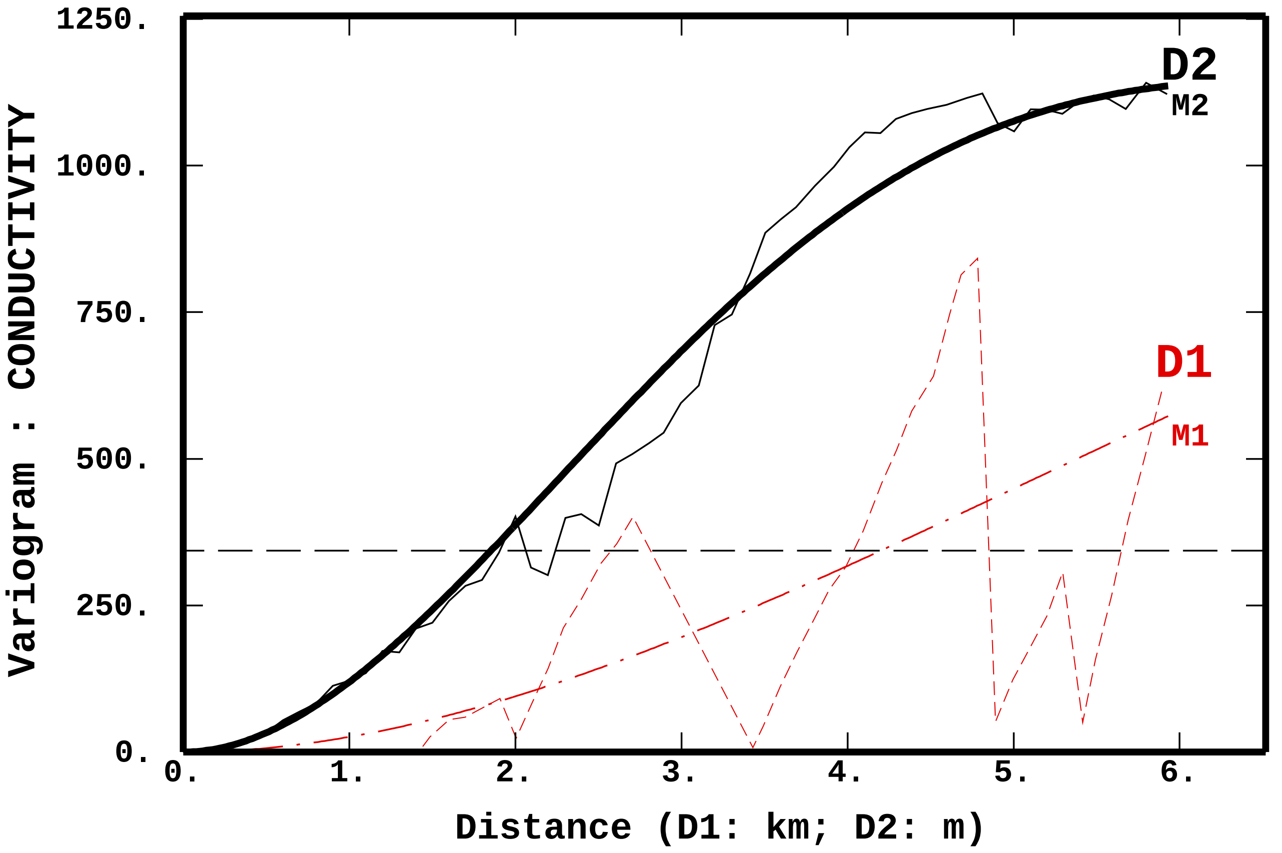
where θ_1, θ_2, θ_3 are Euler's angles.

2D example: Ebro river vertical section



185 Hydrolab Surveyor III conductivity measurements

2D conductivity variogram model



Experimental variogram for D1=horizontal, D2=vertical.

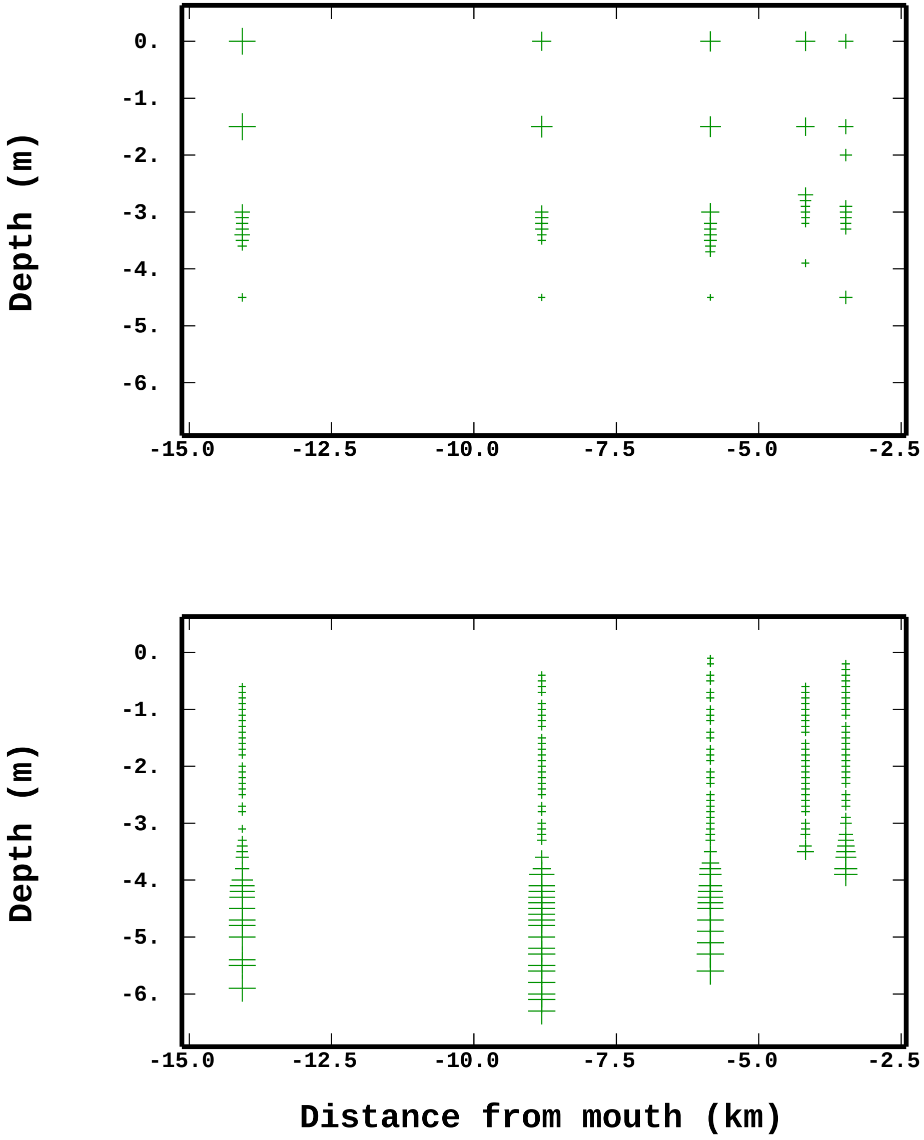
Anisotropic cubic variogram model in both directions (M1, M2).

Abscissa scale: *kilometers* for D1 and *meters* for D2.

Behavior at the origin

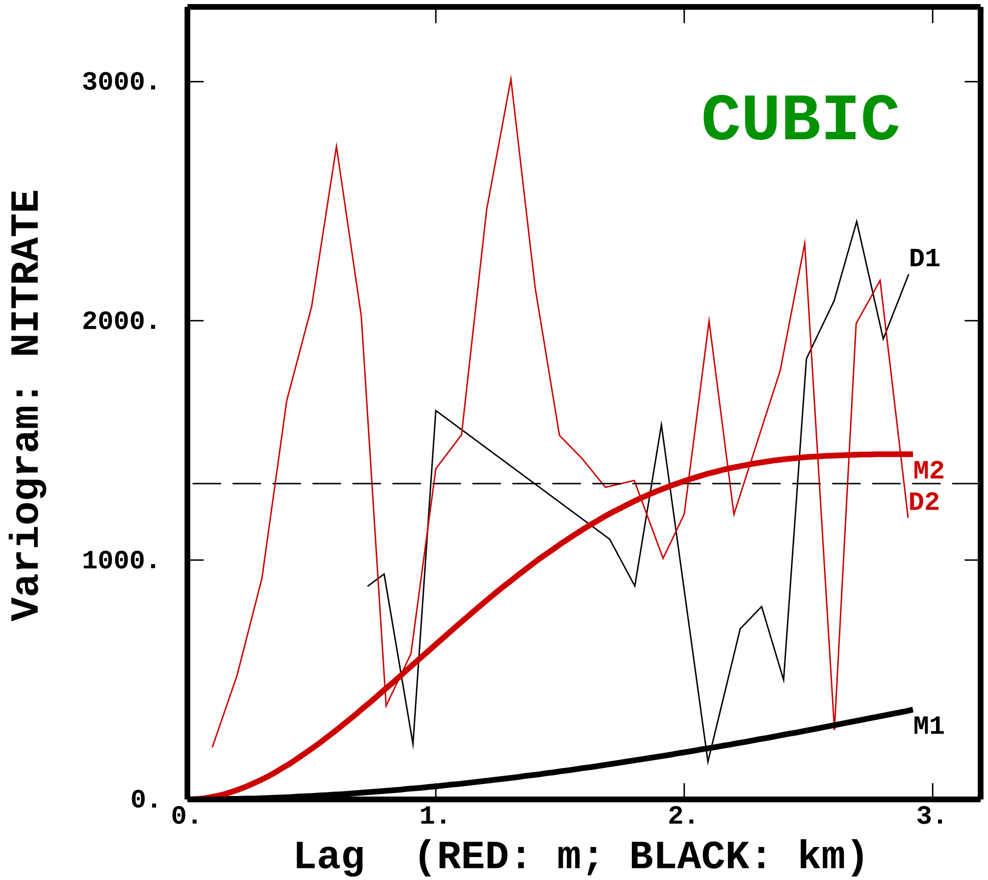
of the variogram

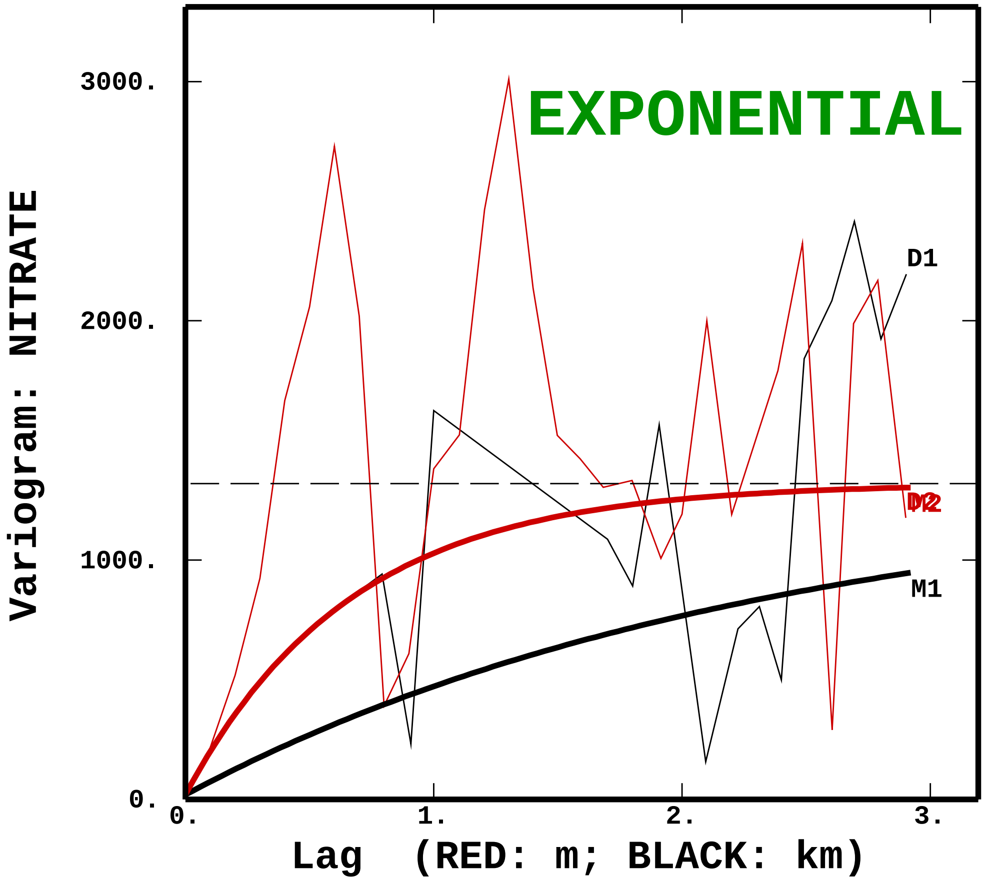
Ebro river: water samples



47 water samples (top) 185 conductivity values (bottom)

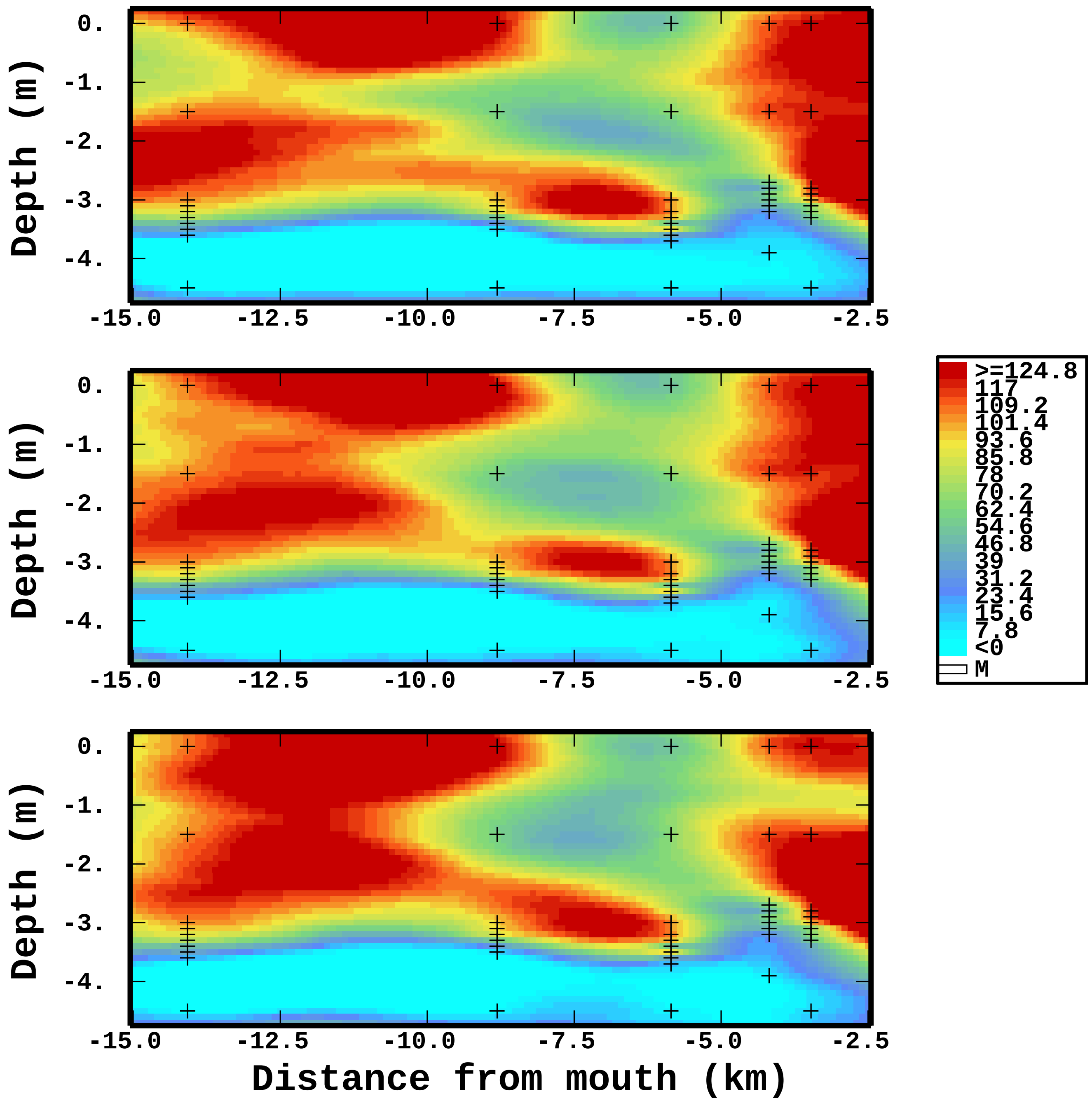
Nitrate variogram: which behavior at origin?



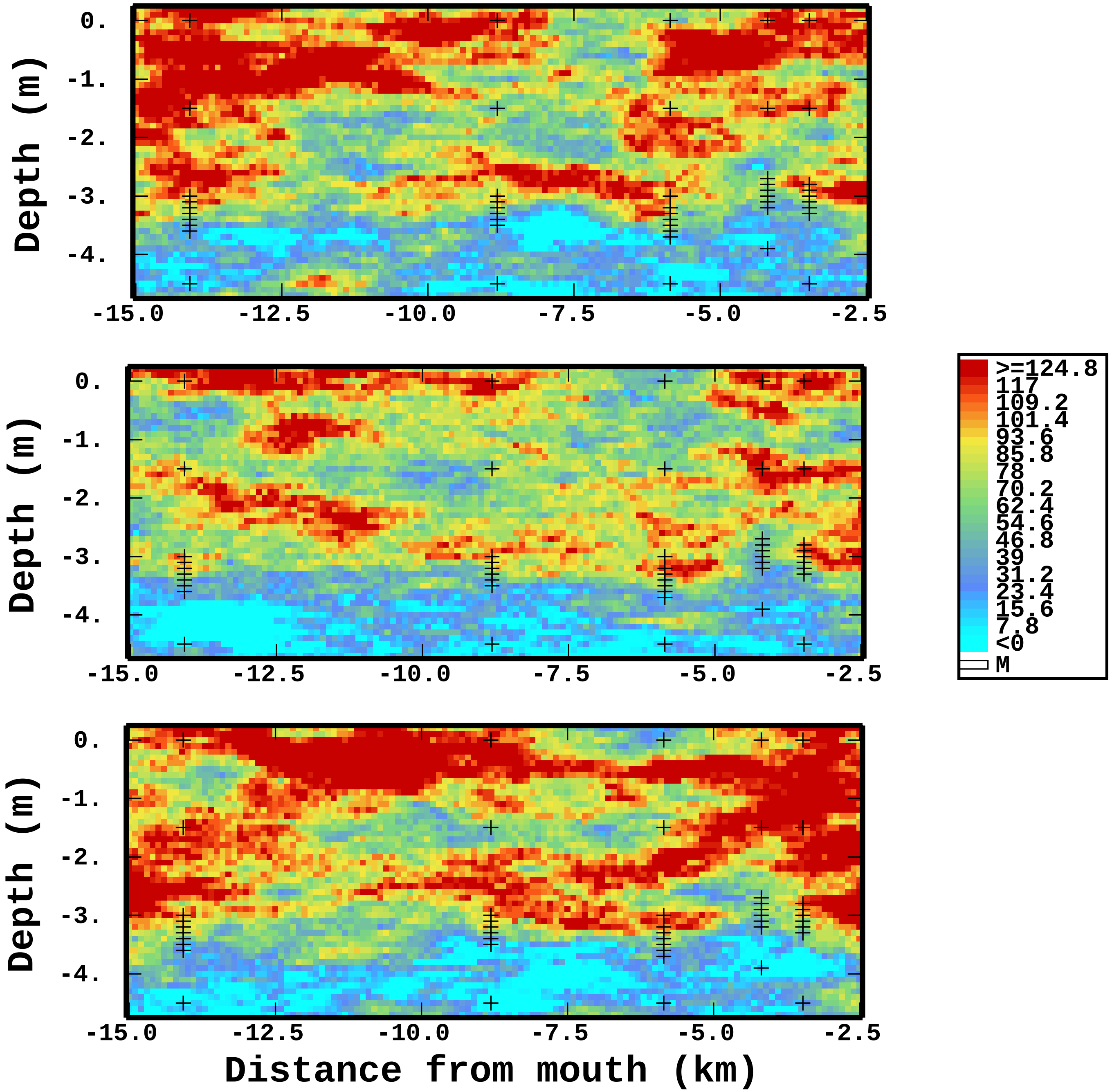


Nitrate experimental variogram with two alternate models.

Cubic variogram: conditional simulations



Exponential model: conditional simulations

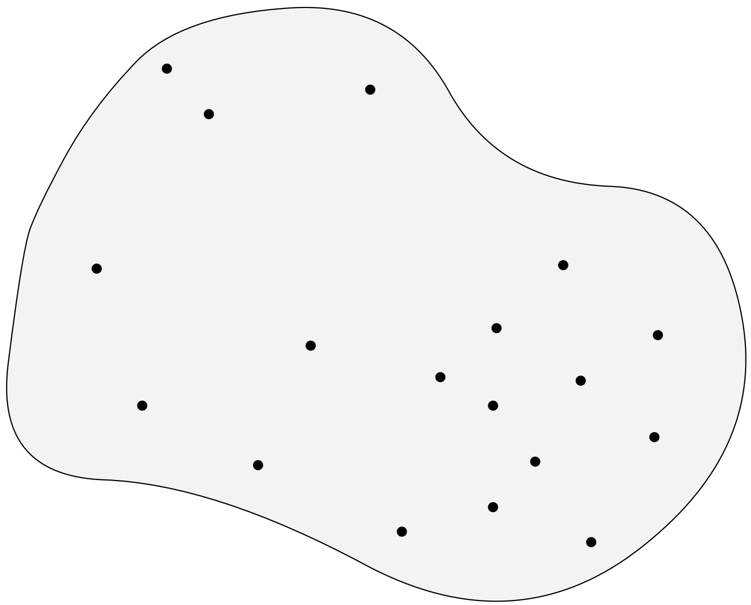


Kriging of the mean

of a random function

Spatially Correlated Data

Sample locations $\mathbf{x}_\alpha$ in a geographical domain:



With spatial correlation we need to consider that:

- sample points have a different number of immediate neighbors,
- distances to neighboring points play a role.

How should samples be weighted in an optimal way?

Estimation of the Mean Value

Using the formula of the arithmetic mean:

$$M^{\star} = \frac{1}{n} \sum_{\alpha=1}^{n} Z(\mathbf{x}_{\alpha})$$

all samples get the same weight: $\dfrac{1}{n}$

We rather need an estimator:

$$M^{\star} = \sum_{\alpha=1}^{n} w_{\alpha} \, Z(\mathbf{x}_{\alpha})$$

with weights w_{α} reflecting the spatial correlation.

Stationary random function

We assume *translation-invariance* of mean and covariance:

$$\forall\,\mathbf{x}\in\mathcal{D}: \quad \mathrm{E}\Big[Z(\mathbf{x})\Big] = m; \qquad \forall\,\mathbf{x}_\alpha,\mathbf{x}_\beta\in\mathcal{D}: \quad C(\mathbf{x}_\alpha,\mathbf{x}_\beta) = C(\mathbf{x}_\alpha - \mathbf{x}_\beta).$$

Unbiased estimation error

The estimation error in our statistical model:

$$\underbrace{M^\star}_{\text{estimated value}} \quad - \quad \underbrace{m}_{\text{true value}}$$

should be zero, on average:

$$\mathrm{E}\Big[\,M^\star - m\,\Big] = 0$$

No bias

No bias is obtained using weights of unit sum:

$$\boxed{\sum_{\alpha=1}^{n} w_\alpha = 1}$$

Consider:

$$\begin{aligned}
\mathrm{E}\Big[M^\star - m\Big] &= \mathrm{E}\Big[\sum_{\alpha=1}^{n} w_\alpha\, Z(\mathbf{x}_\alpha) - m\Big] \\
&= \sum_{\alpha=1}^{n} w_\alpha \underbrace{\mathrm{E}\Big[Z(\mathbf{x}_\alpha)\Big]}_{m} - m \\
&= m \underbrace{\sum_{\alpha=1}^{n} w_\alpha}_{1} - m \quad = \quad 0
\end{aligned}$$

Variance of the estimation error

The variance σ_{E}^2 of the estimation error is:

$$
\begin{aligned}
\operatorname{var}(M^\star - m) &= \mathrm{E}\Big[(M^\star - m)^2\Big] \quad - \quad \underbrace{\mathrm{E}\Big[M^\star - m\Big]^2}_{0} \\
&= \mathrm{E}\Big[\,M^{\star 2} - 2\,m M^\star + m^2\,\Big] \\
&= \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} w_\alpha\, w_\beta\, \mathrm{E}\Big[Z(\mathbf{x}_\alpha)\, Z(\mathbf{x}_\beta)\Big] \\
&\qquad -2\,m \sum_{\alpha=1}^{n} w_\alpha\, \underbrace{\mathrm{E}\Big[Z(\mathbf{x}_\alpha)\Big]}_{m} + m^2
\end{aligned}
$$

$$
\Rightarrow \qquad \boxed{\sigma_{\mathrm{E}}^2 = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} w_\alpha\, w_\beta\, C(\mathbf{x}_\alpha - \mathbf{x}_\beta)}
$$

Minimal estimation variance

We want weights w_α that produce a minimal estimation variance:

$$\boxed{\text{minimum of } \operatorname{var}(M^\star - m) \text{ subject to } \sum_{\alpha=1}^{n} w_\alpha = 1}$$

Method of Lagrange

The objective function φ has $n+1$ parameters:

$$\varphi(w_1, \ldots, w_n, \mu) = \operatorname{var}(M^\star - m) - 2\,\mu \left(\sum_{\alpha=1}^{n} w_\alpha - 1 \right)$$

with μ a Lagrange multiplier. Setting partial derivatives to zero:

$$\forall \alpha : \frac{\partial \varphi(w_1, \ldots, w_n, \mu)}{\partial w_\alpha} = 0, \qquad \frac{\partial \varphi(w_1, \ldots, w_n, \mu)}{\partial \mu} = 0$$

Kriging equations

The method of Lagrange yields the equations for the optimal weights w_α^{KM} of the kriging of the mean:

$$\left\{\begin{array}{rcll} \displaystyle\sum_{\beta=1}^{n} w_\beta^{\mathrm{KM}}\, C(\mathbf{x}_\alpha - \mathbf{x}_\beta) - \mu_{\mathrm{KM}} & = & 0 & \quad \text{for} \quad \alpha = 1, \ldots, n \\ \displaystyle\sum_{\beta=1}^{n} w_\beta^{\mathrm{KM}} & = & 1 & \end{array}\right.$$

The variance at the minimum:

$$\sigma_{\mathrm{KM}}^2 = \mu_{\mathrm{KM}}$$

is equal to the Lagrange multiplier.

Case of no autocorrelation

When the covariance model is a pure *nugget-effect*:

$$C(\mathbf{x}_\alpha - \mathbf{x}_\beta) = \begin{cases} \sigma^2 & \text{if } \mathbf{x}_\alpha = \mathbf{x}_\beta \\ 0 & \text{if } \mathbf{x}_\alpha \neq \mathbf{x}_\beta \end{cases}$$

the kriging of the mean system simplifies to:

$$\begin{cases} w_\alpha^{\mathrm{KM}}\,\sigma^2 & = & \mu_{\mathrm{KM}} & \text{for} \quad \alpha = 1, \ldots, n \\ \displaystyle\sum_{\beta=1}^{n} w_\beta^{\mathrm{KM}} & = & 1 & \end{cases}$$

The solution weights are all equal: $\displaystyle w_\alpha^{\mathrm{KM}} = \frac{1}{n}$

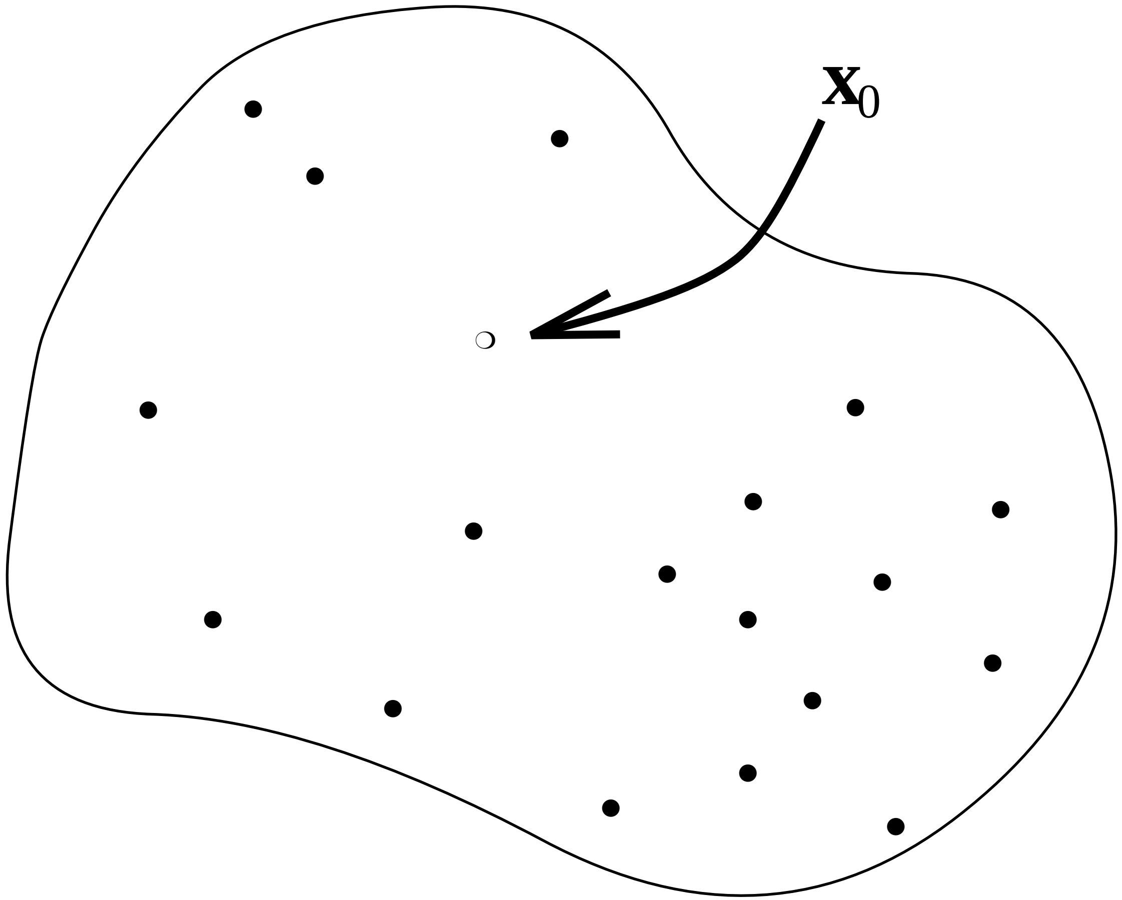
$\Rightarrow \quad \displaystyle M^\star = \frac{1}{n} \sum_{\alpha=1}^{n} Z(\mathbf{x}_\alpha)$ the arithmetic mean! $\quad \mu_{\mathrm{KM}} = \boxed{\sigma^2_{\mathrm{KM}} = \frac{1}{n}\,\sigma^2}$

Ordinary Kriging

at a point in the domain

Estimation at a Point

Sample locations $\mathbf{x}_\alpha$ (dots)
in a domain $\mathcal{D}$:



We wish to estimate a value $Z^\star$ at a point $\mathbf{x}_0$.

Ordinary kriging

The estimate $Z^\star$ is a weighted average of data values $Z(\mathbf{x}_\alpha)$:

$$Z^\star(\mathbf{x}_0) \quad = \quad \sum_{\alpha=1}^{n} w_\alpha \, Z(\mathbf{x}_\alpha) \qquad \text{with} \; \sum_{\alpha=1}^{n} w_\alpha = 1$$

The weights w_α^{OK} of the Best Linear Unbiased Estimator (BLUE) are solution of the system:

$$\begin{cases} \displaystyle\sum_{\beta=1}^{n} w_\beta^{\mathrm{OK}} \, \gamma(\mathbf{x}_\alpha - \mathbf{x}_\beta) + \mu_{\mathrm{OK}} & = \quad \boxed{\gamma(\mathbf{x}_\alpha - \mathbf{x}_0)} \qquad \forall \alpha \\ \displaystyle\sum_{\beta=1}^{n} w_\beta^{\mathrm{OK}} & = \quad 1 \end{cases}$$

Minimal variance: $\sigma_{\mathrm{OK}}^2 = \mu_{\mathrm{OK}} + \boxed{\displaystyle\sum_{\alpha=1}^{n} w_\alpha^{\mathrm{OK}} \, \gamma(\mathbf{x}_\alpha - \mathbf{x}_0)}$

Cross-validation

leaving one out and reestimating it

Cross-validation

Comment: the sound way to cross-validate is to leave out *half* of the data locations and to re-estimate them from the other *half*: this requires many data! For that reason it is often done in the following way (implemented in sotware packages)...

A data value $Z(\mathbf{x}_\alpha)$ is left out and a value $Z^\star(\mathbf{x}_{[\alpha]})$ is estimated at location $\mathbf{x}_\alpha$ by ordinary kriging.

The notation $[\alpha]$ means that the sample at $\mathbf{x}_\alpha$ has not been used for estimating $Z^\star(\mathbf{x}_{[\alpha]})$.

The difference between the data value and the estimated value:

$$Z(\mathbf{x}_\alpha) - Z^\star(\mathbf{x}_{[\alpha]})$$

gives an indication of how well the data value fits into the neigborhood of the surrounding data values.

Average cross-Validation error

If the average of the cross-validation errors is not far from zero:

$$\frac{1}{n}\sum_{\alpha=1}^{n}\Big(Z(\mathbf{x}_\alpha) - Z^\star(\mathbf{x}_{[\alpha]})\Big) \cong 0$$

then there is no systematic bias.

A negative (positive) average error represents
systematic overestimation (underestimation).

Standardized cross-validation error

The kriging standard deviation σ_{K} represents the error predicted by the model.

Dividing the cross-validation error by σ_{K} allows to compare the magnitudes of both errors:

$$\frac{Z(\mathbf{x}_\alpha) - Z^\star(\mathbf{x}_{[\alpha]})}{\sigma_{\mathrm{K}\alpha}}$$

Average squared Standardized Errors

If the average of the squared standardized cross-validation errors is not far from one:

$$\frac{1}{n} \sum_{\alpha=1}^{n} \frac{\Big(Z(\mathbf{x}_\alpha) - Z^\star(\mathbf{x}_{[\alpha]}) \Big)^2}{\sigma^2_{\mathrm{K}\alpha}} \cong 1$$

then the actual estimation error is equal on average to the error predicted by the model.

This quantity gives an idea of the adequacy of the model and of its parameters.

Mapping with kriging

on a regular grid
with irregularly spaced data

Kriging for interpolation

Kriging is an estimation method.

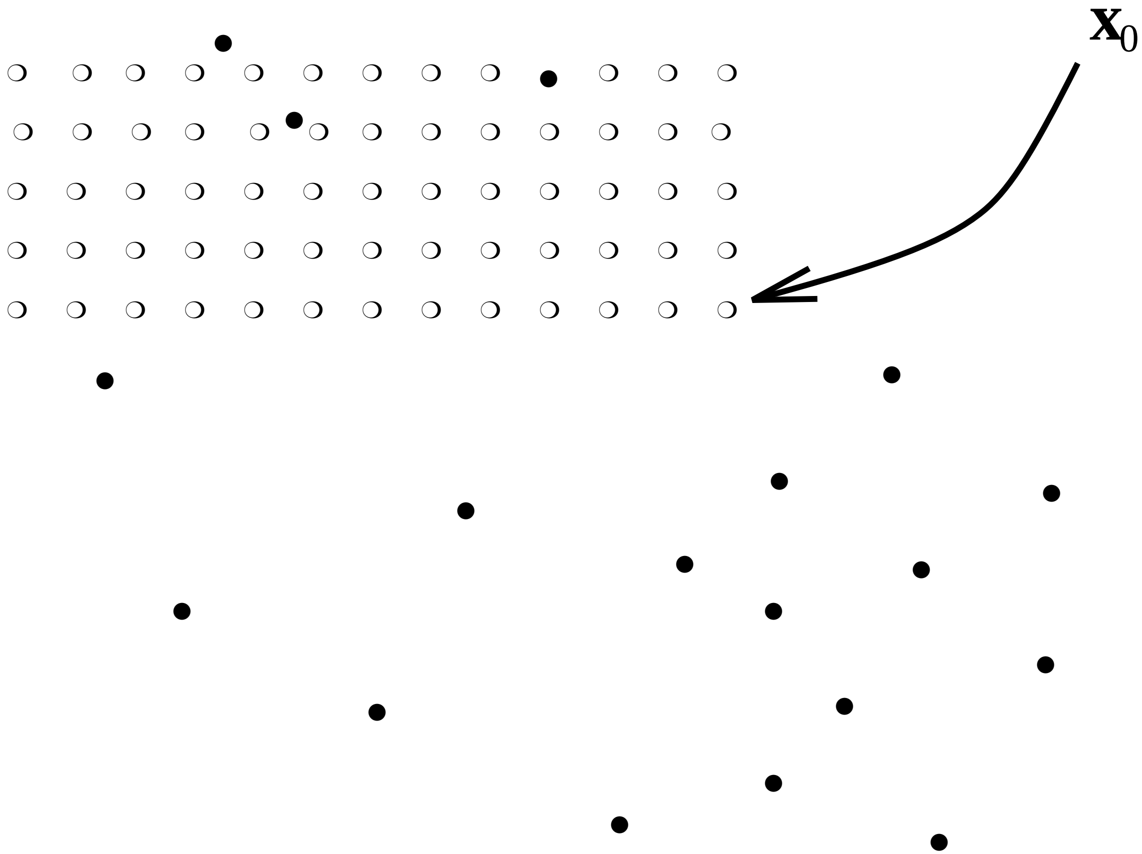
It is not the quickest method to make an interpolation on a regular grid for generating a map.

Its advantages are:

- Kriging integrates the knowledge gained from analysing the spatial structure: the variogram.
- Kriging interpolates exactly: when a sample value is available at the location-of-interest, the kriging solution is equal to that value.
- Kriging provides an indication of the estimation error: the kriging variance.

Generating a map

A regular grid is defined by the computer and at each node of this grid a value is kriged.

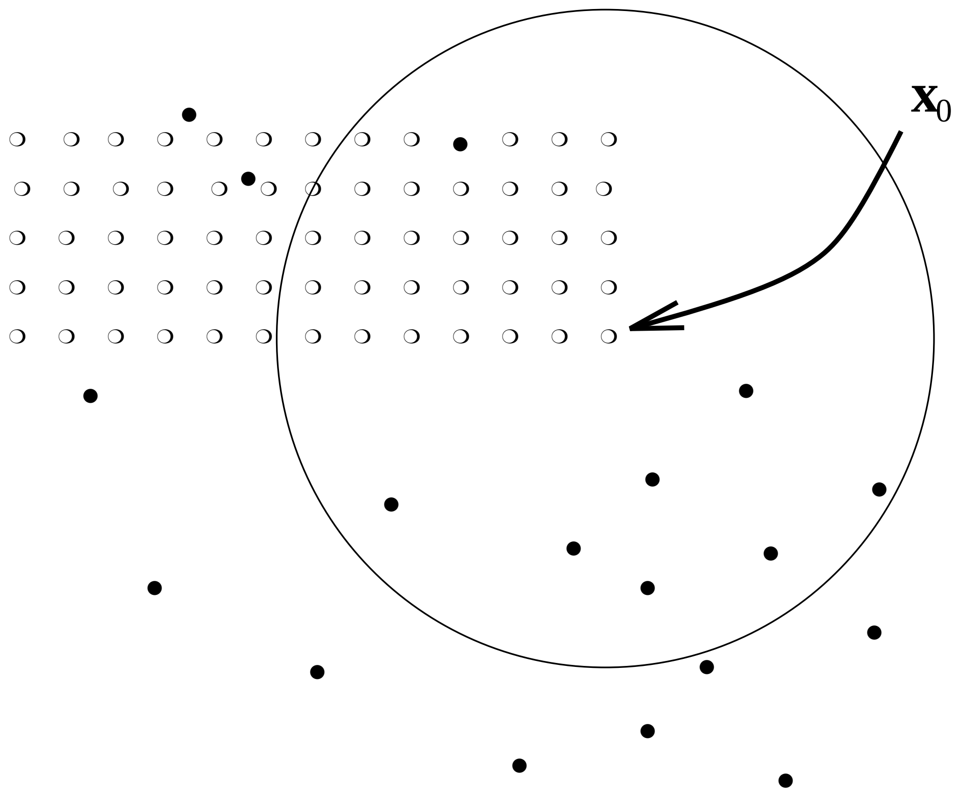


Afterwards a graphical representation of this grid is performed, as a raster of colour squares, as an isoline map, as a bloc diagram. . .

Moving Neighborhood

If all data are used: this is called a *unique* neighborhood.

Using a subset of close data points: a *moving* neighborhood.



To choose the size of the moving neighborhood, the range of the variogram can give an indication.

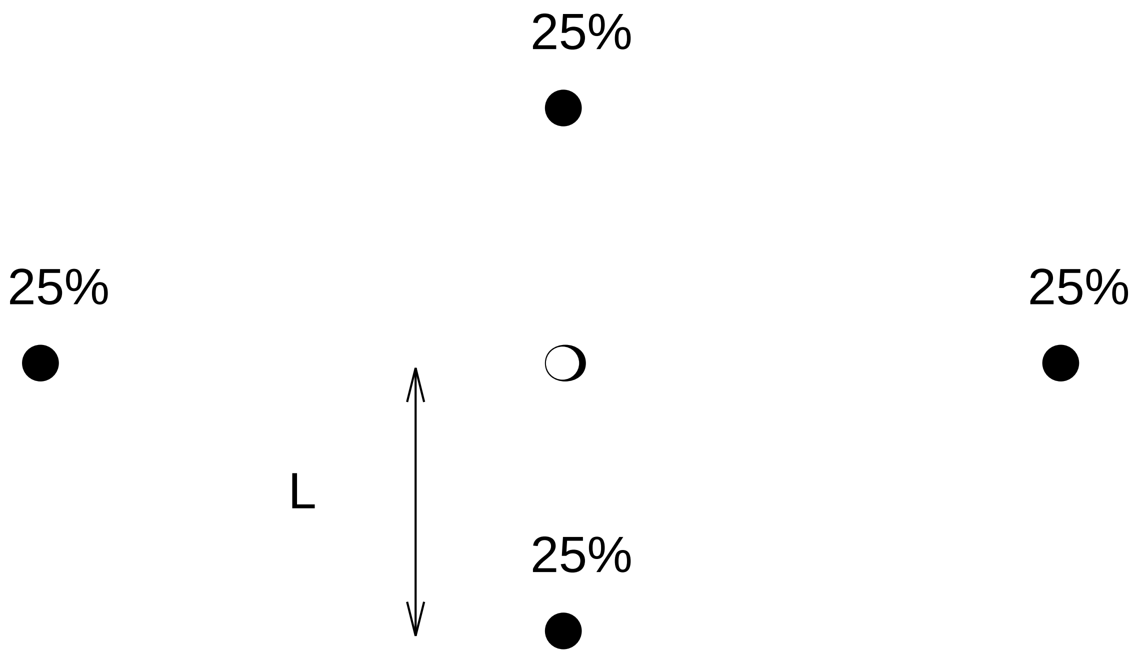
Kriging weights

The shape of the kriging weights

Kriging weights

Nugget-effect model

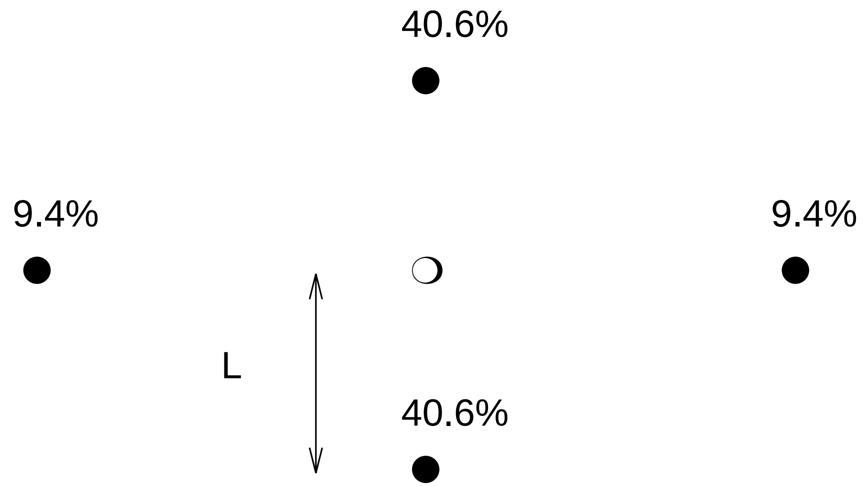
$$\boxed{\sigma^2_{\mathrm{OK}} = 1.25}$$



Isotropic variogram

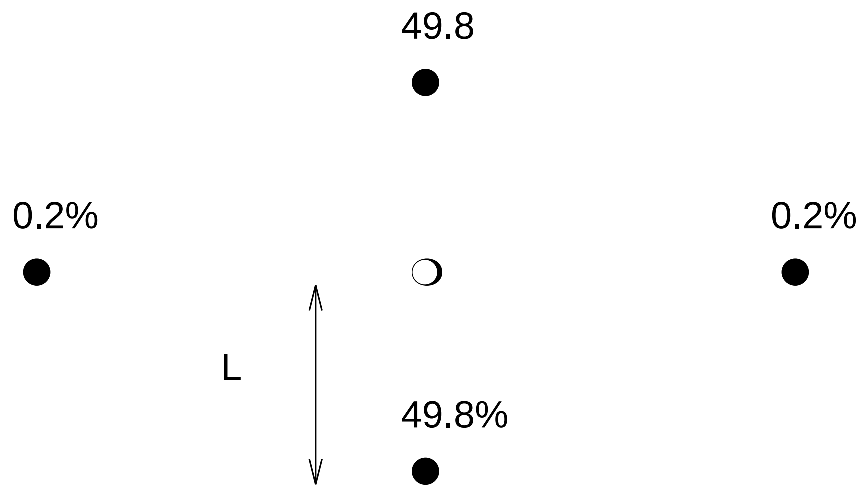
Spherical model with range a/L = 2

$$\boxed{\sigma^2_{\mathrm{OK}} = .84}$$



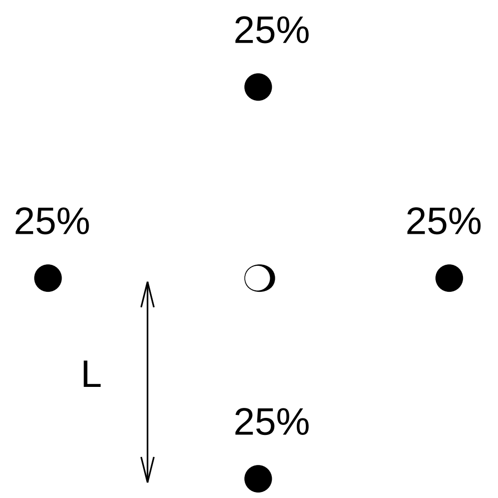
Gaussian model with range a/L = 1.5

$$\boxed{\sigma^2_{\mathrm{OK}} = .30}$$

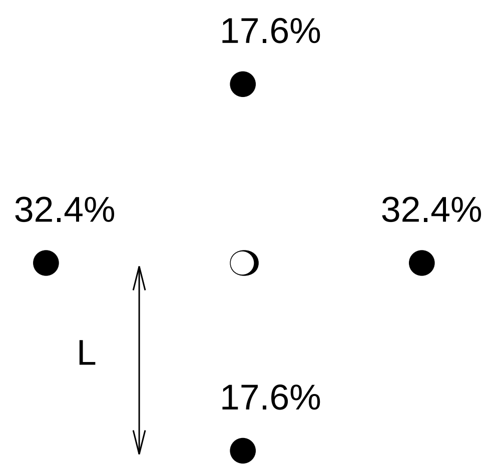


Geometric anisotropy

Spherical with isotropic range

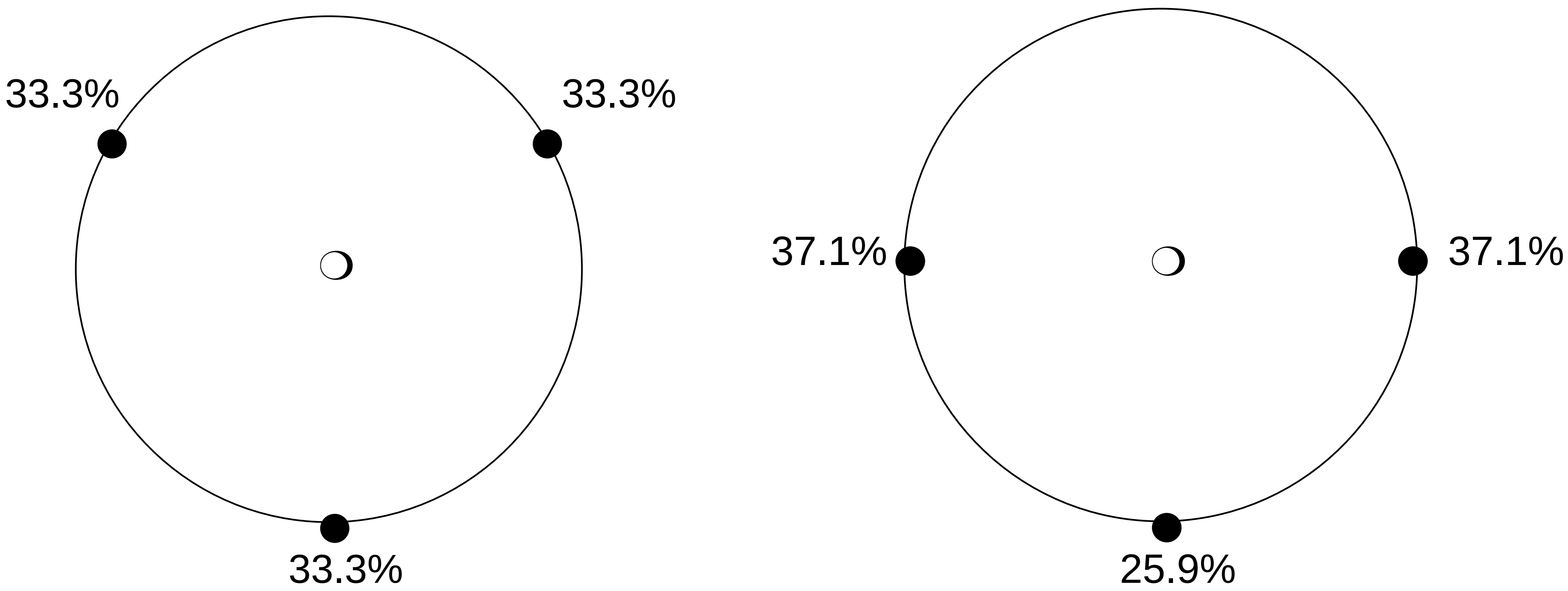


Spherical with horizontal a/L = 1.5 and vertical a/L = .75



Relative position of samples

$\boxed{\sigma^2_{\mathrm{OK}} = .45}$ $\boxed{\sigma^2_{\mathrm{OK}} = .48}$



The left configuration gives a more reliable estimate.

The screen effect

Spherical model with range a/L = 2

$$\boxed{\sigma^2_{\mathrm{OK}} = 1.14}$$

$$\boxed{\sigma^2_{\mathrm{OK}} = 0.87}$$

Adding the sample C screens off the sample B.

Nested variogram

and corresponding linear model of the random function

Nested Variogram Model

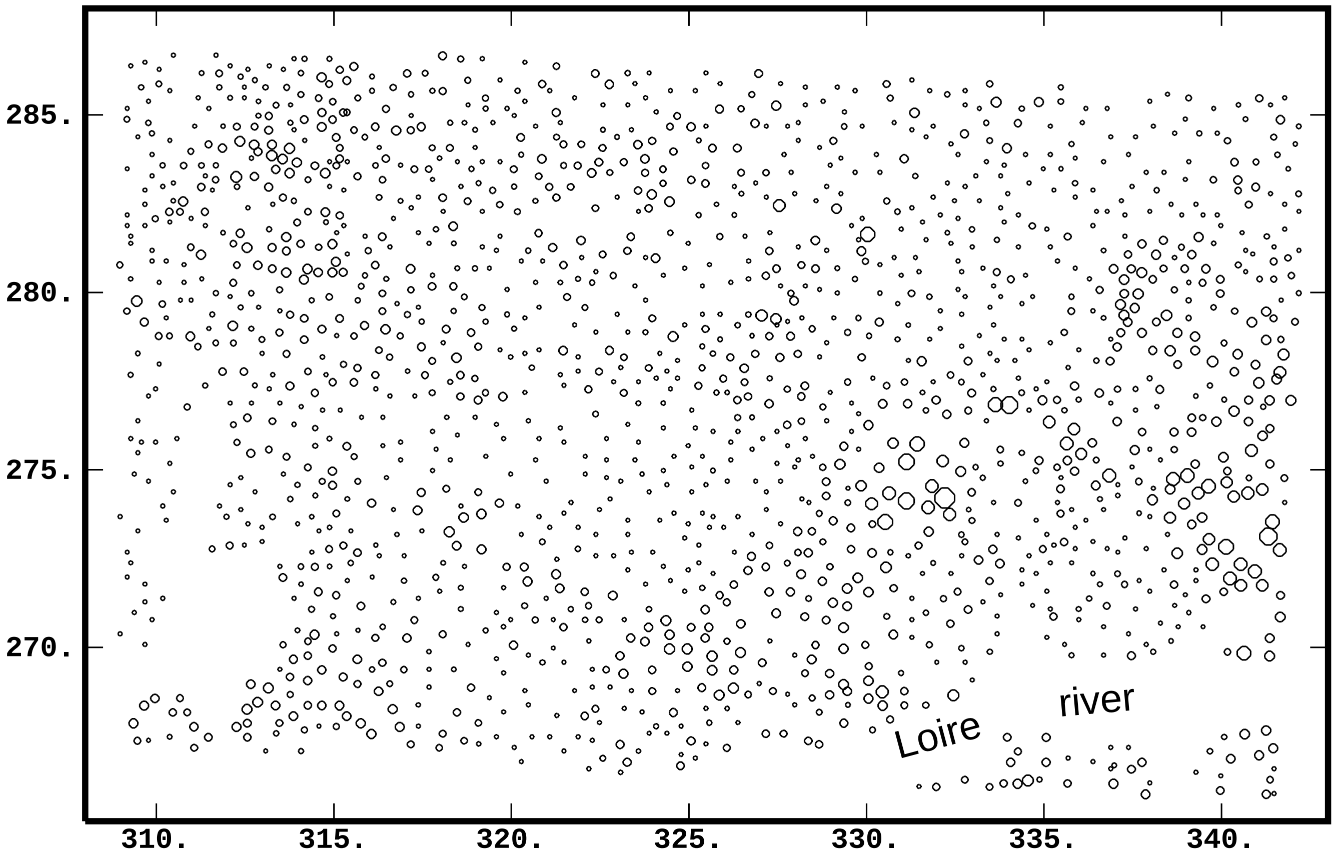
A nested variogram $\gamma(\mathbf{h})$ is composed of
a sum of elementary variograms $\gamma_u(\mathbf{h})$
with $u = 0, \ldots, S$:

$$\gamma(\mathbf{h}) = \gamma_0(\mathbf{h}) + \ldots + \gamma_S(\mathbf{h}) = \sum_{u=0}^{S} \gamma_u(\mathbf{h})$$

Each variogram $\gamma_u(\mathbf{h})$ is build up with a normed variogram $g_u(\mathbf{h})$
multiplied with a coefficient b_u (sill, slope):

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

Example: Arsenic in soil (Loire, France)

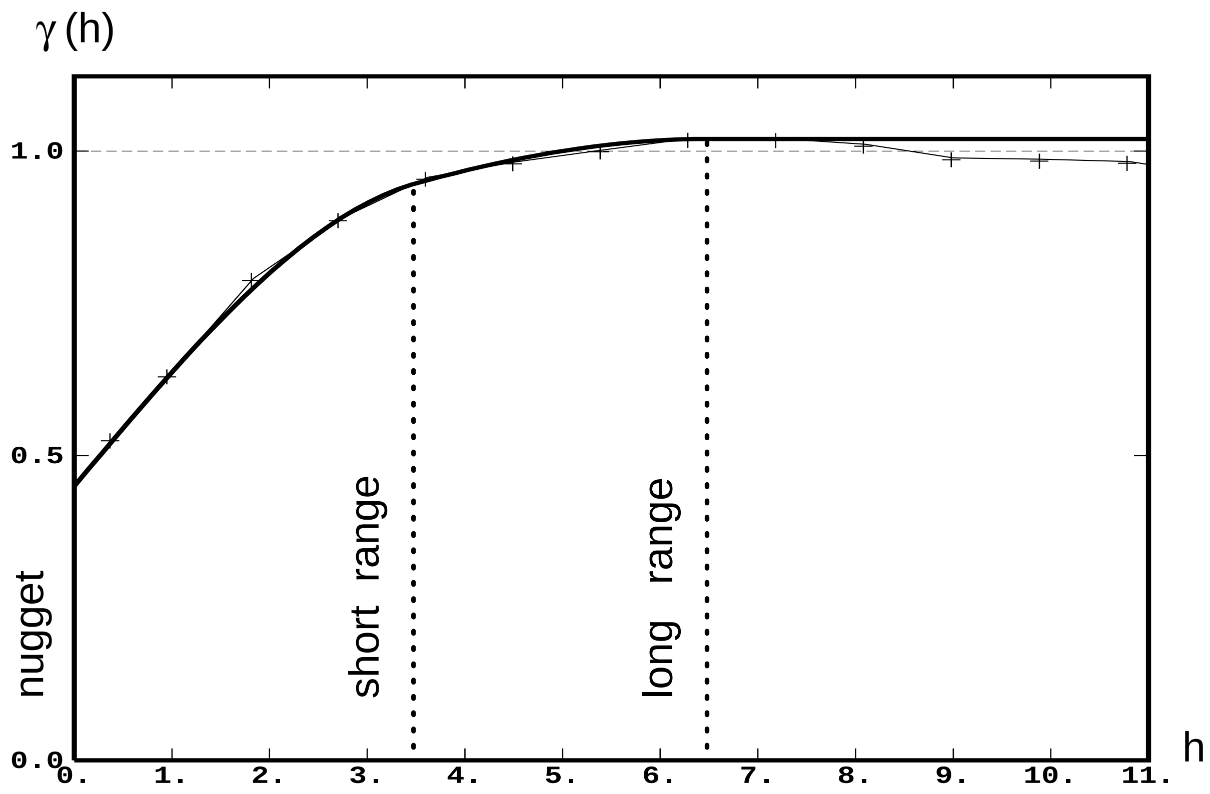


35×25 km^2 region. Dots are proportional to sample value.

Example: Nested Variogram Model

A nugget-effect (nug) and two spherical (sph) structures:

$$\gamma(\mathbf{h}) = b_0 \,\mathsf{nug}(\mathbf{h}) + b_1 \,\mathsf{sph}(\mathbf{h}, a_1) + b_2 \,\mathsf{sph}(\mathbf{h}, a_2)$$



Nested Covariance Function

$$C(\mathbf{h}) = \sum_{u=0}^{S} C_u(\mathbf{h}) = \sum_{u=0}^{S} b_u \, \rho_u(\mathbf{h})$$

where $\rho_u(\mathbf{h})$ are correlation functions.

The $\rho_u(\mathbf{h})$ characterize the spatial correlation at different scales of index u.

The coefficents b_u represent a decomposition of the total variance σ^2 into variances at different spatial scales:

$$C(0) = \sigma^2 = \sum_{u=0}^{S} b_u$$

Regionalization Model

$Z(\mathbf{x})$ built up with uncorrelated components $Y_u(\mathbf{x})$ of zero mean, with covariance functions $C_u(\mathbf{h})$.

Example:

$$Z(\mathbf{x}) = Y_1(\mathbf{x}) + Y_2(\mathbf{x}) + m \qquad \text{with} \quad Y_1 \perp Y_2$$

The covariance function of $Z(\mathbf{x})$ is nested:

$$C(\mathbf{h}) = C_1(\mathbf{h}) + C_2(\mathbf{h})$$

Linear Model with $S+1$ components

$$Z(\mathbf{x}) \;=\; \sum_{u=0}^{S} Y_u(\mathbf{x}) + m$$

with $Y_u \perp Y_v$ for $u \neq v$

Corresponding nested covariance model:

$$C(\mathbf{h}) = \sum_{u=0}^{S} C_u(\mathbf{h}) = \sum_{u=0}^{S} b_u\, \rho_u(\mathbf{h})$$

- *Can components Y_u be extracted from samples $Z(\mathbf{x}_\alpha)$?*

Kriging Spatial Components

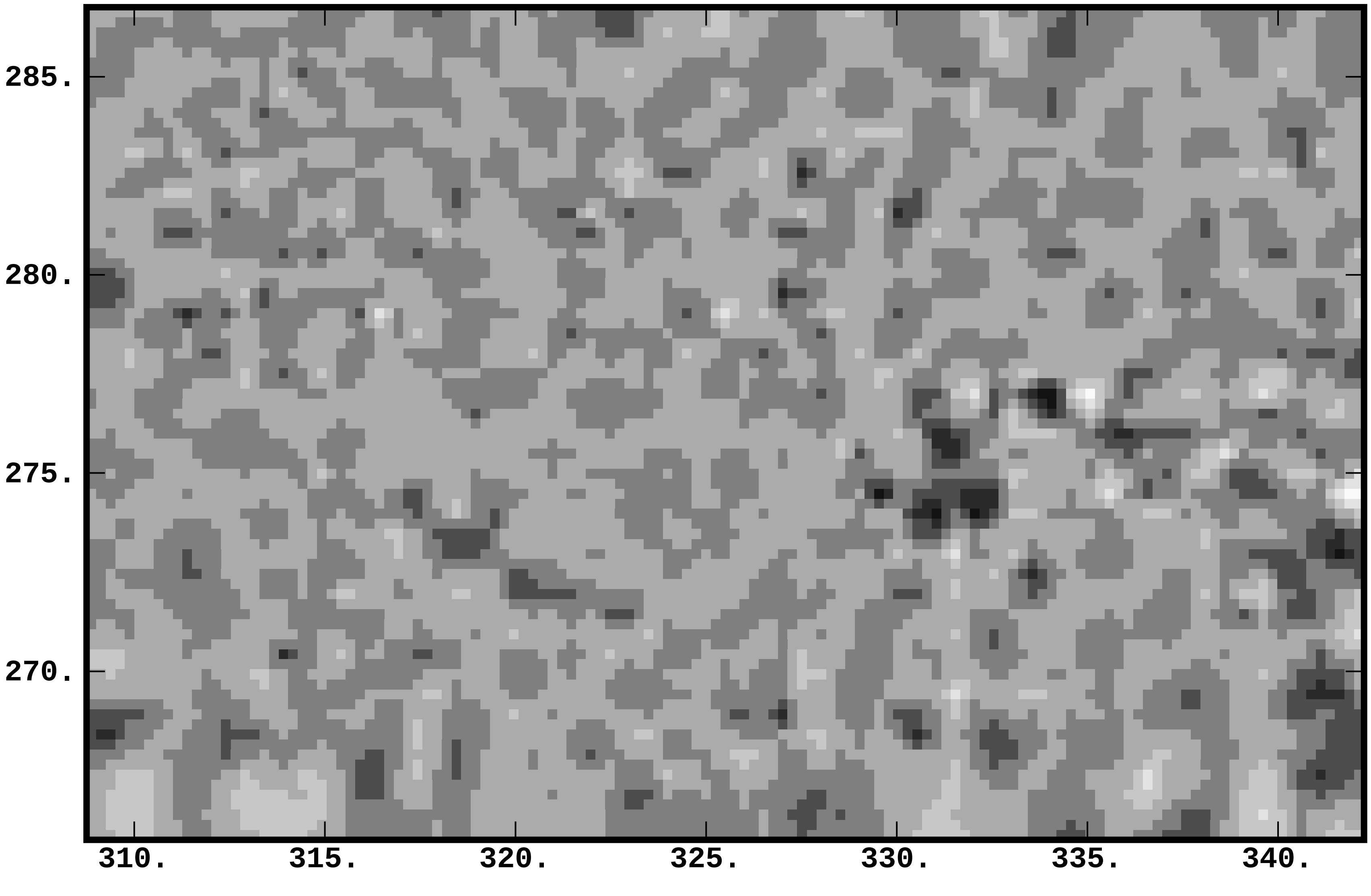
A component $Y_1(\mathbf{x})$ at $\mathbf{x}_0$ is estimated from n data:

$$Y_1^\star(\mathbf{x}_0) = \sum_{\alpha=1}^{n} w_\alpha \, Z(\mathbf{x}_\alpha)$$

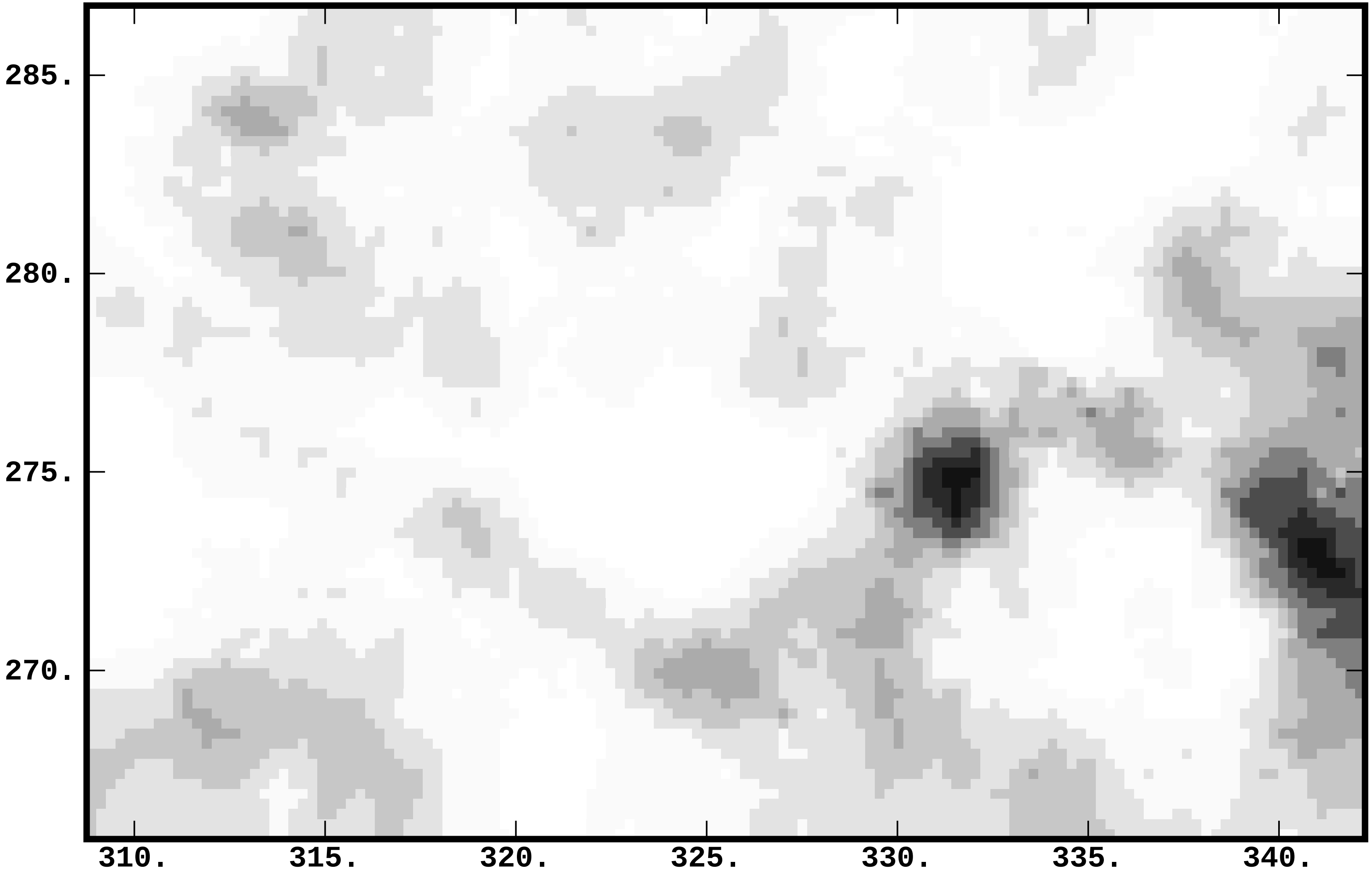
- "No bias" with $\displaystyle\sum_{\alpha=1}^{n} w_\alpha = 0$: *this filters the mean* m
- Minimizing the "estimation variance":

$$\begin{cases} \displaystyle\sum_{\beta=1}^{n} w_\beta^1 \, C(\mathbf{x}_\alpha - \mathbf{x}_\beta) - \mu_1 = C_1(\mathbf{x}_\alpha - \mathbf{x}_0) & \text{for } \alpha = 1, \ldots, n \\ \displaystyle\sum_{\beta=1}^{n} w_\beta = 0 \end{cases}$$

Example: Short-range Component of As



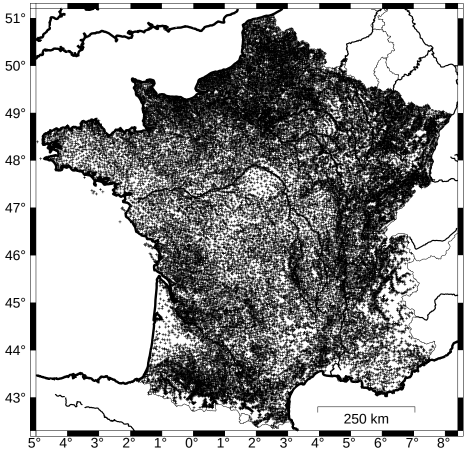
Example: Long-range Component of As



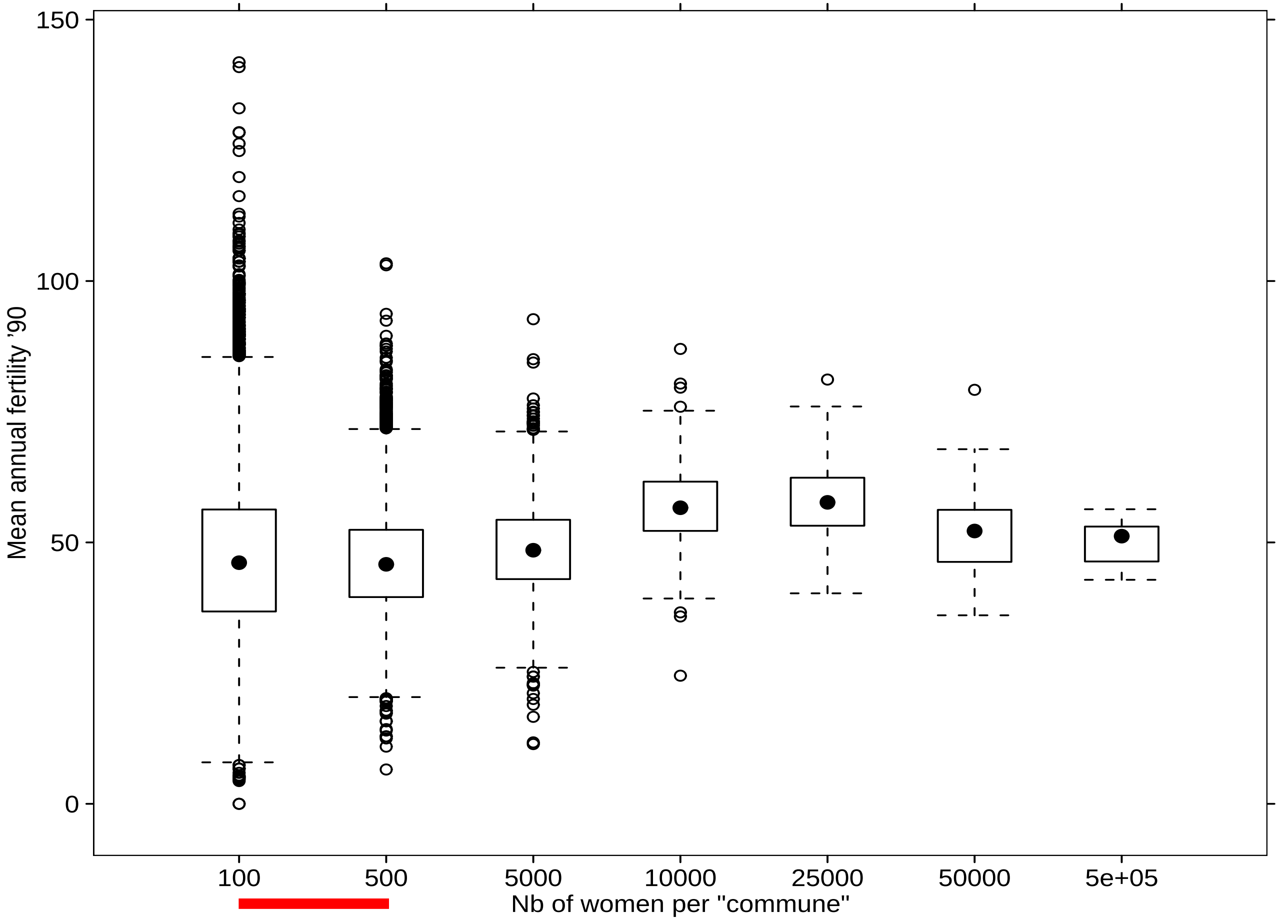
Demographic application

fertility data

Demographic application: fertility 1990



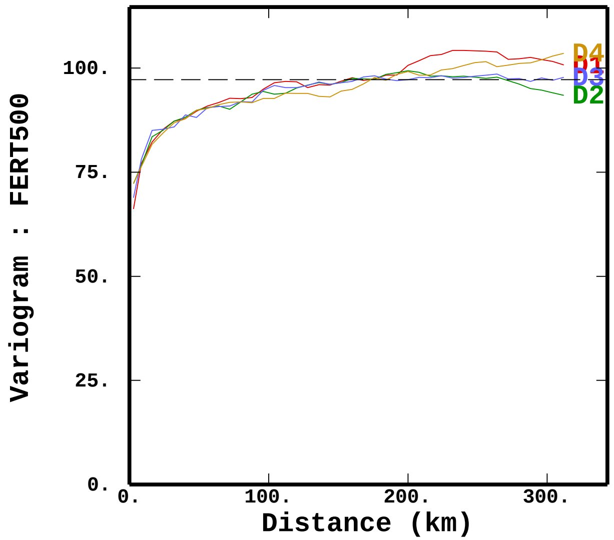
Communes de France

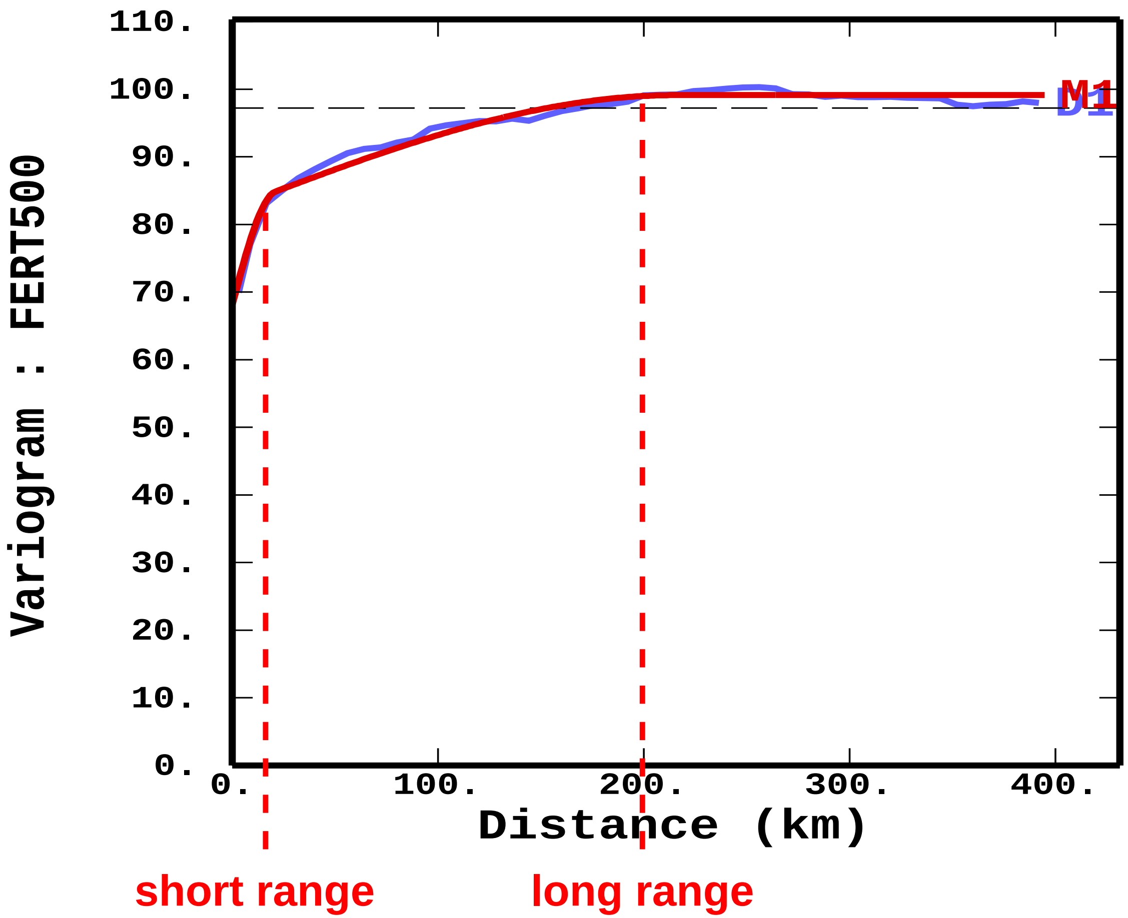


FERT500 class

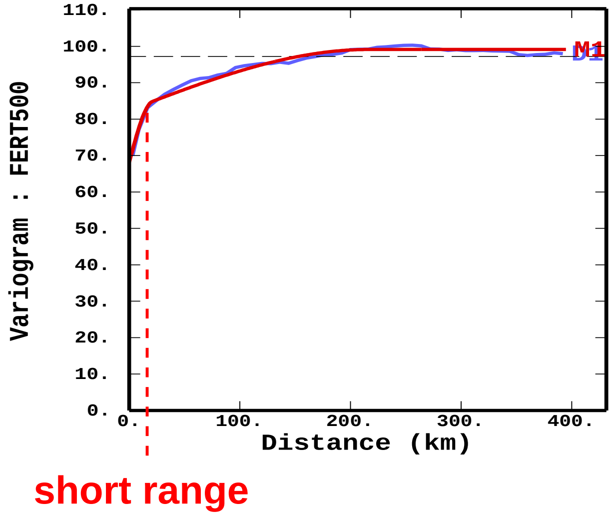
Data provided by INSEE (www.insee.fr)

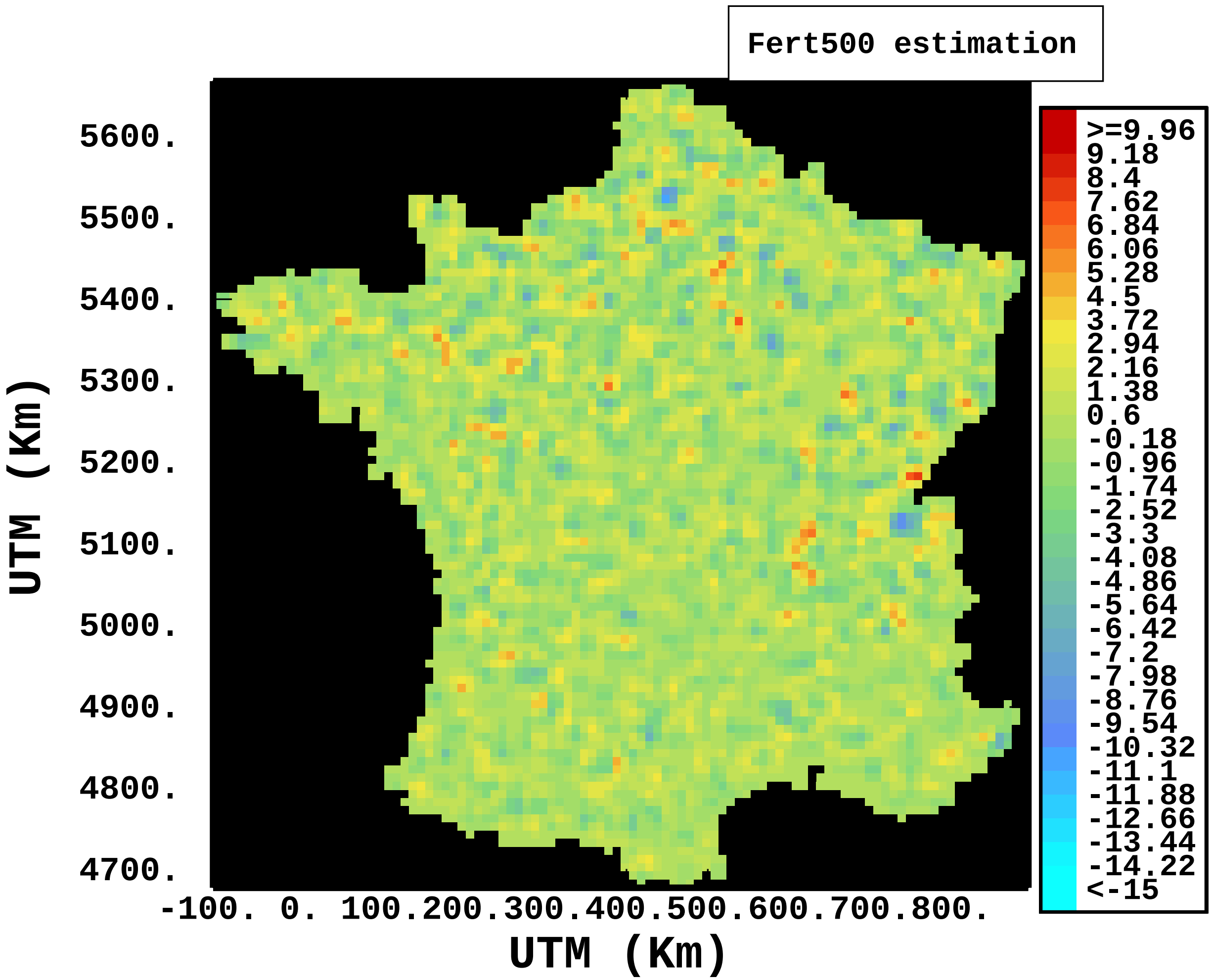
Variograms: class 100-500 women / commune



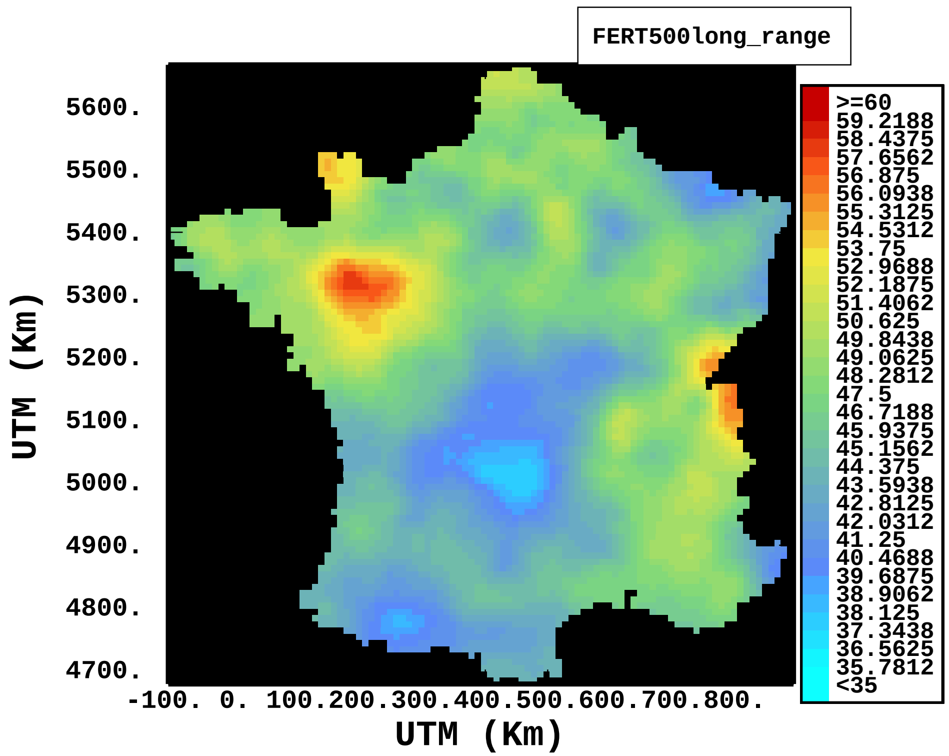


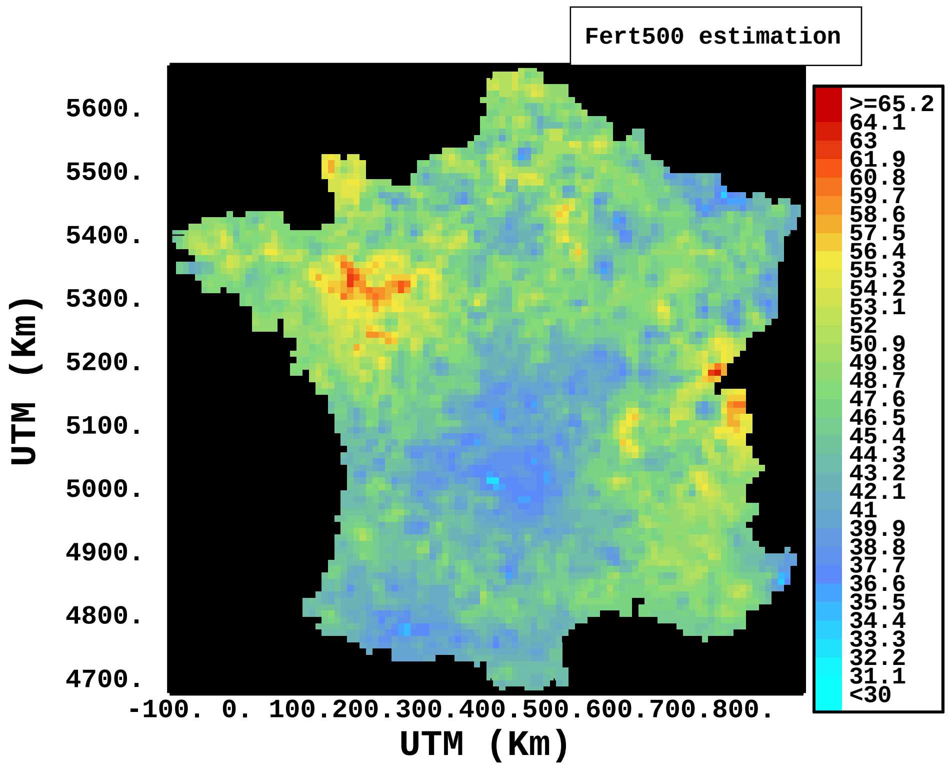
Kriging: short range effect only





Kriging





Long range effect

Short + long range

Kriging with external drift

Drift

Translation-invariant drift: polynomials, trigonometric functions

External Drift: an auxiliary variable known everywhere

Conditions

1. Auxiliary variable $s(\mathbf{x})$ known everywhere in the domain $\mathcal{D}$.

2. The relation to the variable of interest is linear:

$$\mathrm{E}\Big[\,Z(\mathbf{x})\,\Big] = b_0 + b_1\, s(\mathbf{x})$$

External Drift method

$$Z^{\star}(\mathbf{x}_0) \;=\; \sum_{\alpha=1}^{n} w_\alpha \, Z(\mathbf{x}_\alpha)$$

$$\Rightarrow \qquad \mathrm{E}\Big[\, Z^{\star}(\mathbf{x}_0) \Big] \;=\; \sum_{\alpha=1}^{n} w_\alpha \Big(b_0 + b_1 \, s(\mathbf{x}_\alpha) \Big)$$

Constraint: no bias

The constraint

$$\boxed{\sum_{\alpha=1}^{n} w_\alpha = 1}$$

has the effect that the coefficients b_0 and b_1 are filtered out:

$$\mathrm{E}\Big[\,Z^\star(\mathbf{x}_0)\Big] \;=\; \mathrm{E}\Big[\,Z(\mathbf{x}_0)\Big]$$

$$\Longrightarrow \qquad b_0 + b_1 \sum_{\alpha=1}^{n} w_\alpha\, s(\mathbf{x}_\alpha) \;=\; b_0 + b_1\, s(\mathbf{x}_0)$$

$$\Longrightarrow \qquad \sum_{\alpha=1}^{n} w_\alpha\, s(\mathbf{x}_\alpha) \;=\; s(\mathbf{x}_0)$$

Interpolation of external drift

This second constraint:

$$\boxed{\sum_{\alpha=1}^{n} w_\alpha \, s(\mathbf{x}_\alpha) = s(\mathbf{x}_0)}$$

generates weights w_α which interpolate exactly $s(\mathbf{x})$.

Kriging System with linear and external drift

$$
\left\{
\begin{array}{ll}
\displaystyle\sum_{\beta=1}^{n} w_\beta\, C(\mathbf{x}_\alpha - \mathbf{x}_\beta) + \mu_0 + \mu_1\, x_\alpha^1 + \mu_2\, x_\alpha^2 + \mu_3\, s(\mathbf{x}_\alpha) \;=\; C(\mathbf{x}_\alpha - \mathbf{x}_0), \;\forall\alpha & \\
\displaystyle\sum_{\beta=1}^{n} w_\beta = 1 & \\
\displaystyle\sum_{\beta=1}^{n} w_\beta\, x_\beta^1 = x_0^1 & \text{(longitude)} \\
\displaystyle\sum_{\beta=1}^{n} w_\beta\, x_\beta^2 = x_0^2 & \text{(latitude)} \\
\displaystyle\sum_{\beta=1}^{n} w_\beta\, s(\mathbf{x}_\beta) = s(\mathbf{x}_0) & \text{(external drift)}
\end{array}
\right.
$$

Kriging temperature

with elevation as external drift

Temperature Data

Agriculture: temperature conditions the growth of plants

Region: Scotland (without the Shetland and Orkney Islands)

Data: average January temperatures (1961-1980)

Stations: 146 sites, all below 400 m altitude

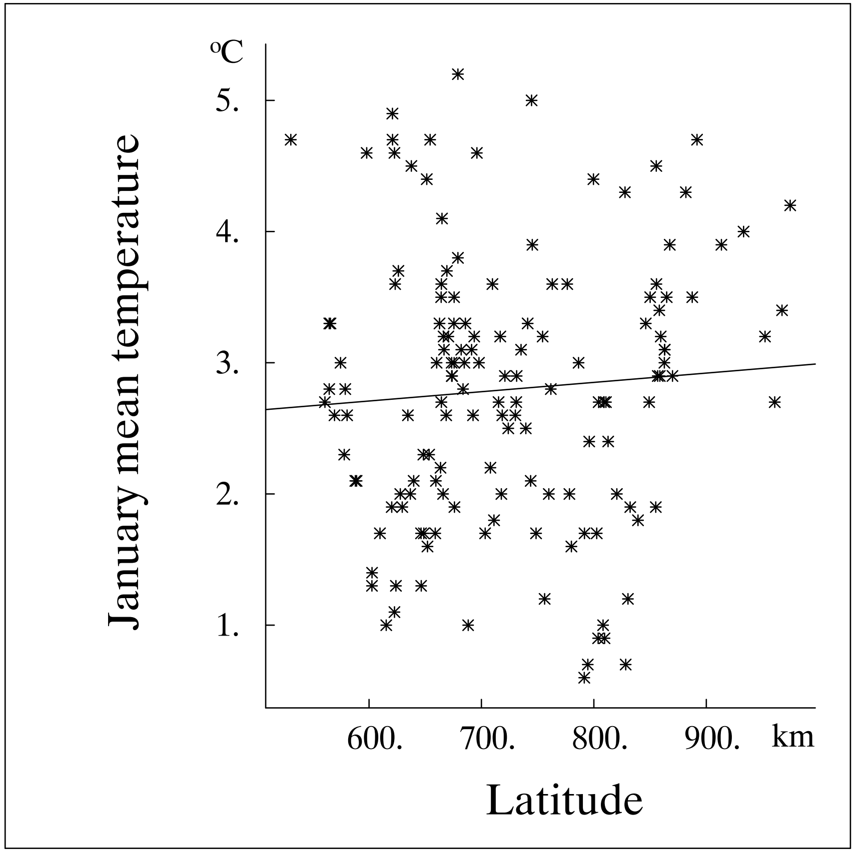
Highest point: Ben Nevis, 1344 m

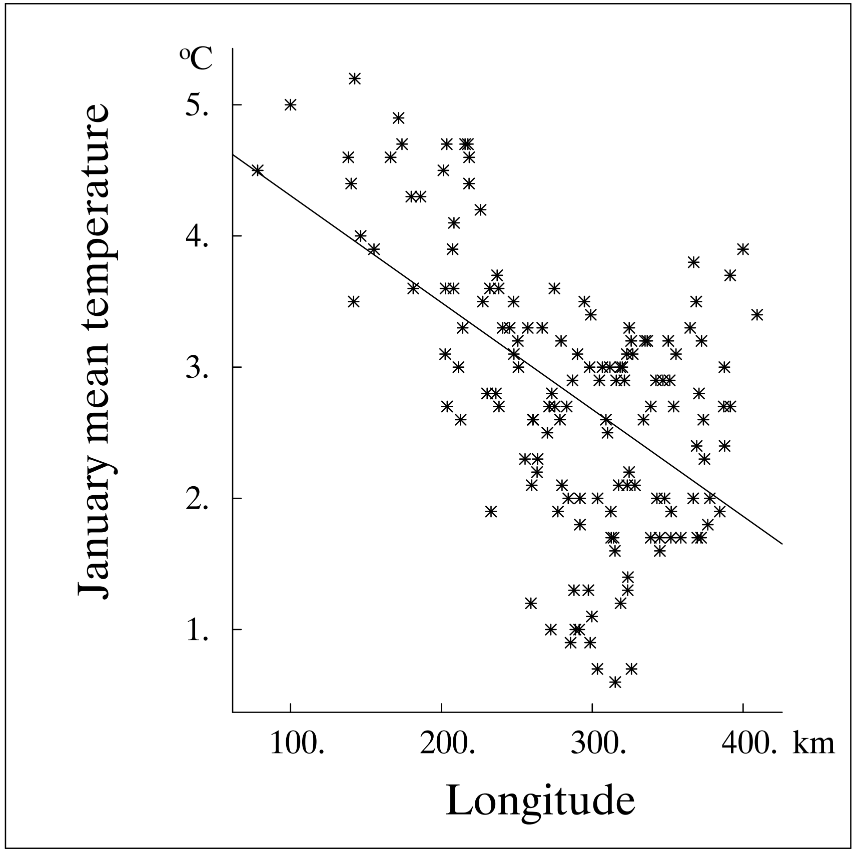
Elevation: at 3035 nodes of a regular grid

Highest elevation: 1272 m

<u>Reference:</u> Int. J. Clim., 14, 77–91, 1994

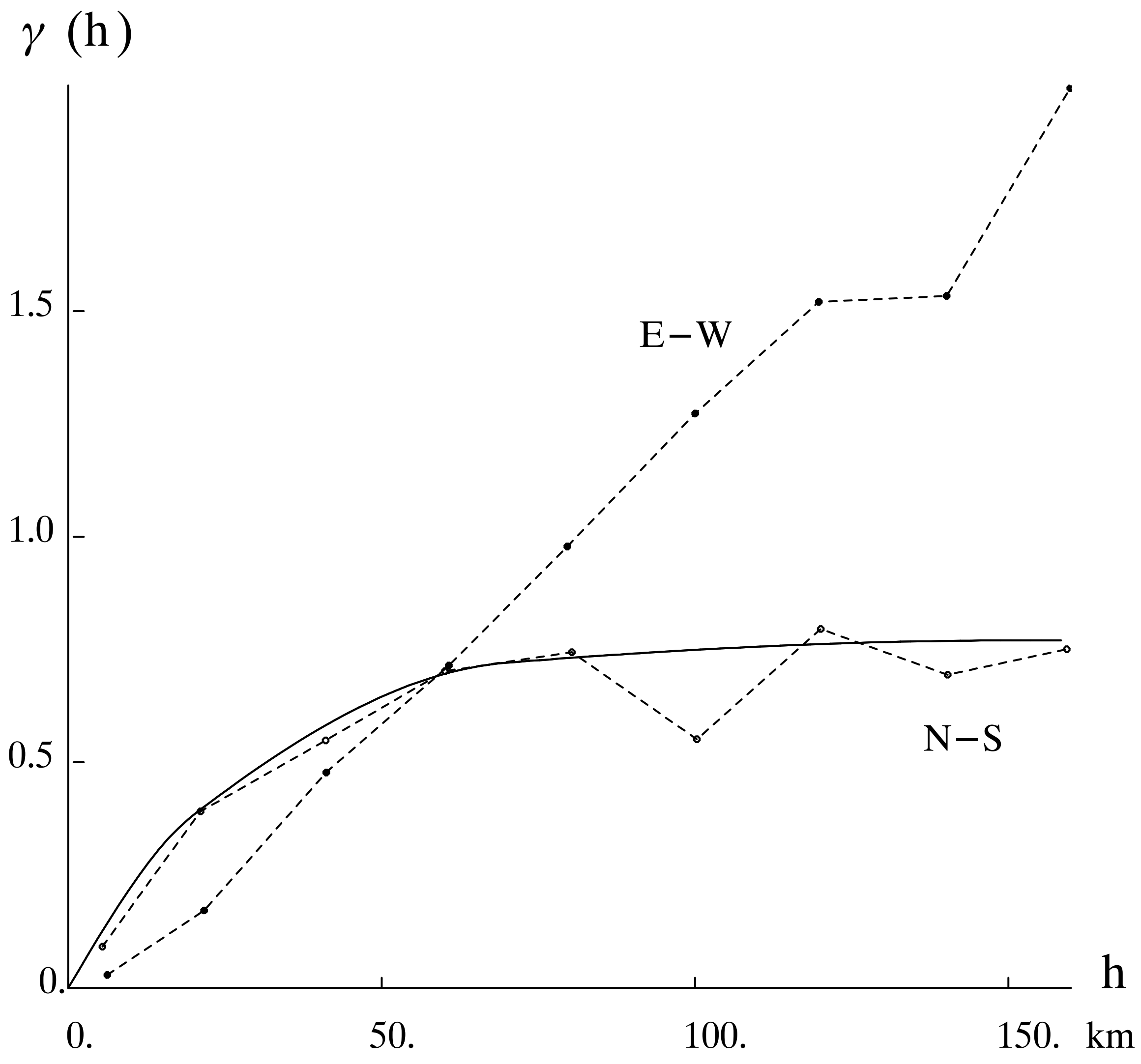
January temperature vs latitude / longitude





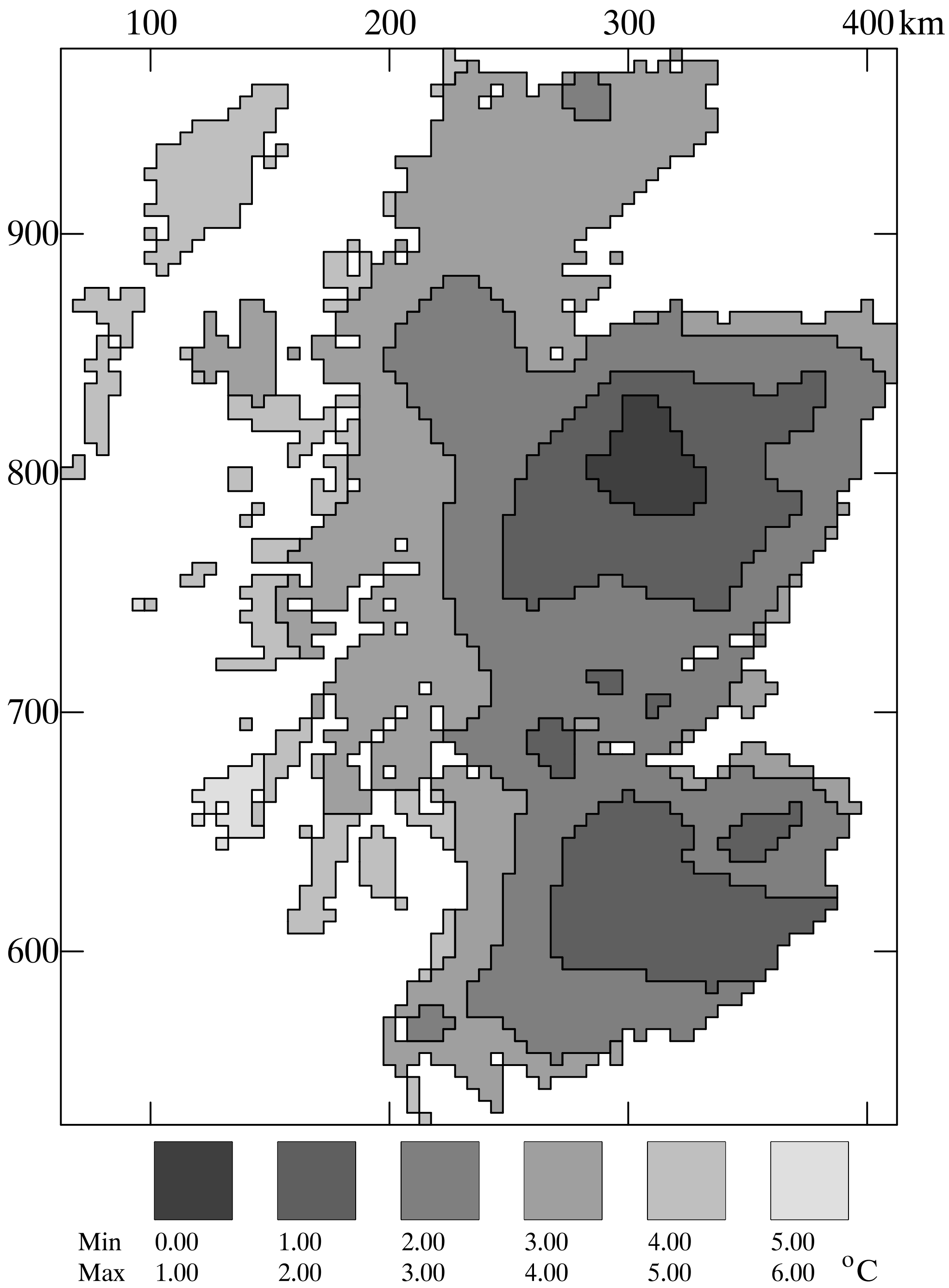
Scatter diagrams of temperature with latitude and longitude: there is a systematic decrease from west to east, while there is not much trend in the north-south direction.

E-W and N-S temperature variograms

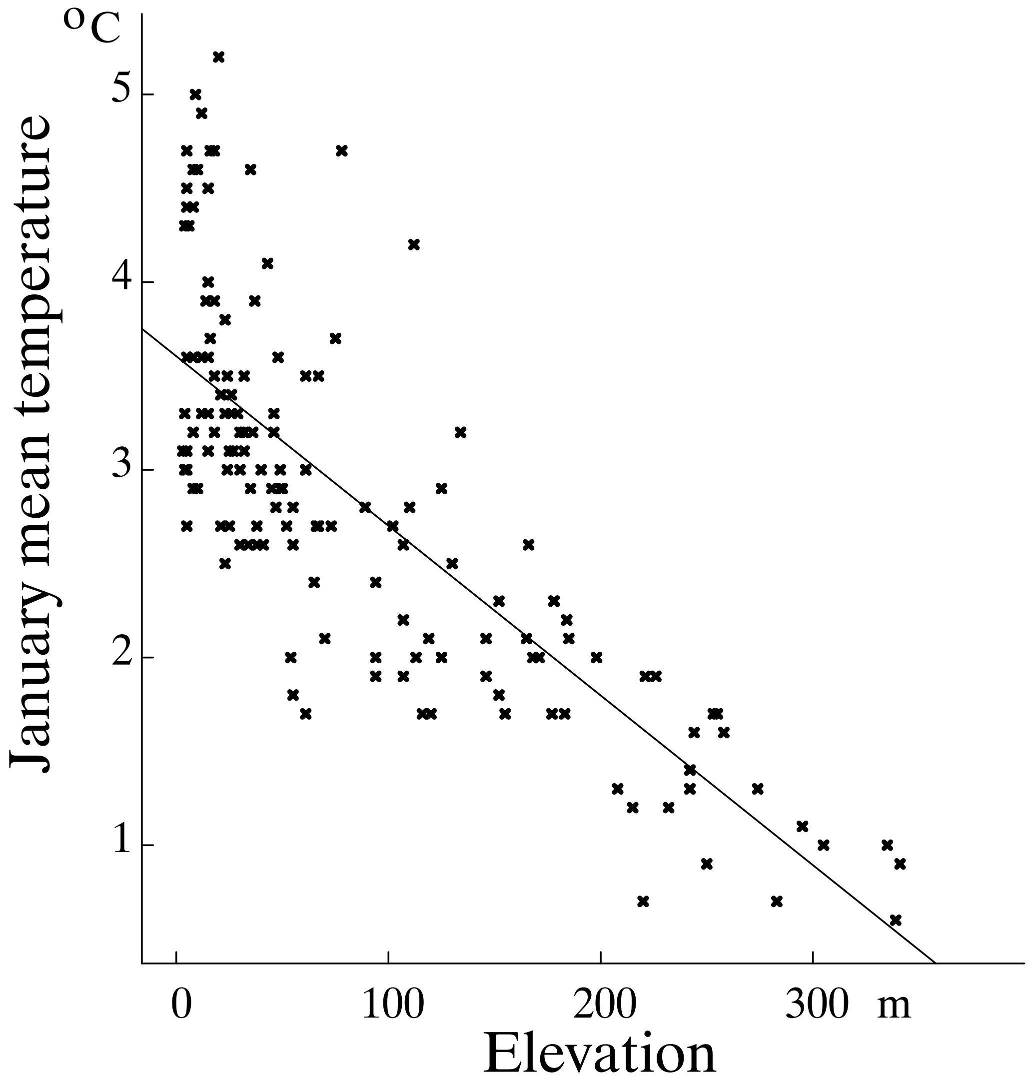


The model is fitted in the direction without drift

Kriging mean January temperature

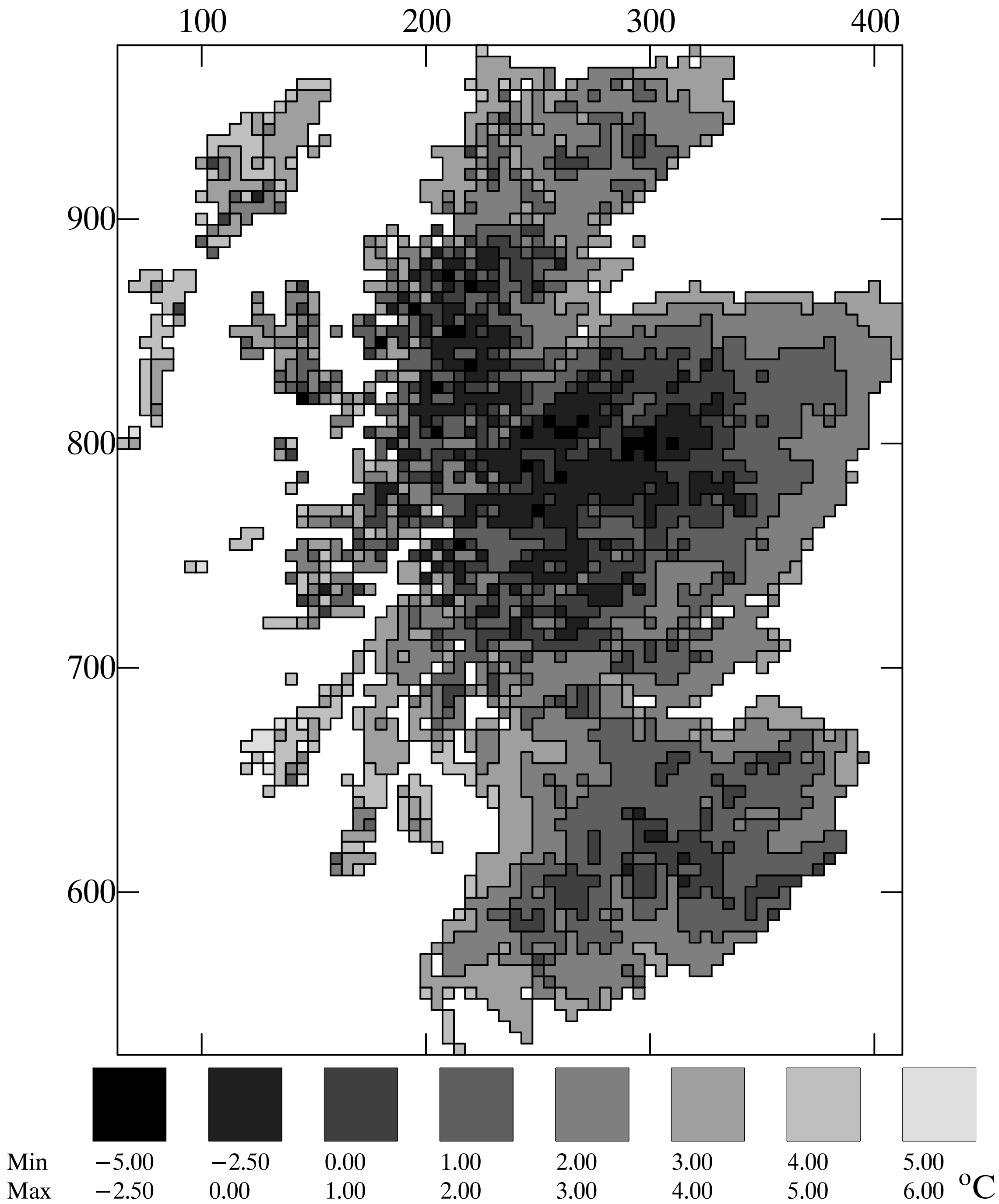


Temperature vs elevation

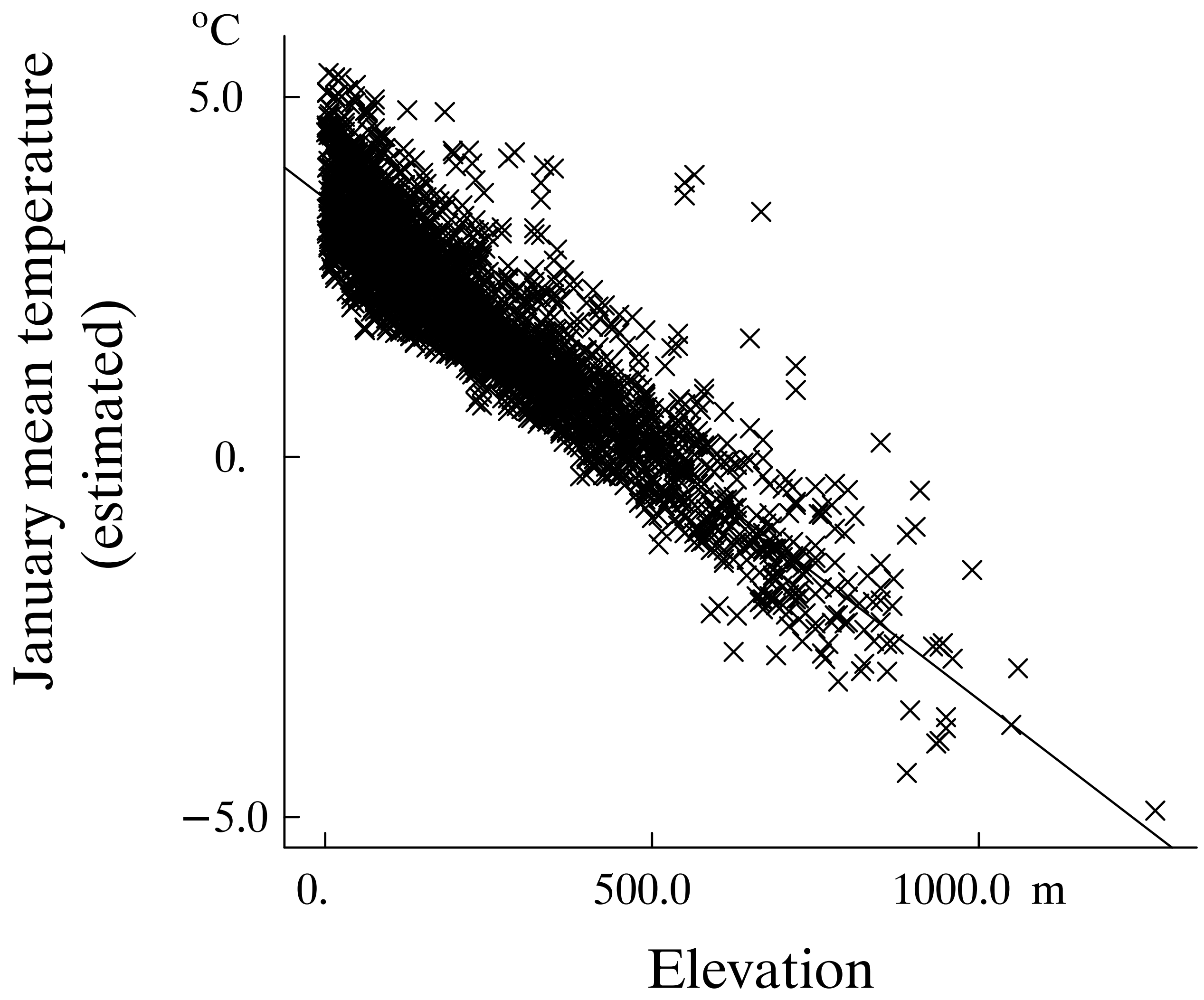


- Coastal influence below 50m, then linear relation.
- Temperatures only available below 400m.

Kriging temperature with elevation as drift



Temperature estimates vs elevation



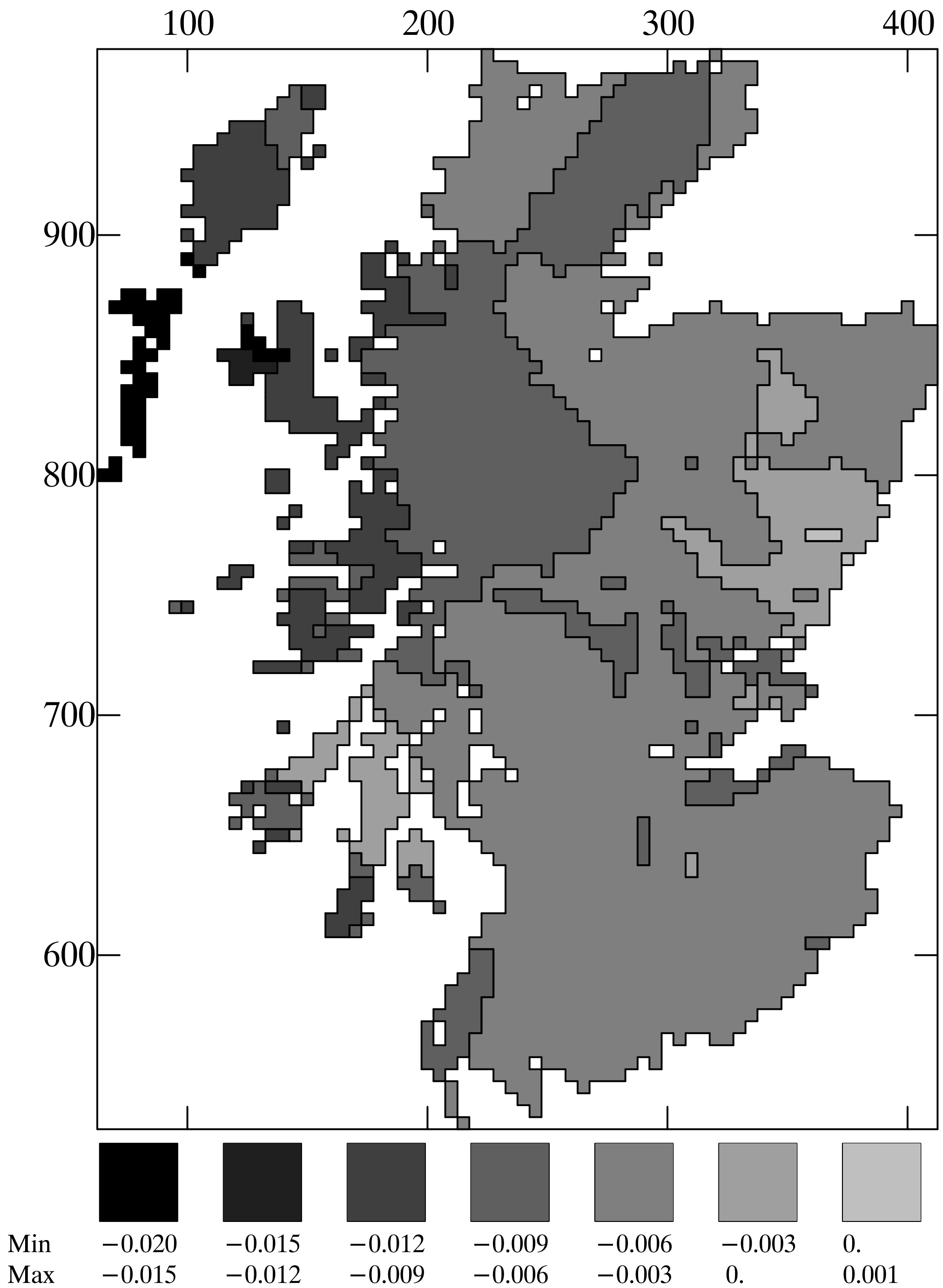
The estimated values above 400m are linearly extrapolated outside the range of the data !

Estimated external drift coefficient

$$b_1^\star = \sum_{\alpha=1}^{n} w_\alpha \, Z(\mathbf{x}_\alpha)$$

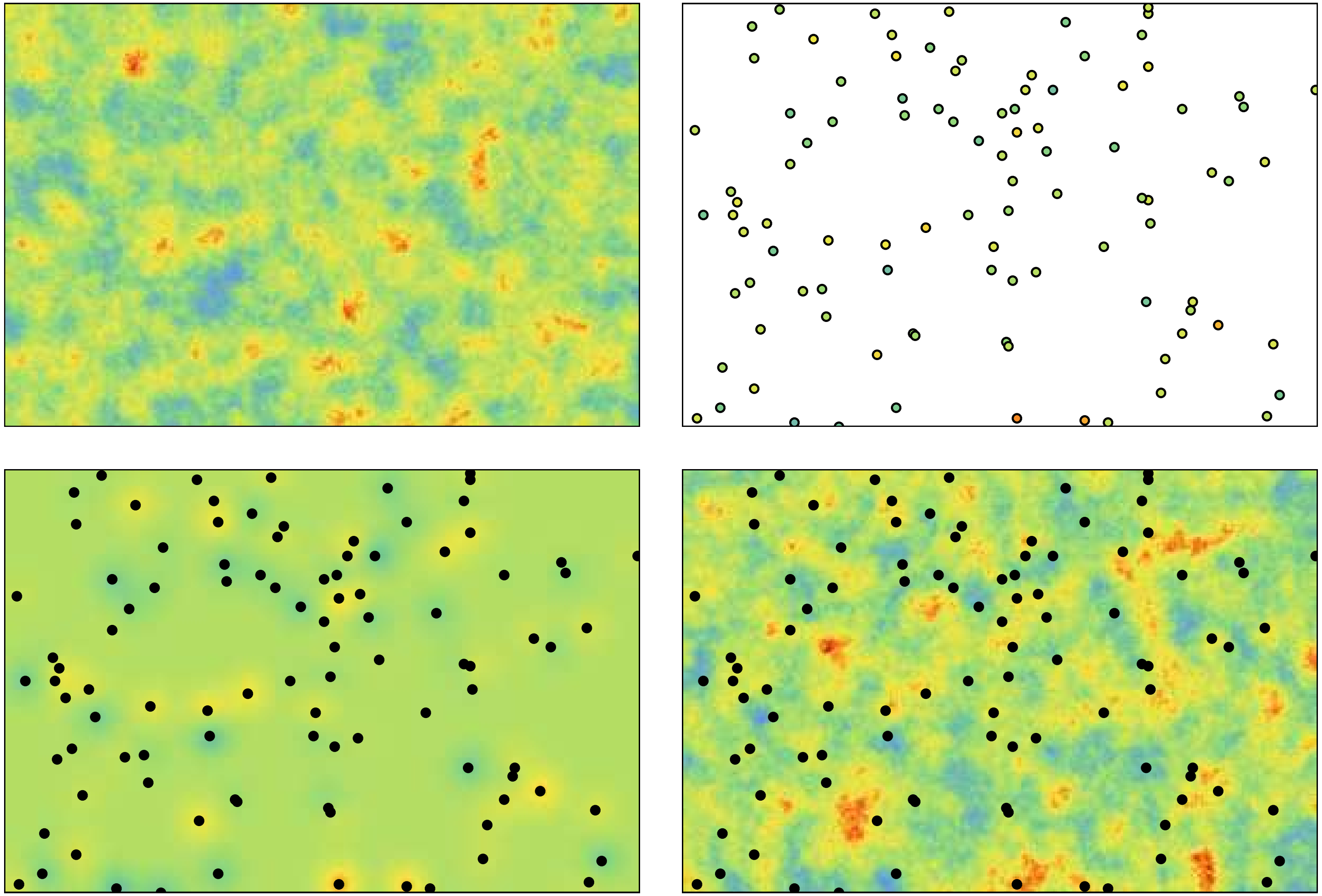
$$\begin{cases} \displaystyle\sum_{\beta=1}^{n} w_\beta \, C(\mathbf{x}_\alpha - \mathbf{x}_\beta) + \mu_0 + \mu_1 \, x_\alpha^1 + \mu_2 \, x_\alpha^2 + \mu_3 \, s(\mathbf{x}_\alpha) & = \boxed{0}, \ \forall \alpha \\ \displaystyle\sum_{\beta=1}^{n} w_\beta = \boxed{0} & \\ \displaystyle\sum_{\beta=1}^{n} w_\beta \, x_\beta^1 = \boxed{0} & \text{(longitude)} \\ \displaystyle\sum_{\beta=1}^{n} w_\beta \, x_\beta^2 = \boxed{0} & \text{(latitude)} \\ \displaystyle\sum_{\beta=1}^{n} w_\beta \, s(\mathbf{x}_\beta) = \boxed{1} & \text{(external drift)} \end{cases}$$

Estimated external drift coefficient



Conditional simulation

Conditional simulation vs Kriging



Model $m = 0$, $C = sph(1, 20)$. Simulation field 300×200.

Simulation (TL), conditioning data points (TR), simple kriging (BL) and conditional simulation (BL)

Change of support

geostatistical simulation of O_3

CASE STUDY: Geostatistical simulation of O_3

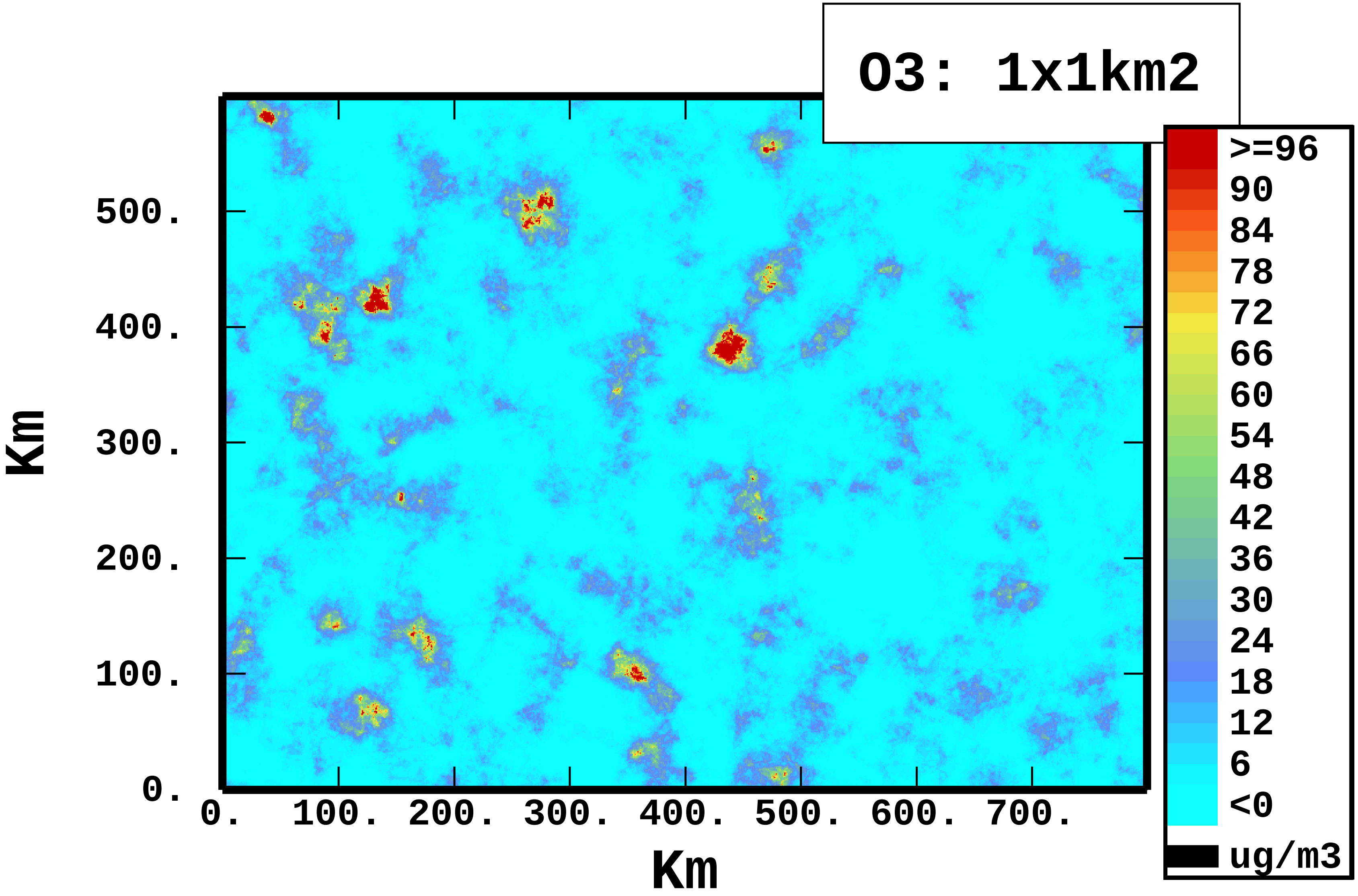
Simulation of realizations a lognormal random function

Region 800×600 Km^2

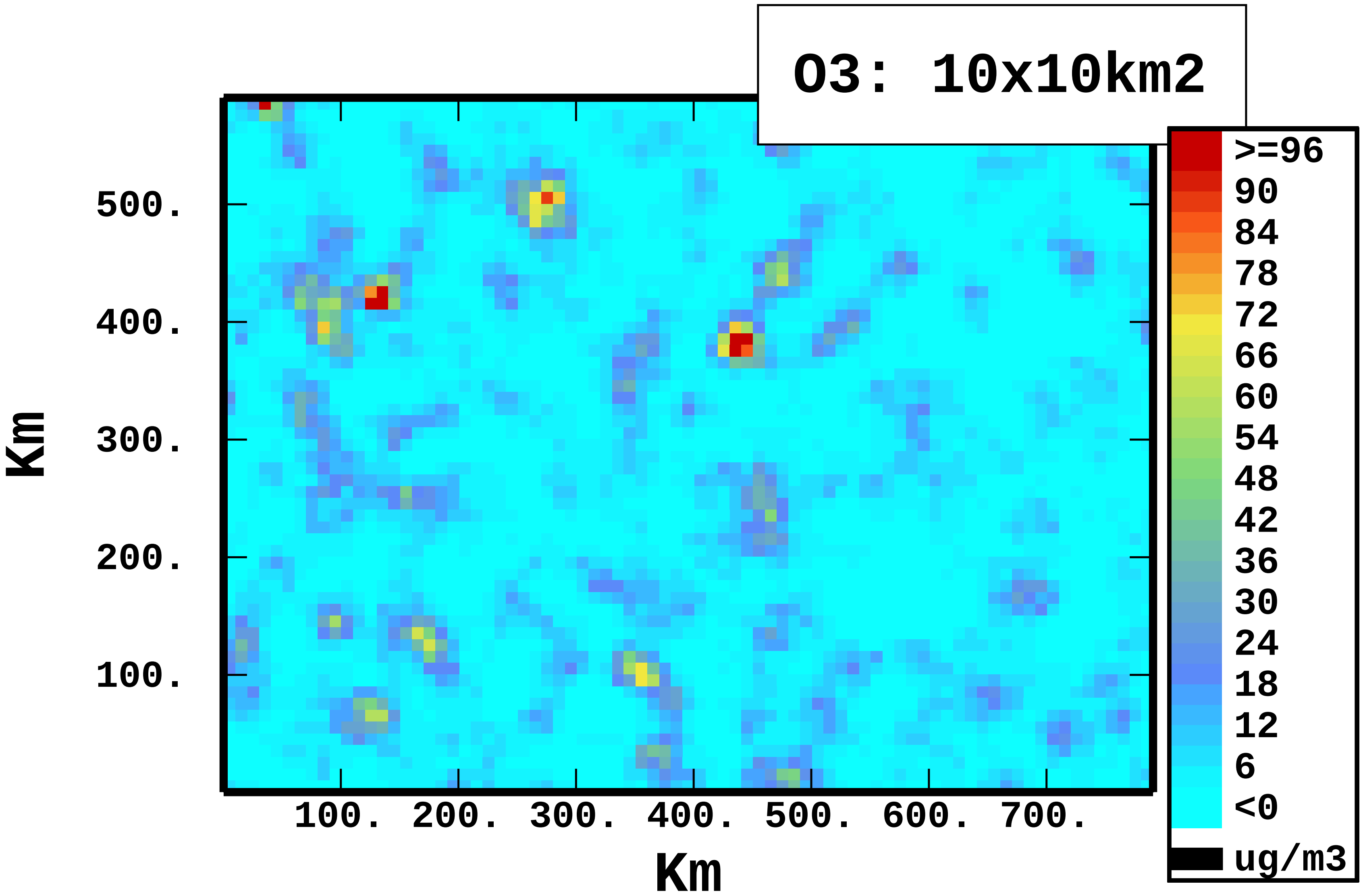
Cells 1×1 Km^2

Variogram with a range of 50 Km

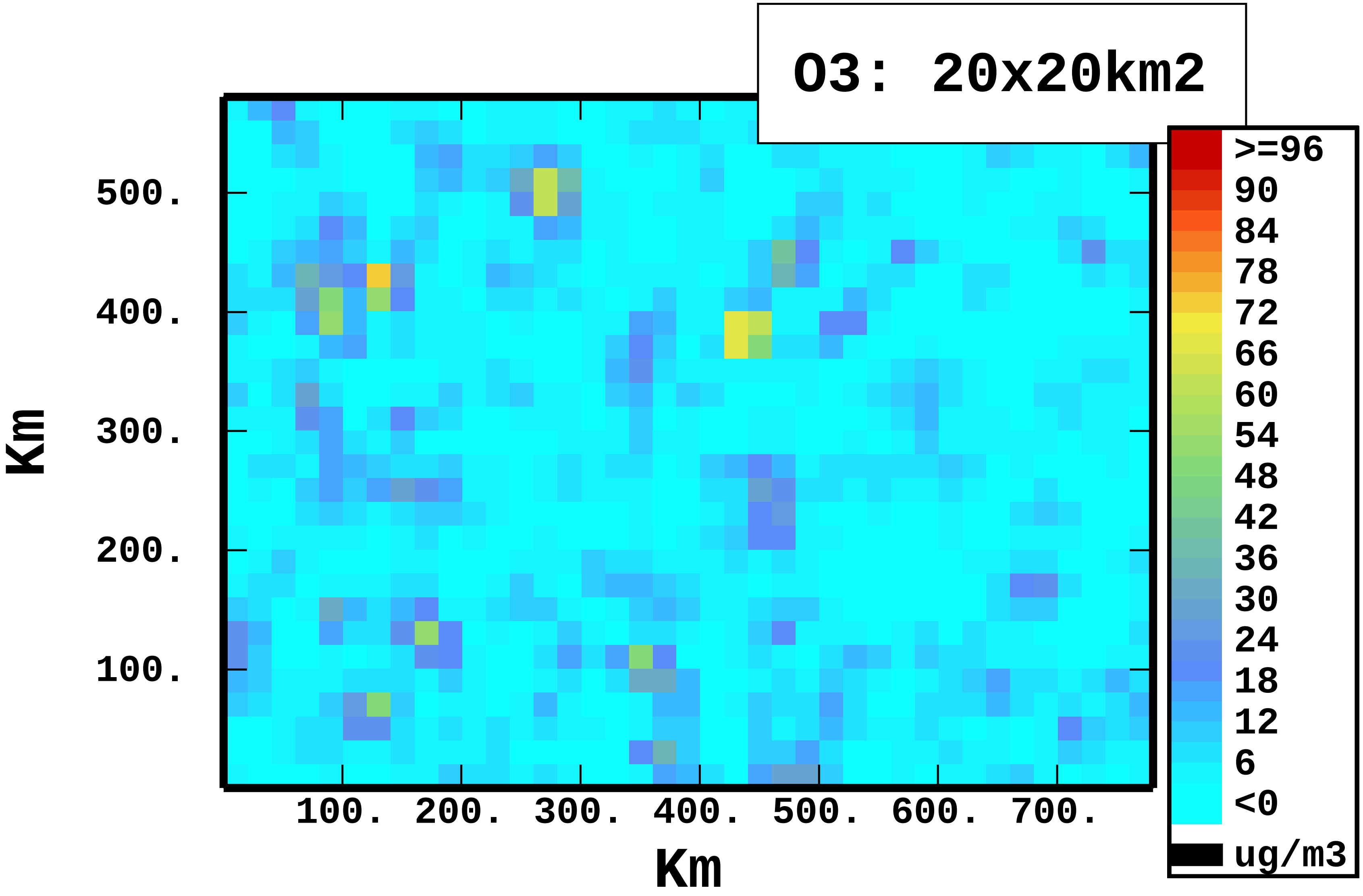
Simulation of Ozone: 1×1 Km^2 support



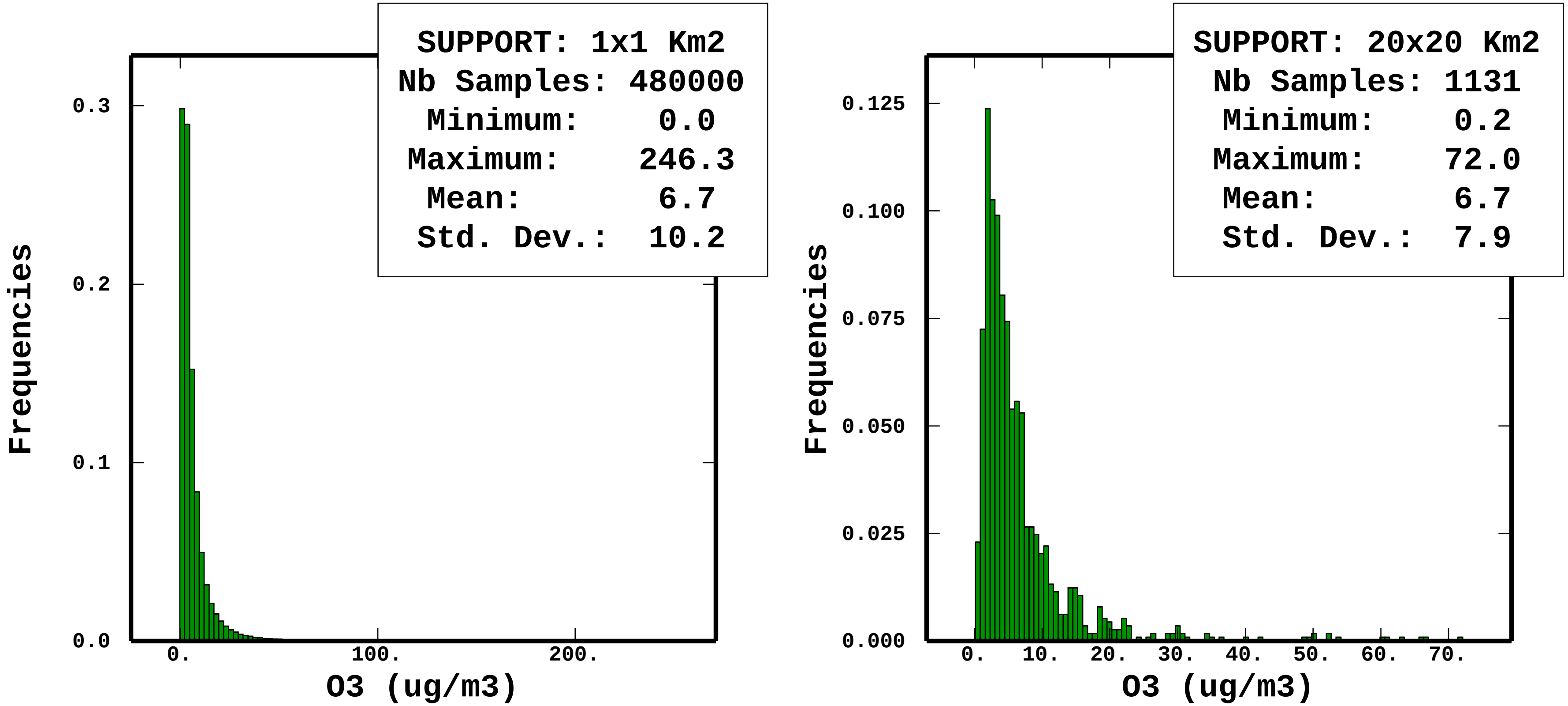
Simulation of Ozone: 10$\times$10 Km^2 support



Simulation of Ozone: 20×20 Km^2 support



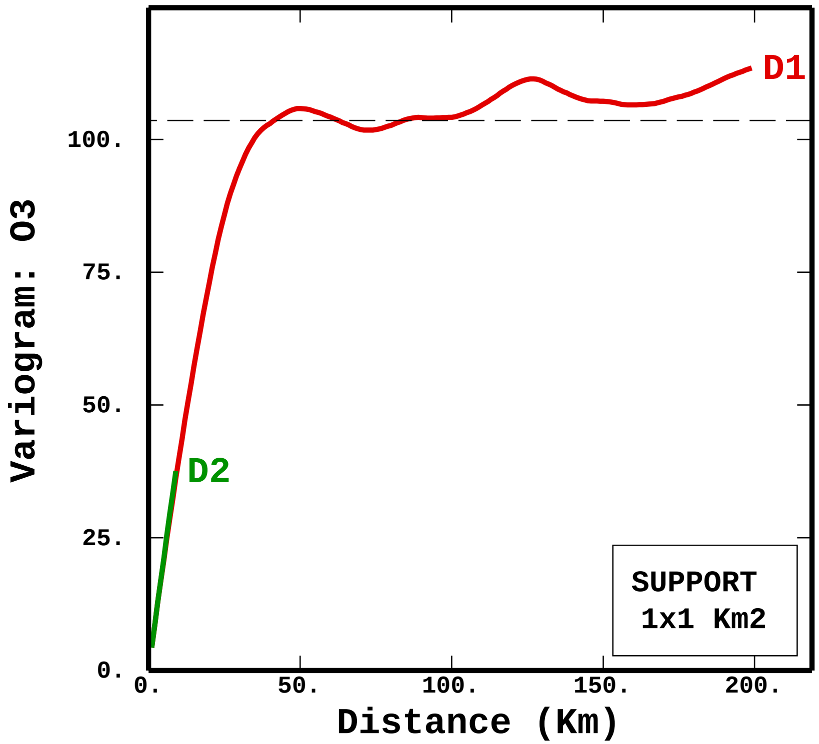
Simulation of Ozone

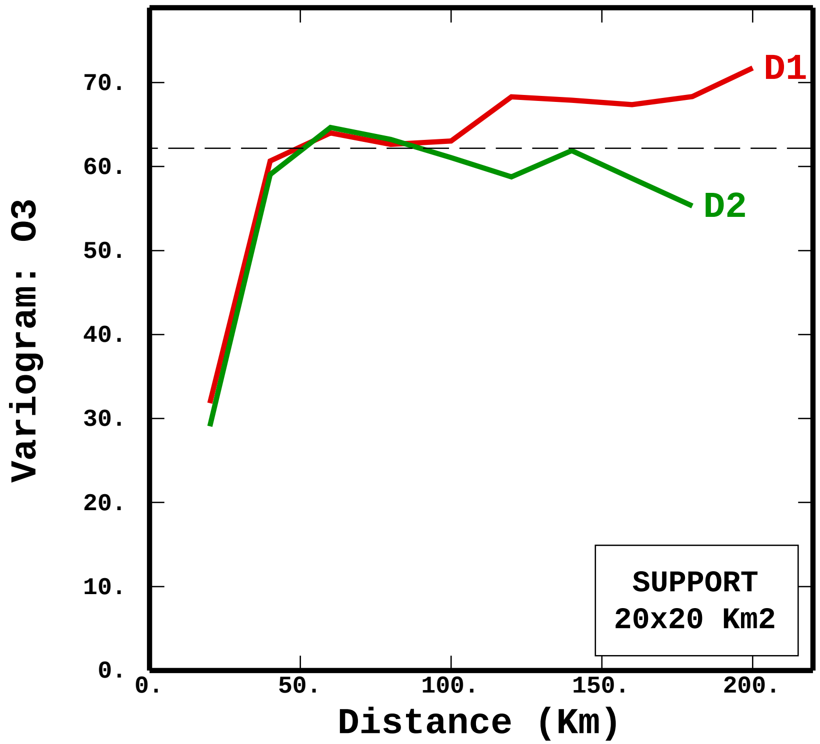


Increasing the support: the means are equal,

but the extremes and the variance are reduced

Simulation of Ozone

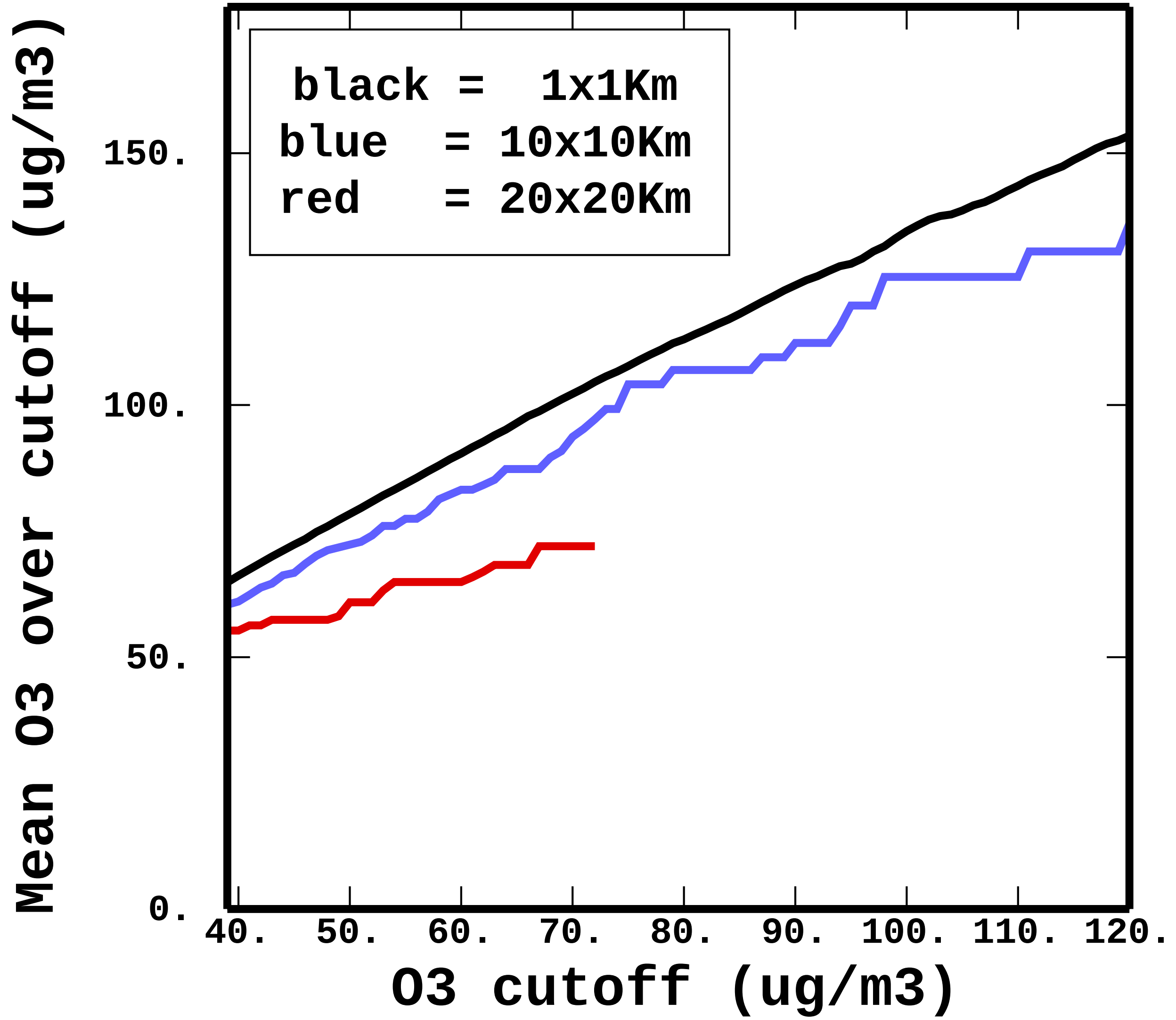




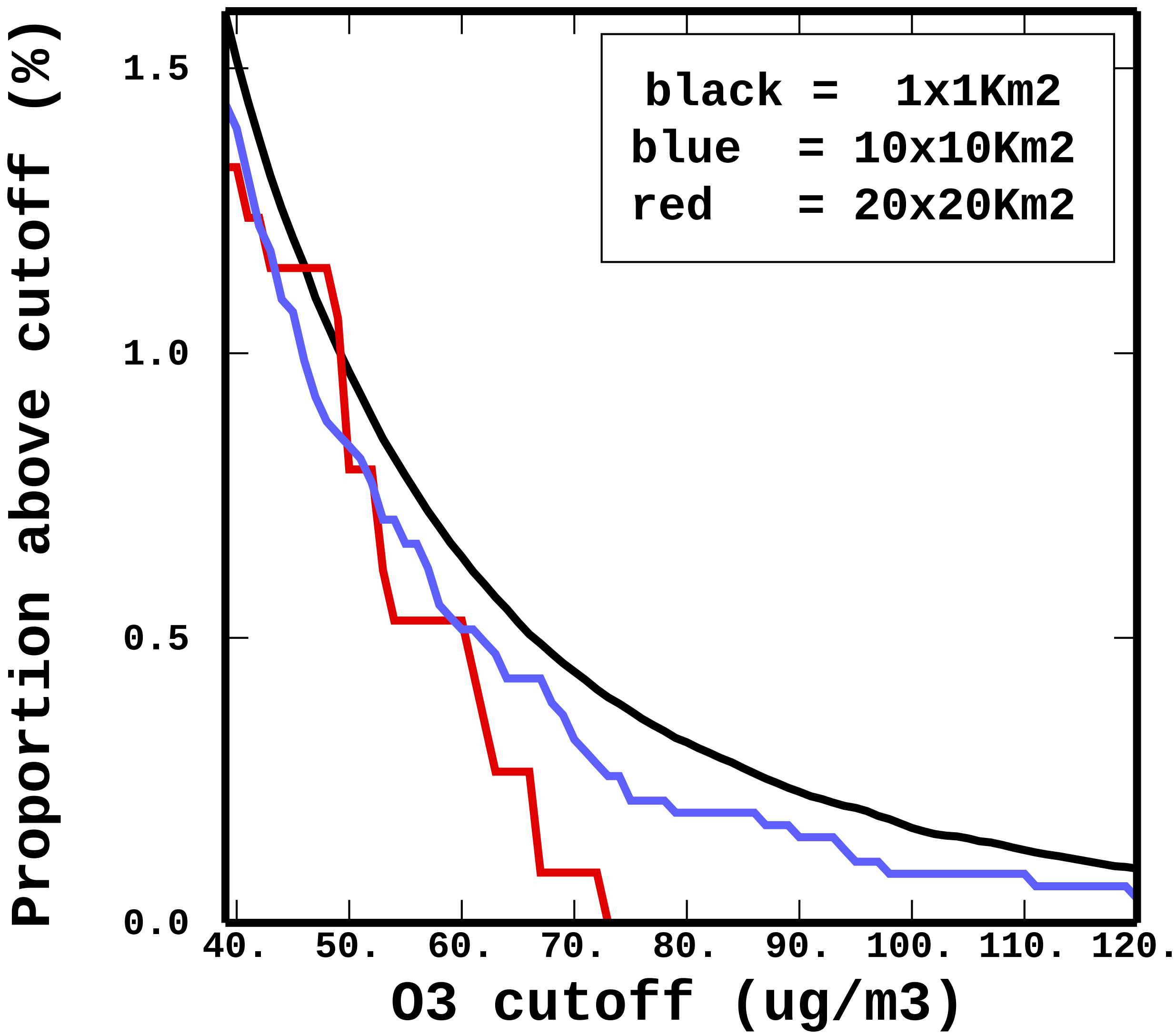
Increasing the support:

the range increases

Simulation of Ozone



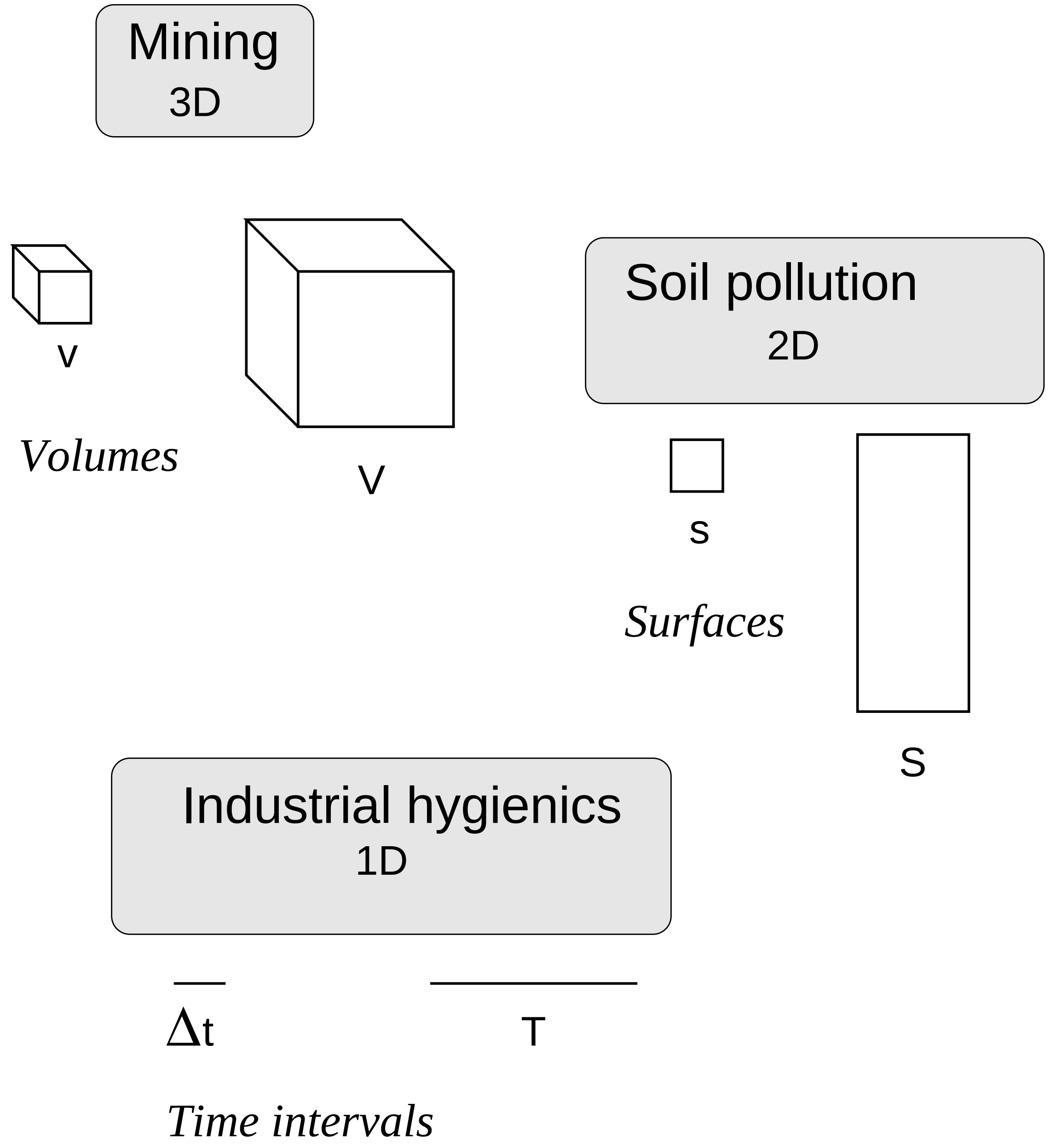
Simulation of Ozone



Change of support

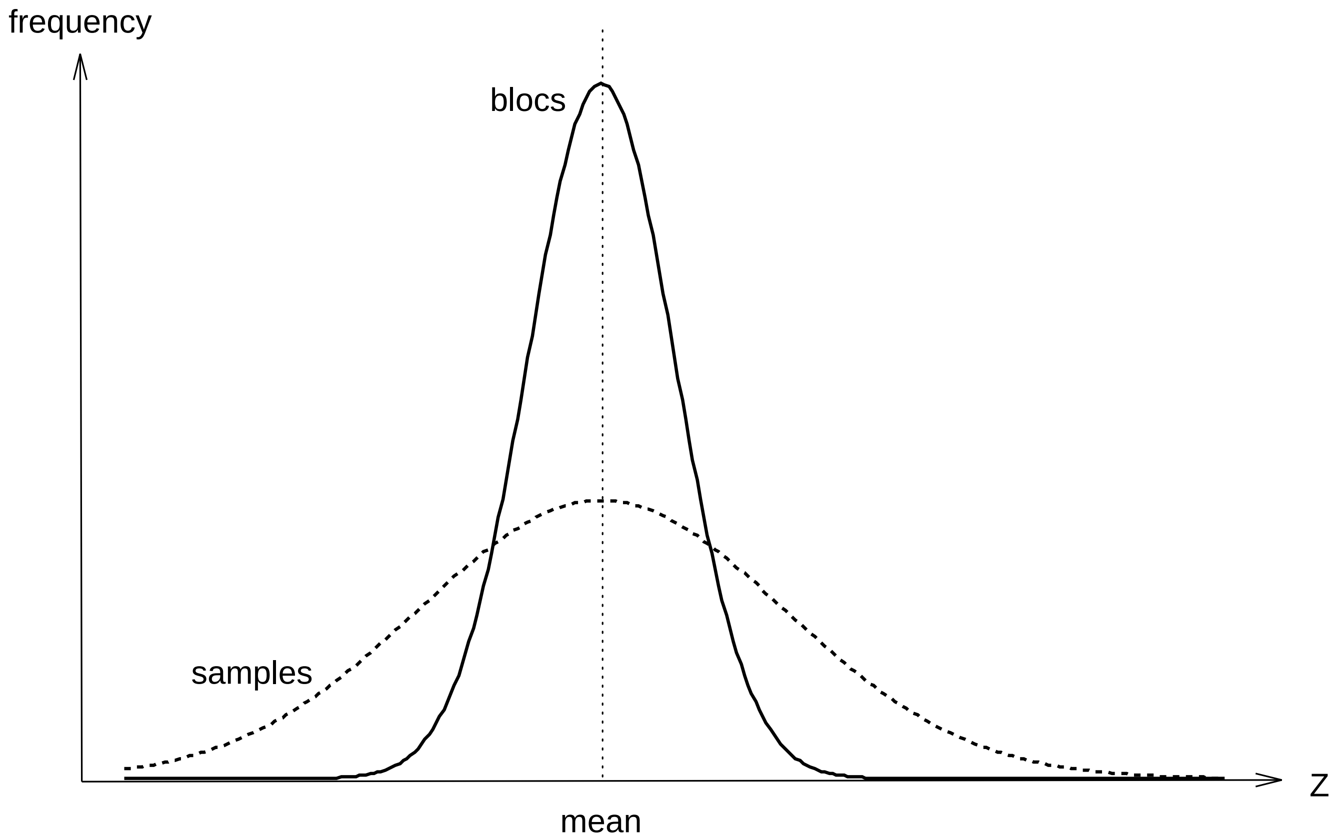
concept

TOPIC: The Support of a Random Function



The Effect of Changing the Support

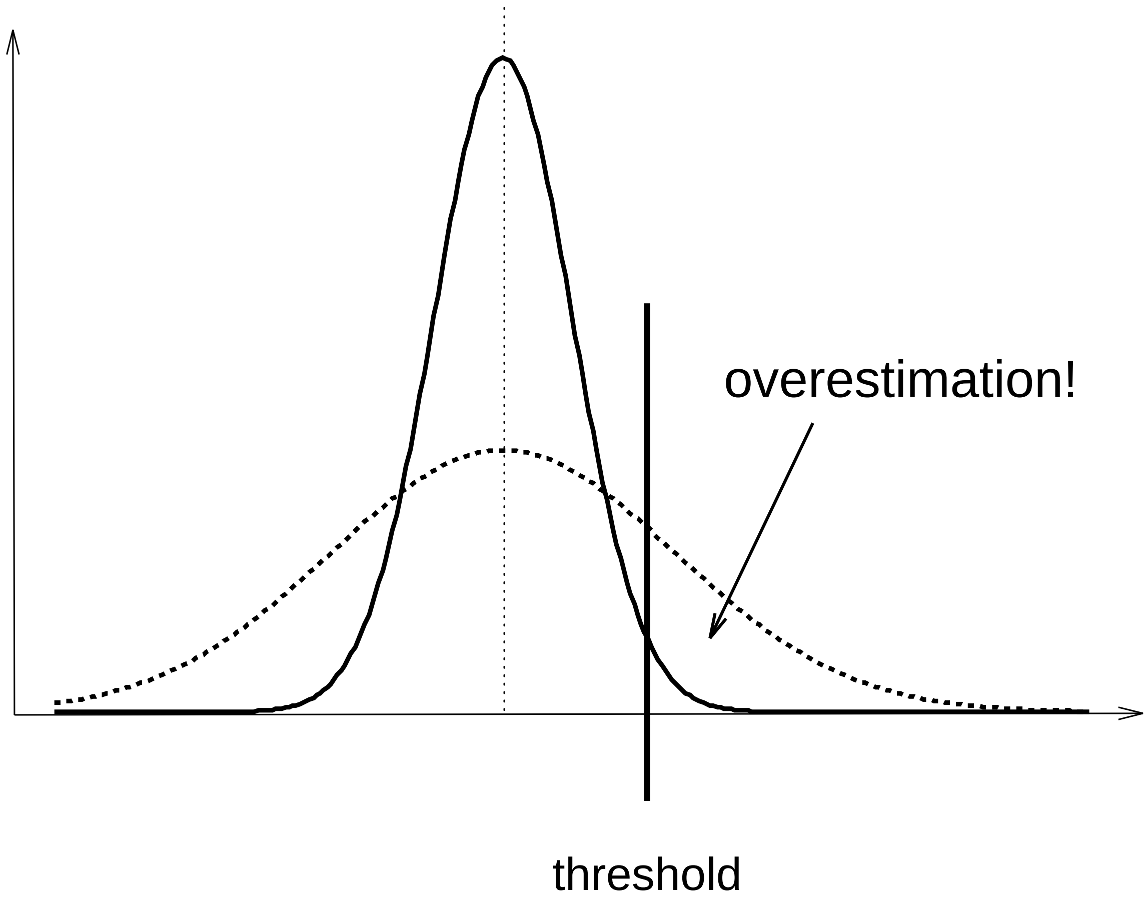
Distribution of samples on small volumes (cm^3) is different from that of model output averages over large blocks (m^3):



- The mean of both distributions is the same,
- the distribution of the block values is narrower.

Neglecting the Support Effect

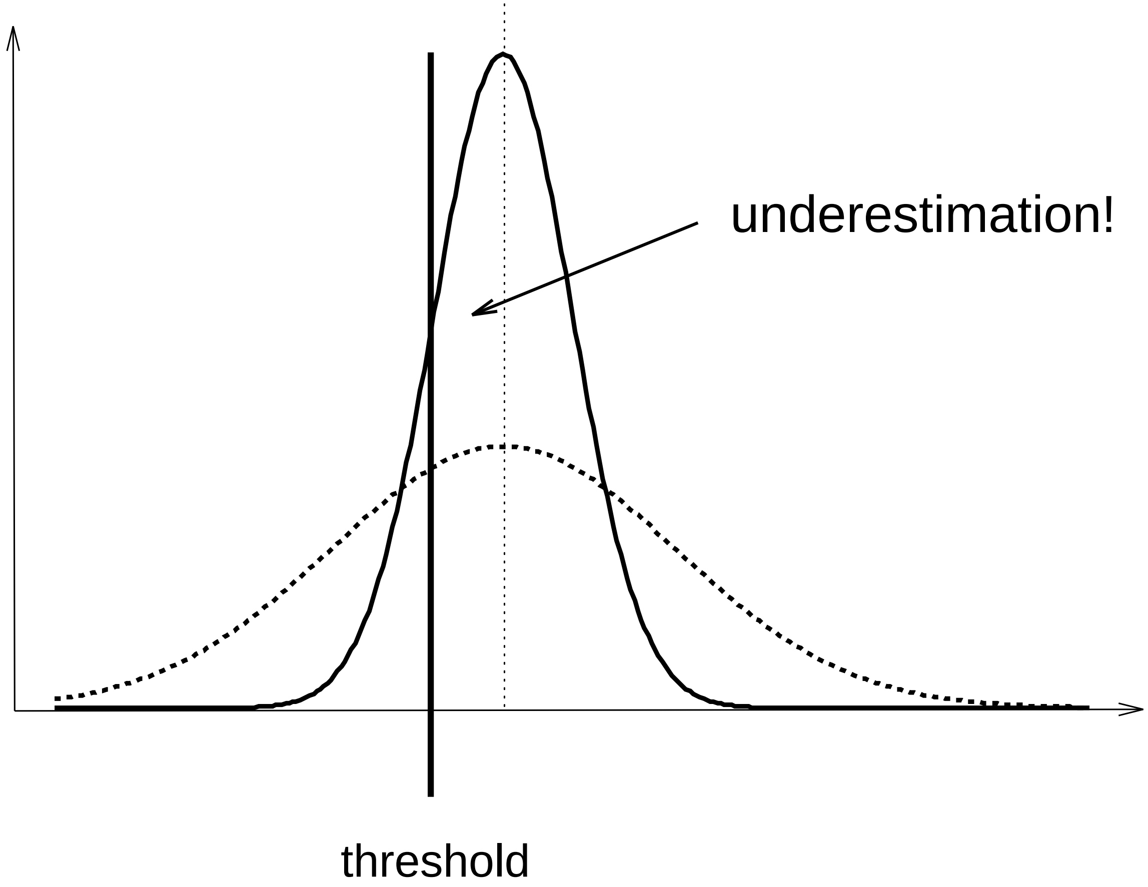
We are often interested in what is above a threshold:



Neglecting the support effect may lead to a systematic over-estimation. . .

Neglecting the Support Effect

. . . or to systematic under-estimation:



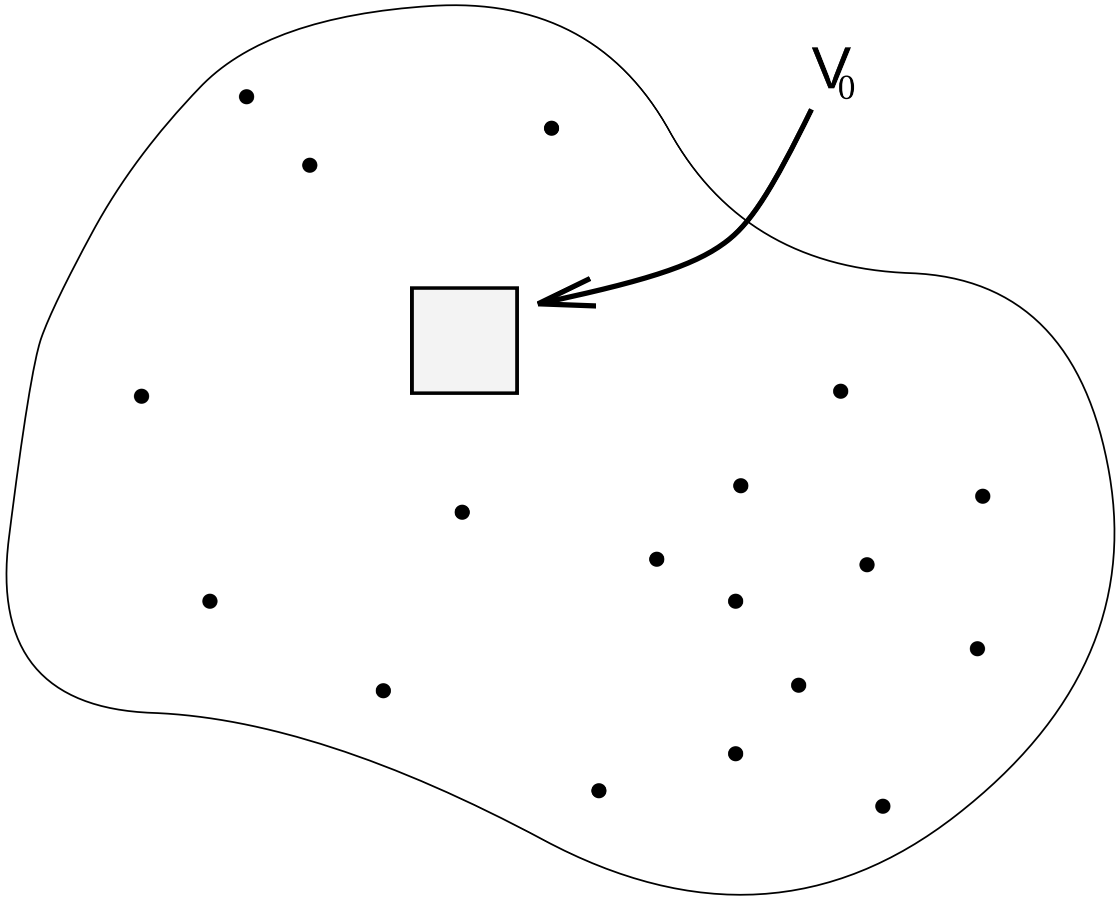
$\Rightarrow$ A good estimation method should incorporate a change of support model.

Kriging of a Block average

(centered at a point in the domain)

Estimation of a block value

Sample locations $\mathbf{x}_\alpha$ (dots)
in a domain $\mathcal{D}$:



We wish to estimate the spatial average $Z^\star$ for a block V_0.

Block Kriging

The block value $Z^\star(V_0)$ is estimated as a weighted average of the data values $Z(\mathbf{x}_\alpha)$:

$$Z^\star(V_0) \;=\; \sum_{\alpha=1}^{n} w_\alpha \, Z(\mathbf{x}_\alpha) \qquad \text{with} \sum_{\alpha=1}^{n} w_\alpha = 1$$

The optimal weights w_α^{OK} are obtained from the sytem:

$$\begin{cases} \displaystyle\sum_{\beta=1}^{n} w_\beta^{\mathrm{OK}} \, \gamma(\mathbf{x}_\alpha - \mathbf{x}_\beta) + \mu_{\mathrm{OK}} & = \boxed{\overline{\gamma}(V_0, \mathbf{x}_\alpha)} \qquad \forall\alpha \\ \displaystyle\sum_{\beta=1}^{n} w_\beta^{\mathrm{OK}} & = 1 \end{cases}$$

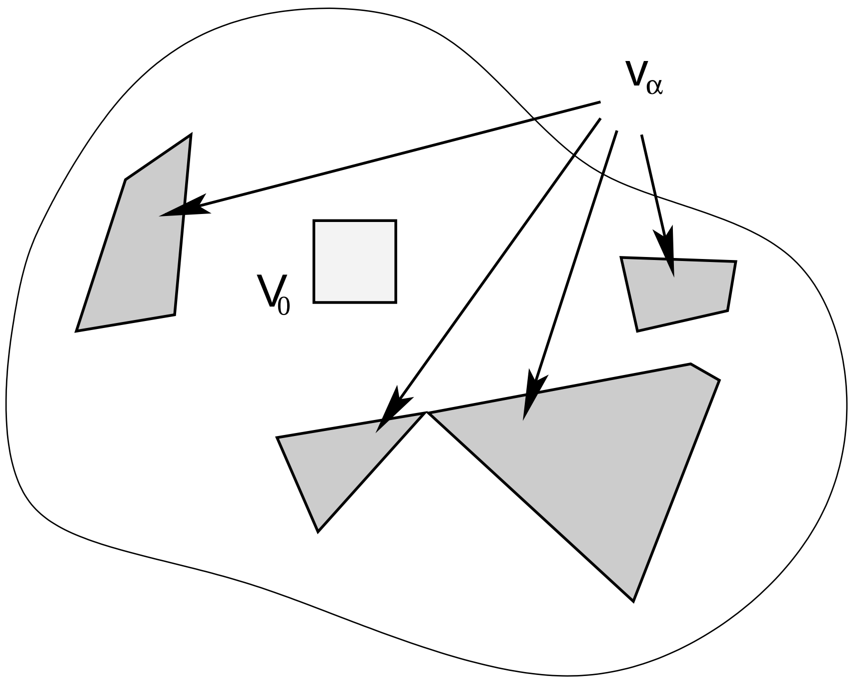
Kriging variance: $\sigma^2_{\mathrm{OK}} = \mu_{\mathrm{OK}} - \overline{\gamma}(V_0, V_0) + \displaystyle\sum_{\alpha=1}^{n} w_\alpha^{\mathrm{OK}} \, \overline{\gamma}(V_0, \mathbf{x}_\alpha)$

Block kriging with non-point data

In applications the data can be averaged on blocks V_α.
We then use average variograms between these blocks:

$$\overline{\gamma}\Big(V_\alpha, V_\beta \Big) \;=\; \frac{1}{|V_\alpha|\,|V_\beta|} \int\limits_{\mathbf{x}\in V_\alpha} \int\limits_{\mathbf{y}\in V_\beta} \gamma(\mathbf{x}-\mathbf{y})\, d\mathbf{x}\, d\mathbf{y}$$

This requires the knowledge of the point variogram.



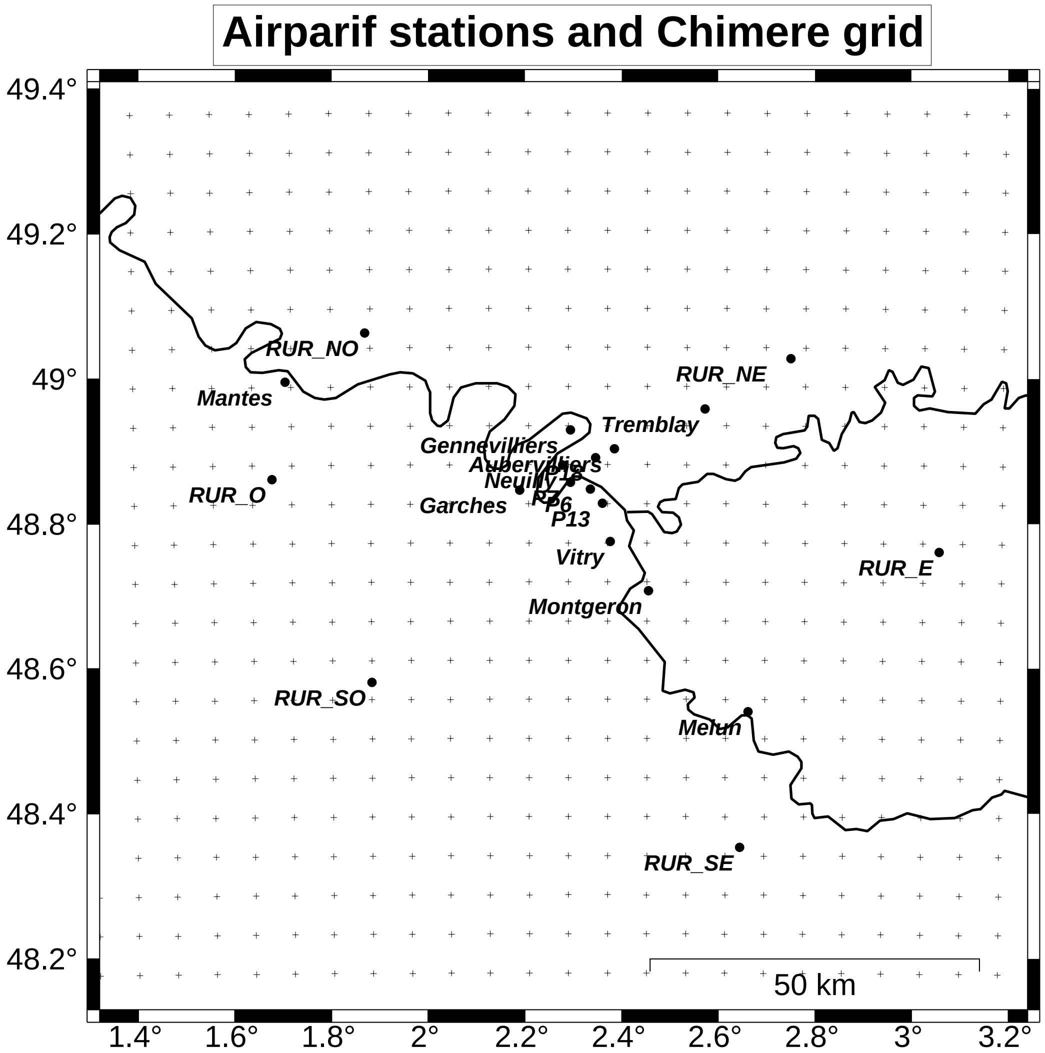
Change of support

risk of exceeding ozone alert level

Change of support

- The variability of spatial or temporal data depends on the averaging volume/interval(= the support)
  Increasing support, the variability decreases (reduction of variance, extremes...)
- Observations are on point support as compared to the cells of a numerical model.
  End-users are often interested by a support of different (intermediate) size $\longrightarrow$ blocks
- It is thus necessary to describe statistically how variability changes as a function of support.
  If the distribution is monomodal and not too asymmetrical, an affine correction may suffice. Otherwise, non-linear geostatistics or geostatistical simulation are needed
- Applications: data aggregation, estimation of small block statistics, downscaling. . .

Ozone in Paris on 17 july 1999 at 15h UTC



19 Airparif stations; 25×25 grid with cells of size 6×6 km^2

Air quality regulations

Two ozone thresholds refering to a support of 1 hour:

$\longrightarrow$ Swiss alert level: 120 μg/m^3

$\longrightarrow$ European alert level: 180 μg/m^3

Time support is always specified, yet regulations do not contain any indication about the spatial support !

Suppose the air quality experts agree on the following spatial decision support:

a block of 1×1 km^2 size

(instead of the CHIMERE 6×6 km^2 cell).

We need to model the point-block-cell change of support.

Discrete Gaussian point-block model

(due to Georges MATHERON, 1976)

$\underline{\mathbf{x}}$ is a point randomly located in a block v.

$$\mathrm{E}\big[\, Z(\underline{\mathbf{x}}) \mid Z(v) \,\big] \;=\; Z(v),$$

is known as *Cartier*'s relation.

For a Gaussian point anamorphosis (station data),

$$Z(\mathbf{x}) \;=\; \varphi(Y(\mathbf{x})) = \sum_{k=0}^{\infty} \frac{\varphi_k}{k!}\, H_k(Y(\mathbf{x}))$$

with Hermite polynomials H_k and coefficients φ_k,
the block anamorphosis $\varphi_v(Y(v))$ comes as:

$$\varphi_v(Y(v)) \;=\; \mathrm{E}\big[\, \varphi(Y(\underline{\mathbf{x}})) \mid Y(v) \,\big] = \sum_{k=0}^{\infty} \frac{\varphi_k}{k!}\, r^k\, H_k(Y(v)).$$

Point-block-cell correlations

The Gaussian block anamorphosis is:

$$\varphi_v(Y(v)) \;=\; \sum_{k=0}^{\infty} \frac{\varphi_k}{k!}\, r^k\, H_k(Y(v)),$$

with r being the point-block coefficient $(0 \leq r \leq 1)$.

r can be computed from the block dispersion variance (which is calculated from the station data variogram):

$$\operatorname{var}(Z(v)) \;=\; \operatorname{var}(\varphi_v(Y(v))) = \sum_{k=1}^{\infty} \frac{\varphi_k^2}{k!}\, r^{2k}$$

We get in the same way a point-cell coefficient r'.

And finally the block-cell coefficient $r_{vV} = r'/r$.

Uniform conditioning

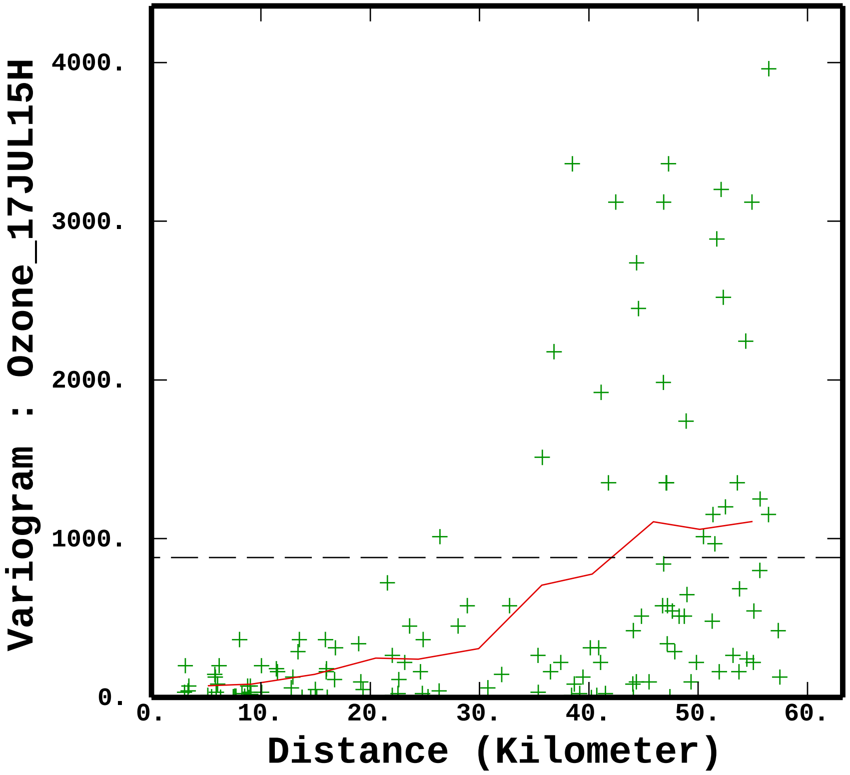
It consists in taking the conditional expectation of a non-linear function of blocks knowing the cell value containing them.

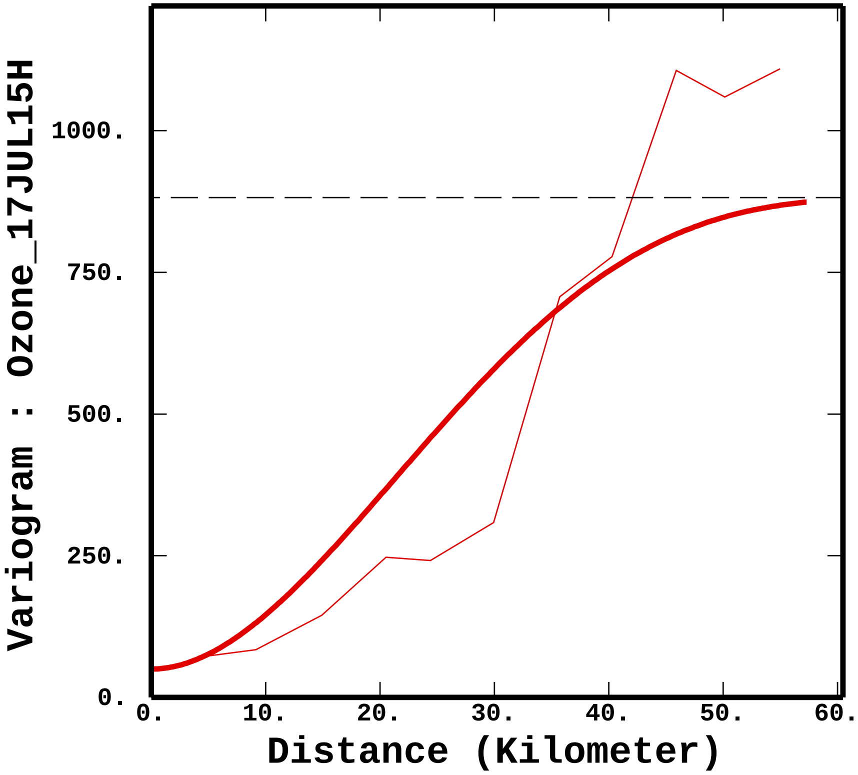
The proportion of blocks $v \in V_0$ above the threshold z_c knowing the cell value $Z(V_0)$ is:

$$\mathrm{E}\Big[\, \boldsymbol{1}_{Z(\underline{v})\geq z_c} \mid Z(V_0) \Big] \;=\; 1 - G\left(\frac{y_c - r_{vV}\, Y(V_0)}{\sqrt{1 - r_{vV}{}^2}}\right).$$

G is the Gaussian distribution.

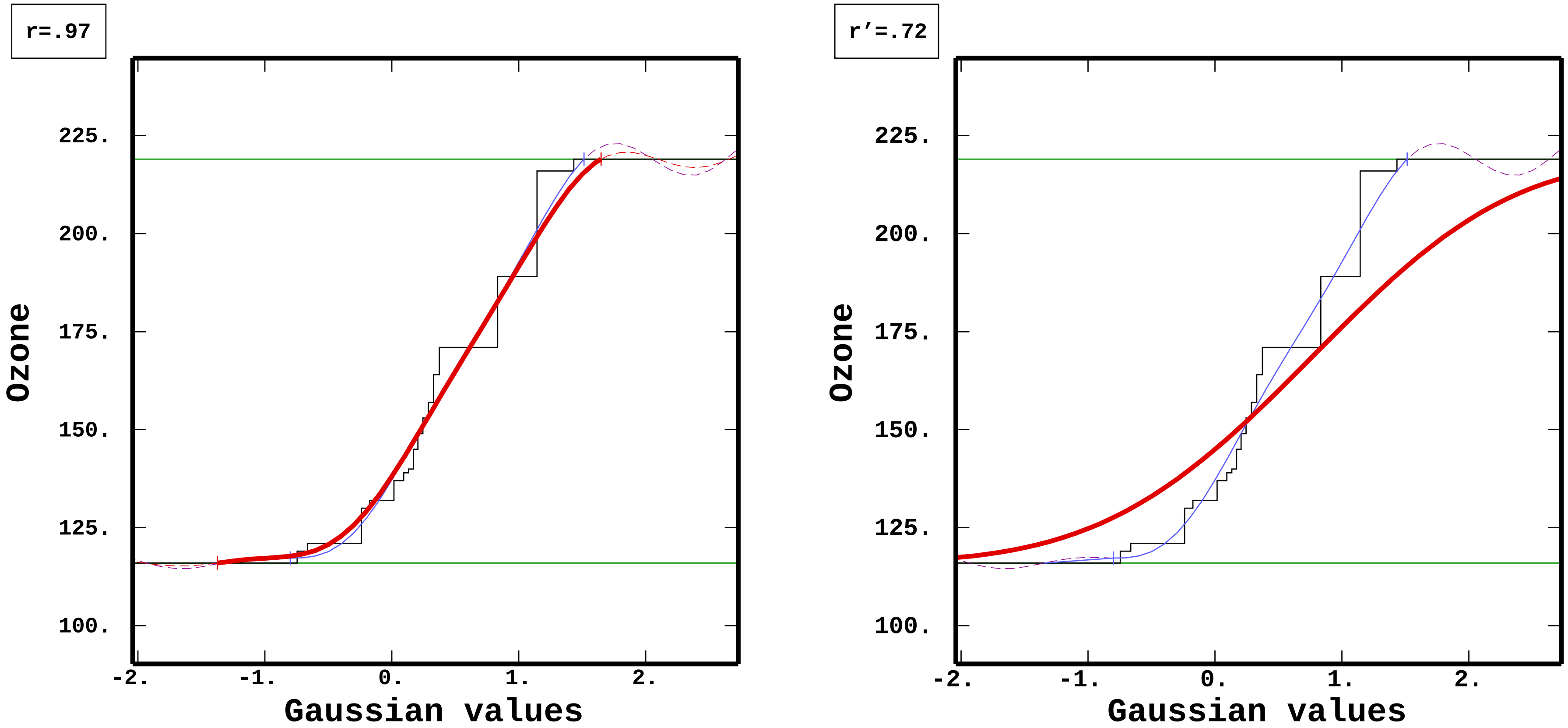
Variogram of Airparif measurements





Nugget-effect + cubic model.
Sill = variance.

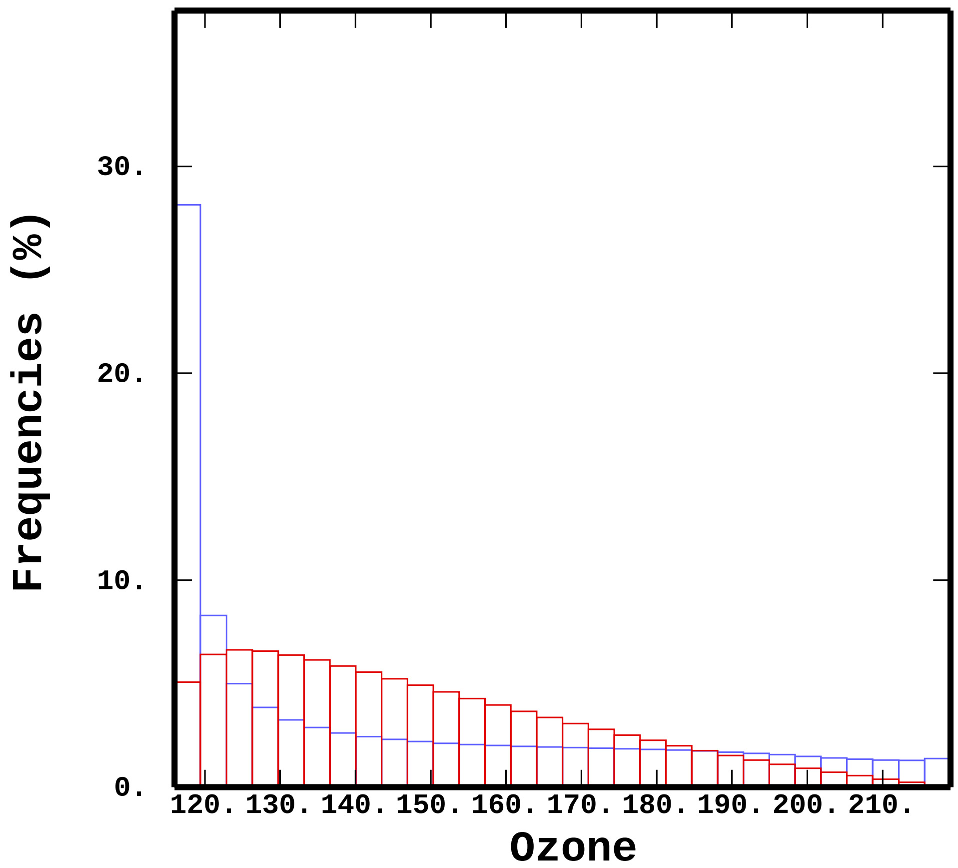
Anamorphosis of Airparif measurements



Anamorphosis of block values (r=.97)
close to the anamorphosis of point values.

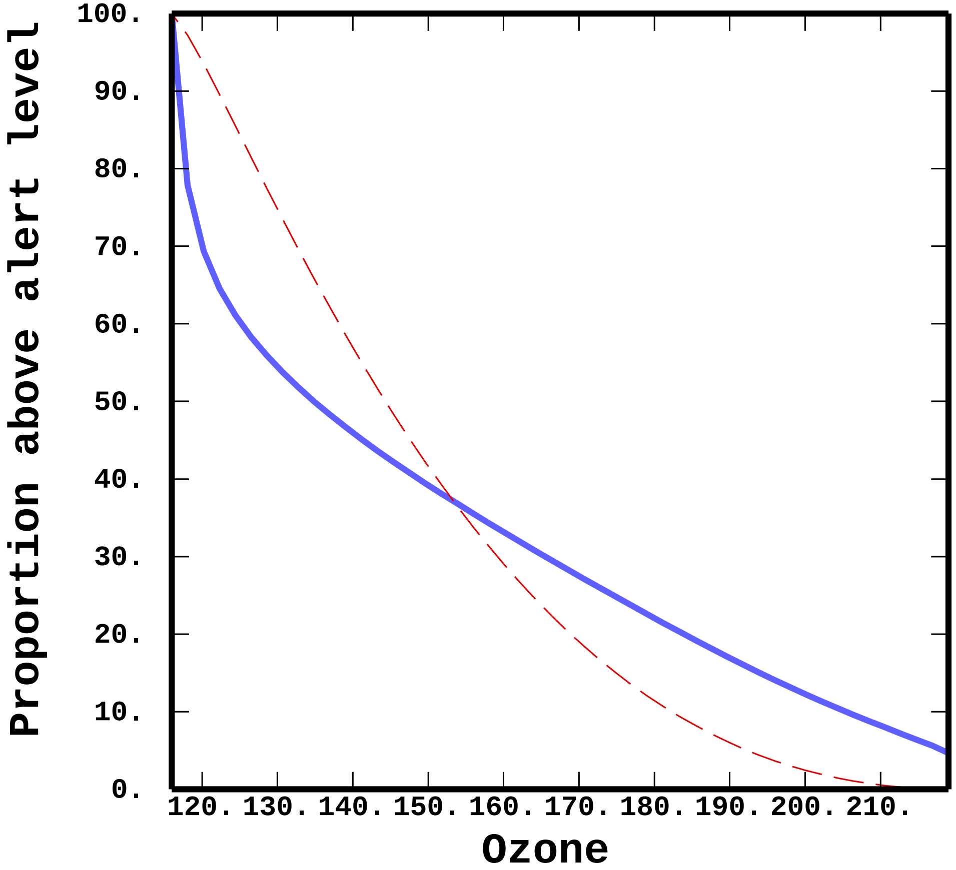
Anamorphosis of cell values (r'=.72).

Histograms



Histograms of blocks (blue) and cells (red)
on the basis of the change-of-support model.

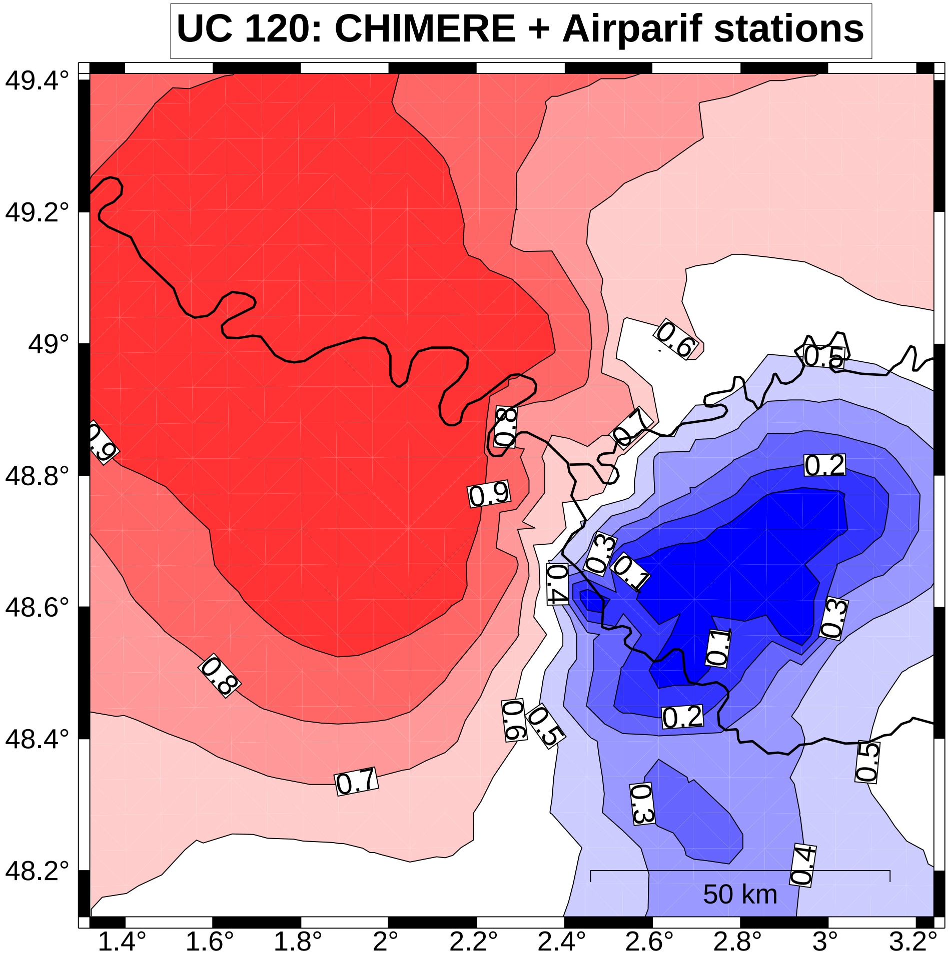
Proportion of values above threshold



Proportions of blocks (blue) and cells (red).

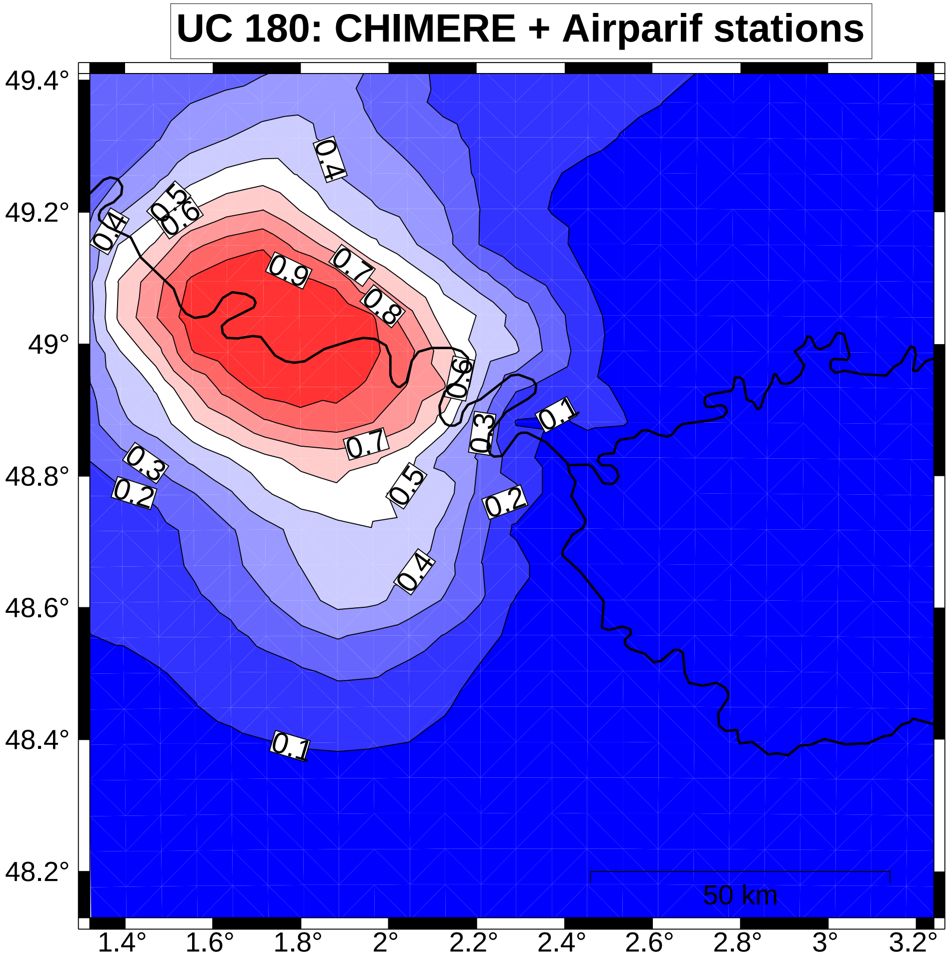
Depending on the threshold, the difference can be important !

Uniform conditioning by CHIMERE



Exceedance probabilities for 1×1 km^2 support
with the Swiss threshold of 120 μg/m^3

Uniform conditioning by CHIMERE

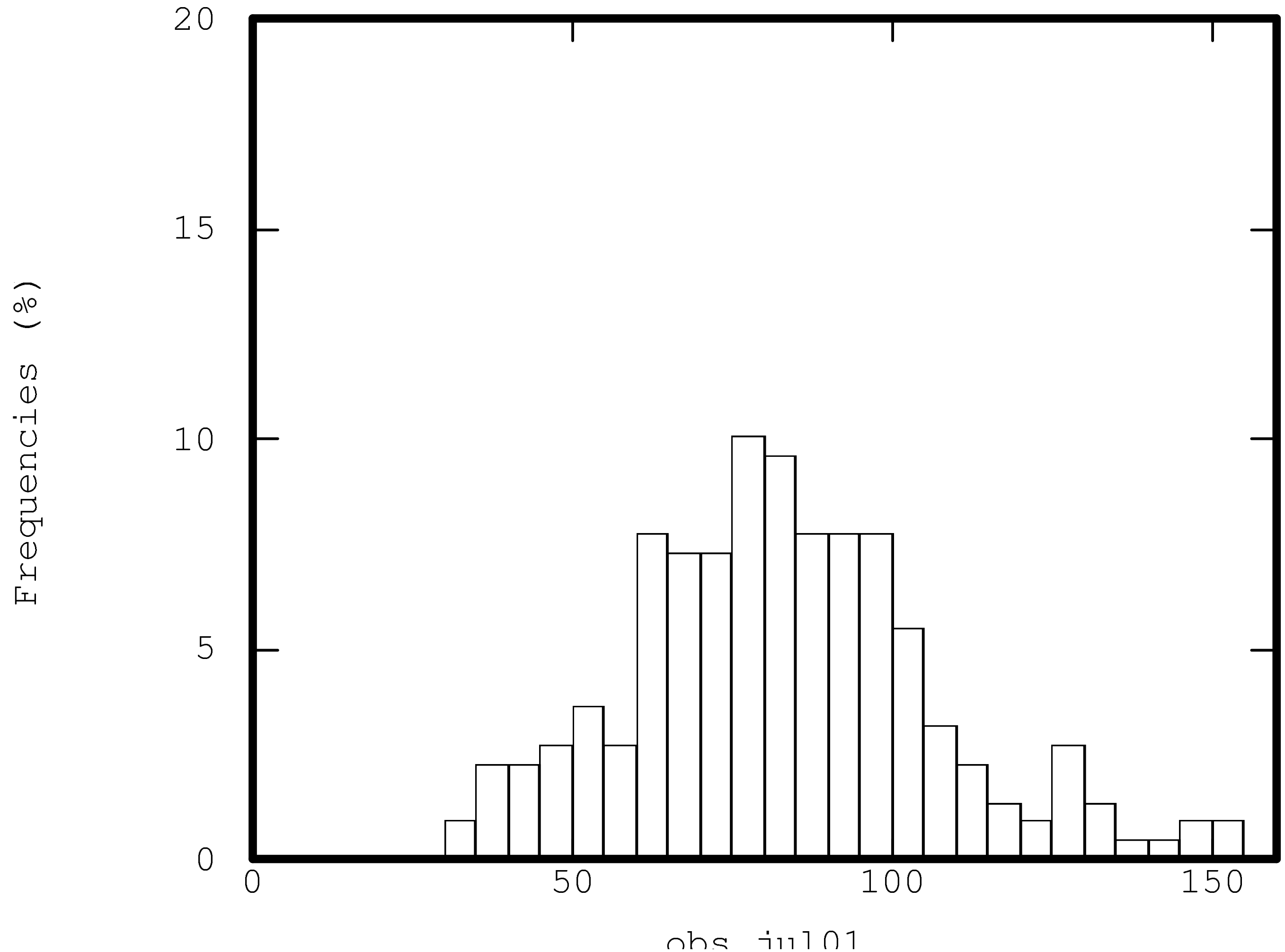


Exceedance probabilities for 1×1 km^2 support
with the European threshold of 180 μg/m^3

Precipitation in SE Norway

geostatistical downscaling

Histogram of precipitation: July 2001

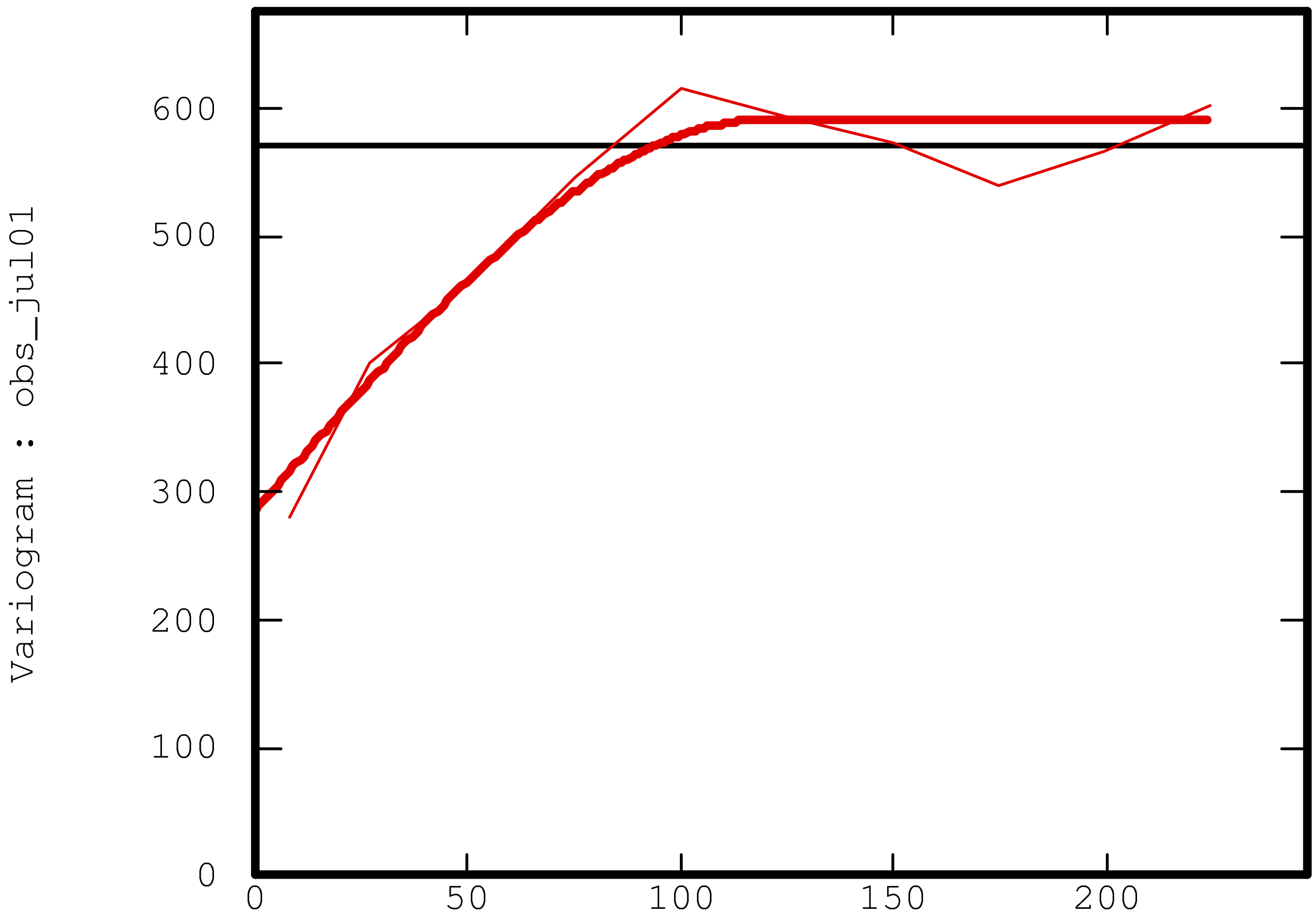


a

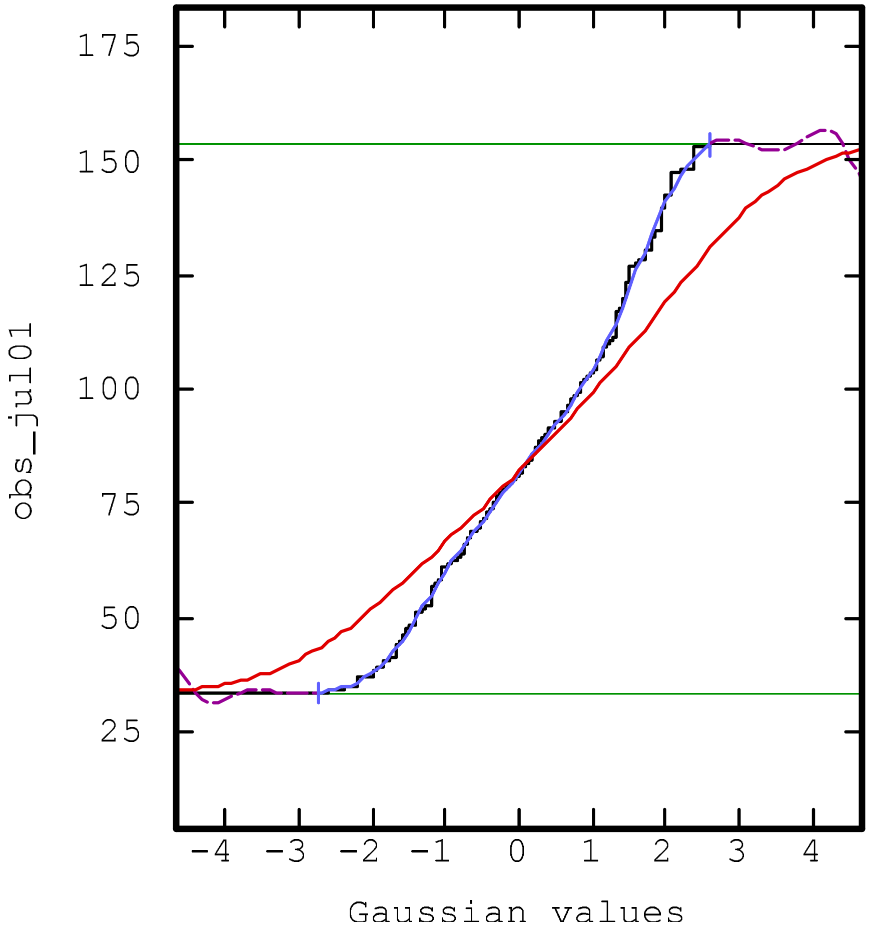
---

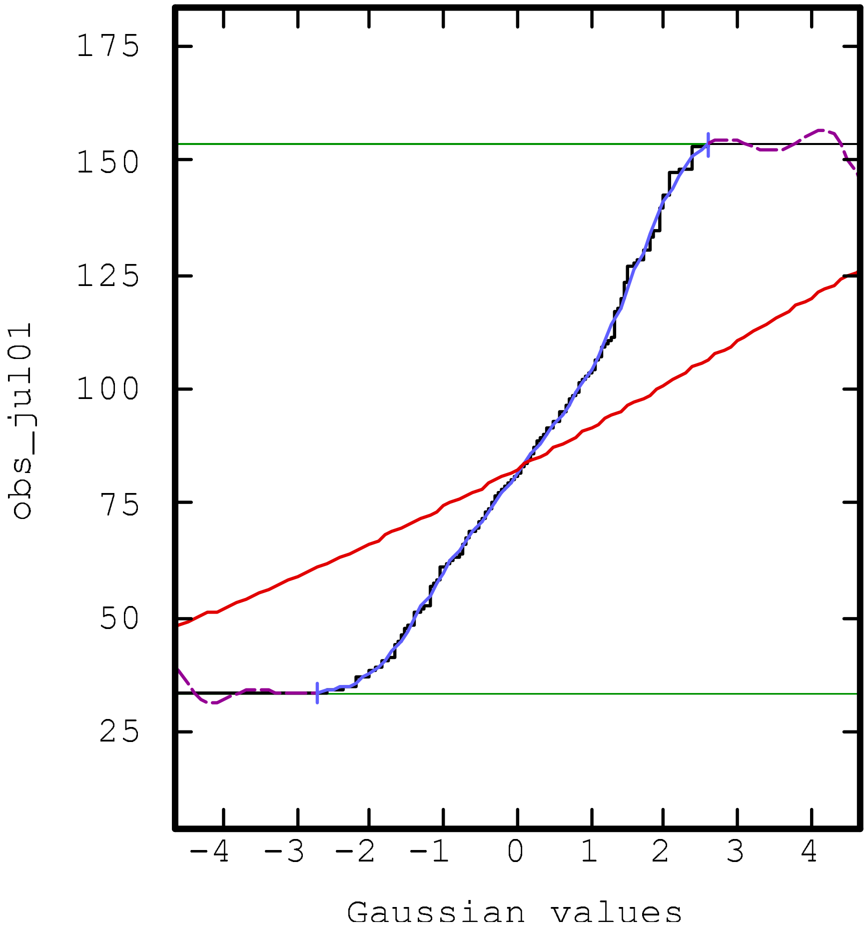
[a] Cf. Fanny Duffourg (2004).

Variogram of precipitation



Block and cell anamorphosis





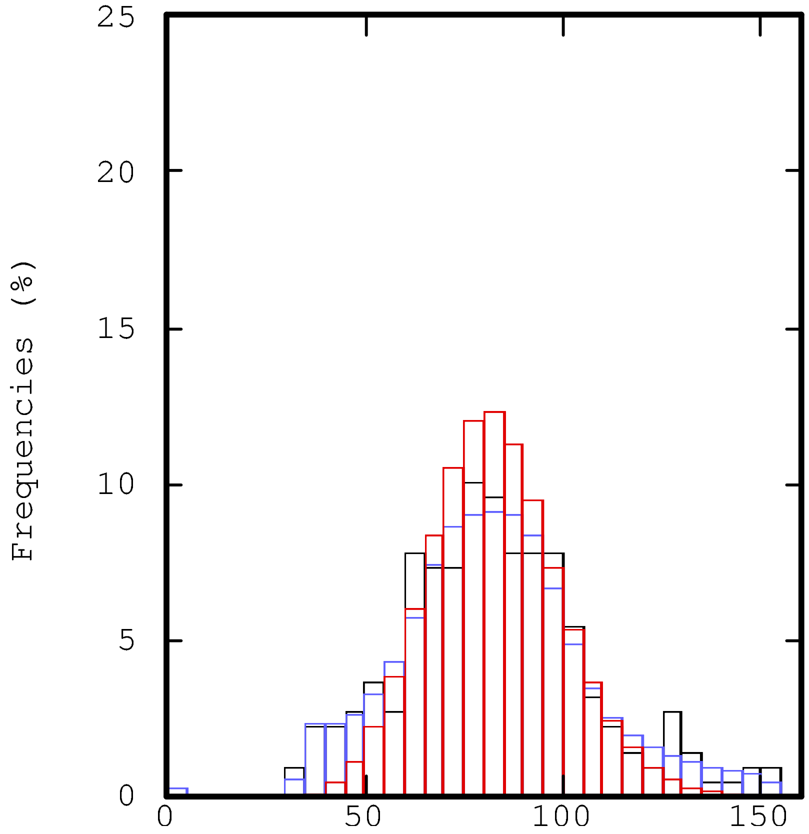
r=.7

$10 \times 10 \mathrm{km}^2$ blocks

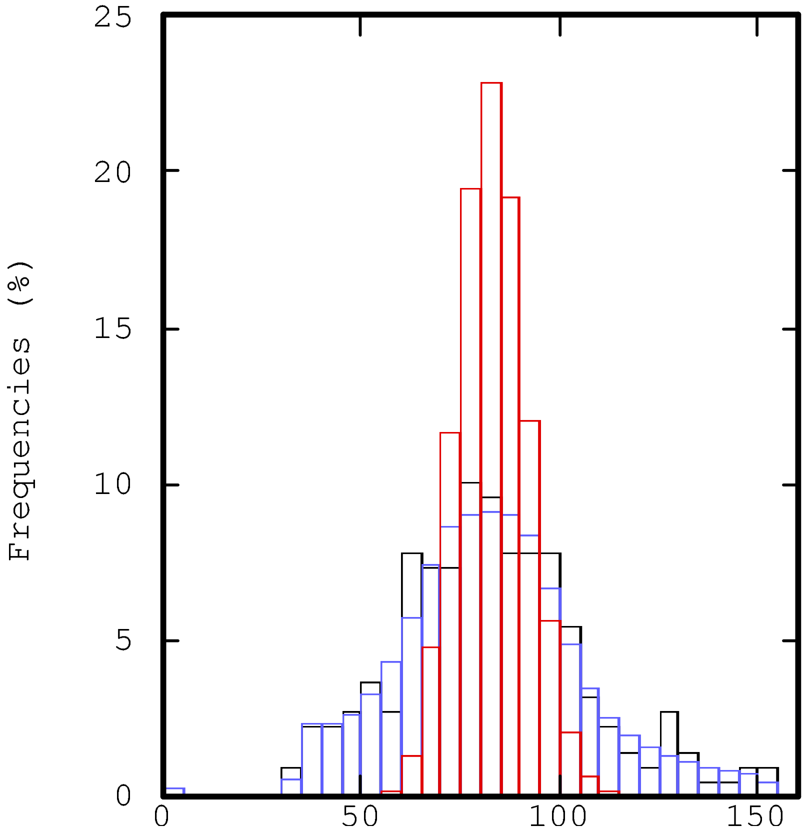
r=.365

NCEP cells

Reconstructed histograms

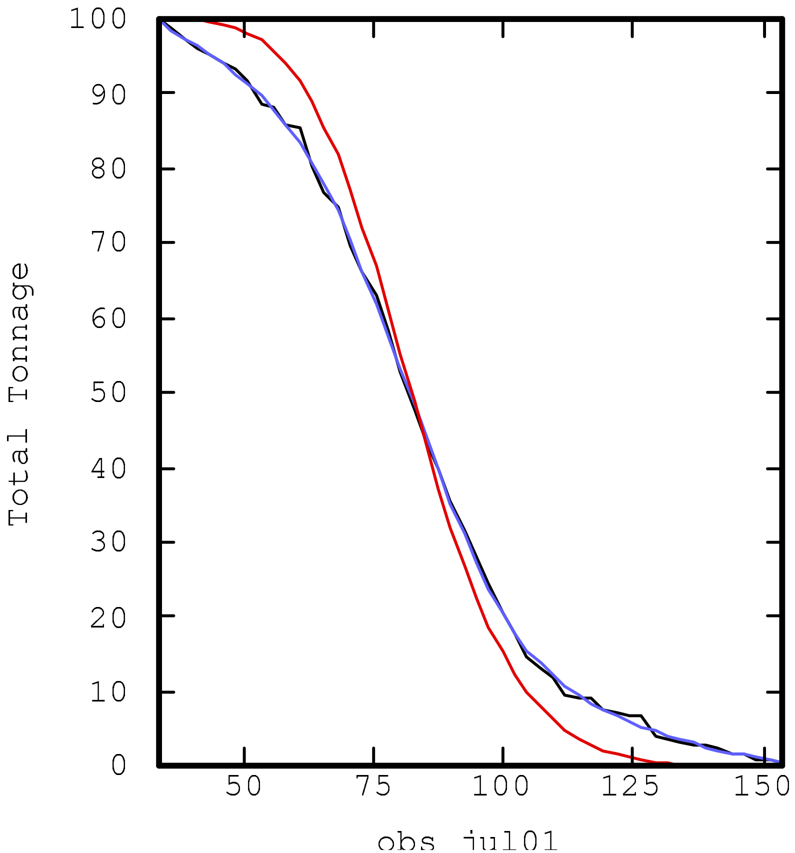


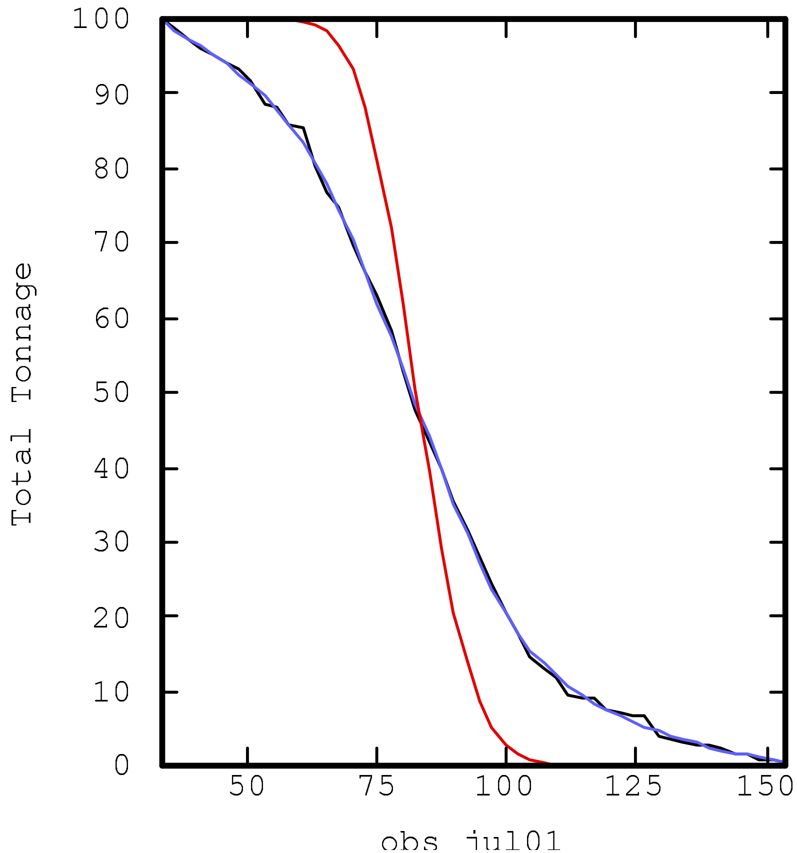
10×10 km^2 blocks



101×212km^2 NCEP cells

Proportion above threshold



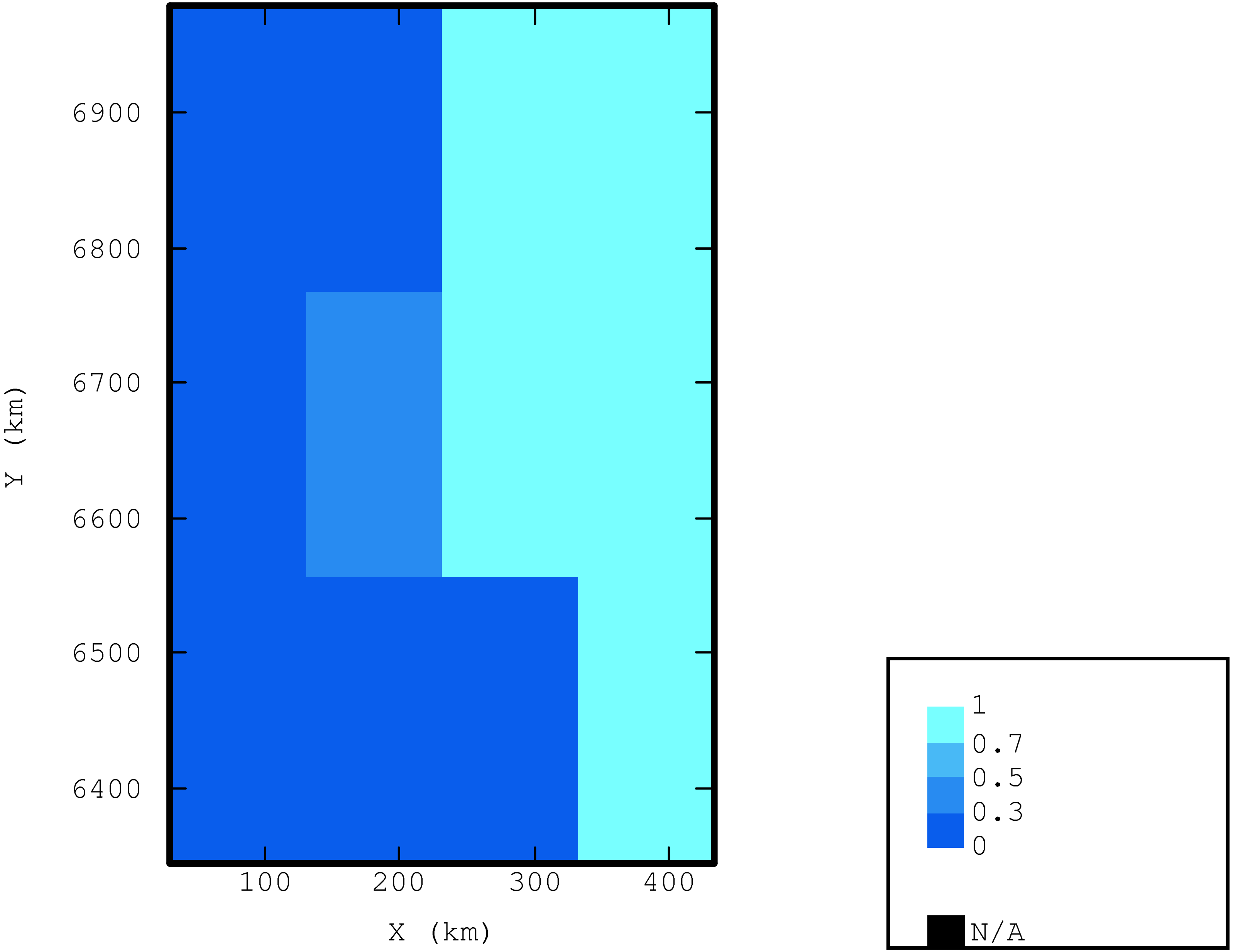


$10 \times 10\text{km}^2$ blocks

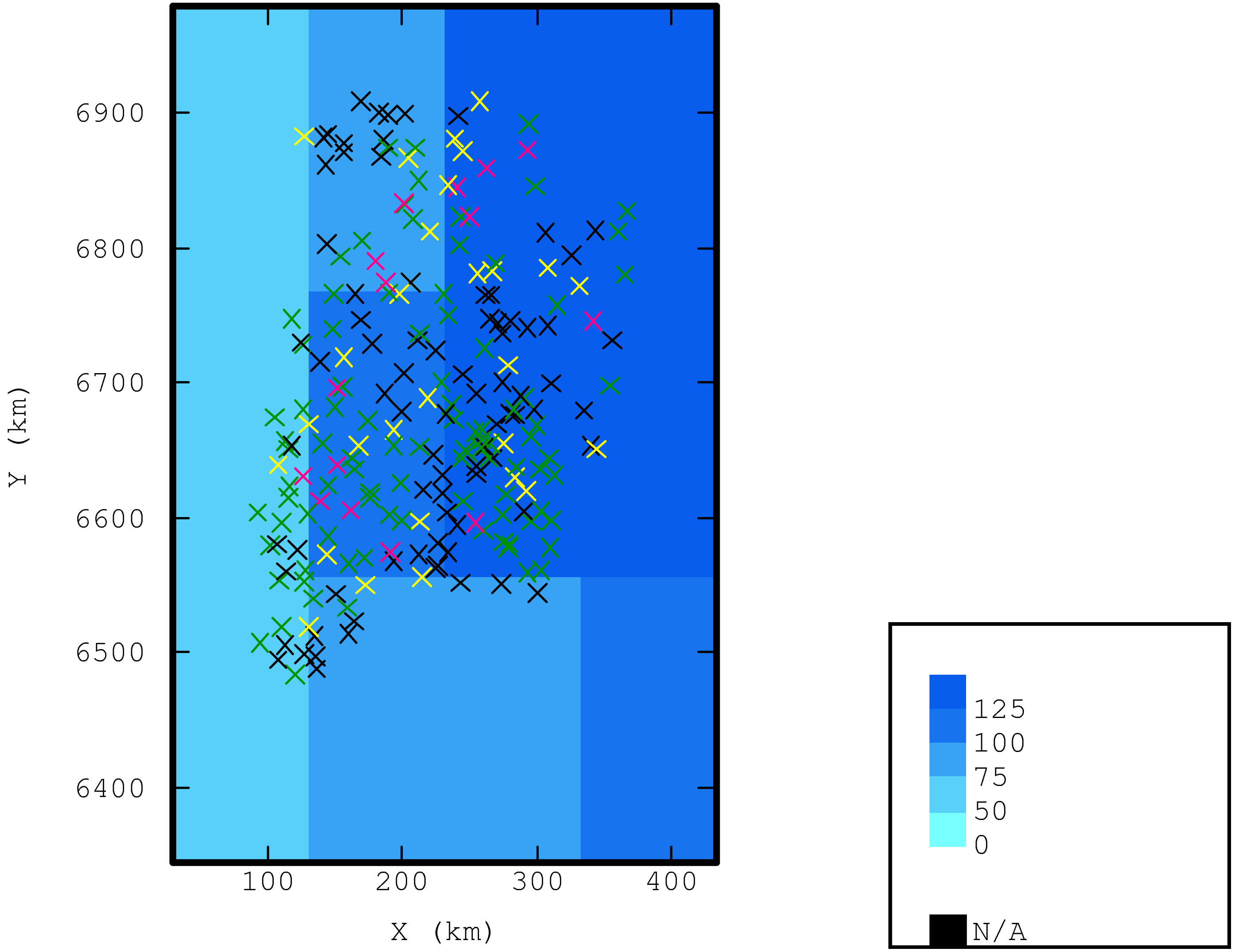
NCEP cells

A threshold of 100mm will be used

Proportion blocks >100mm within NCEP cells



NCEP cells and station values



Color codes: 0 < x < 75mm < x < 100mm < x < 125mm < x

References

[1] J. P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 1999.

[2] C. Lantuéjoul. *Geostatistical Simulation: Models and Algorithms*. Springer-Verlag, Berlin, 2002.

[3] H. Wackernagel. *Multivariate Geostatistics: an Introduction with Applications*. Springer-Verlag, Berlin, 3rd edition, 2003.