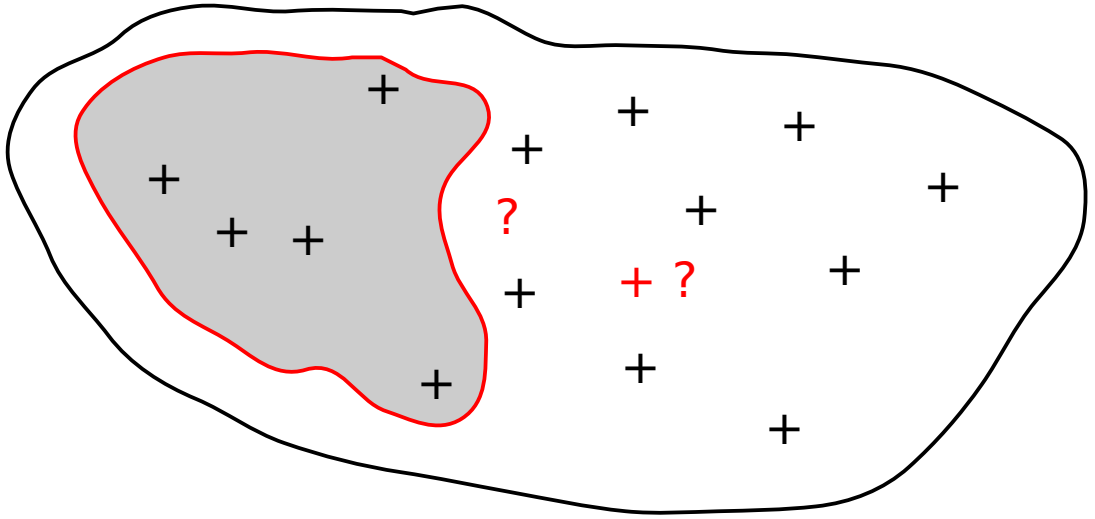
Regionalized variable Z:
- Temperature.
- Pollutant concentration.
- Content of an ore.
- ...



Main questions:

- What are the values where no measurements (interpolation / mean estimation)?
- What is the error associated with these estimates?

Geostatistical framework (hypotheses, tools, methods) → objectives of this course!

Outline:

1. Some deterministic interpolation methods

2. Variogram

3. Estimating the variogram

4. Modeling the variogram

1. Thiessen polygons

Interpolated value = value of closest measurement

$$Z(x_0) = Z(x_{i0})$$
$$i_0 = \text{argmin}_i ||x_0 - x_i||$$

Mean value over domain: $\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} \lambda_i Z(x_i)$ where $\lambda_i = \frac{a_i}{\sum_{j=1}^{n} a_j}$

2. Inverse distance weighting

Weight ~ inverse distance at a given power k

$$Z(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \quad \text{where } \lambda_i = \frac{||x_0 - x_i||^{-k}}{\sum_{j=1}^{n} ||x_0 - x_j||^{-k}}$$

3. Spline methods

Cubic spline: local fitting of 3rd order polynomials on intervals between consecutive points + continuity and differentiation at limits of intervals

There is a variety of other deterministic interpolation methods.

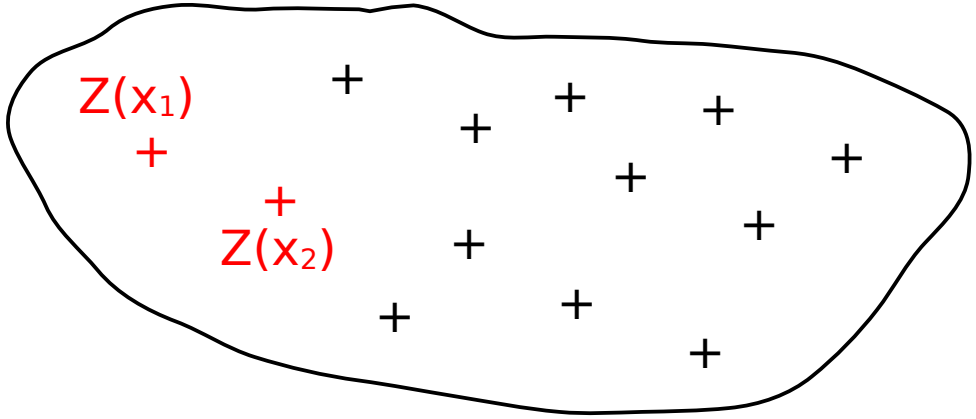
However:

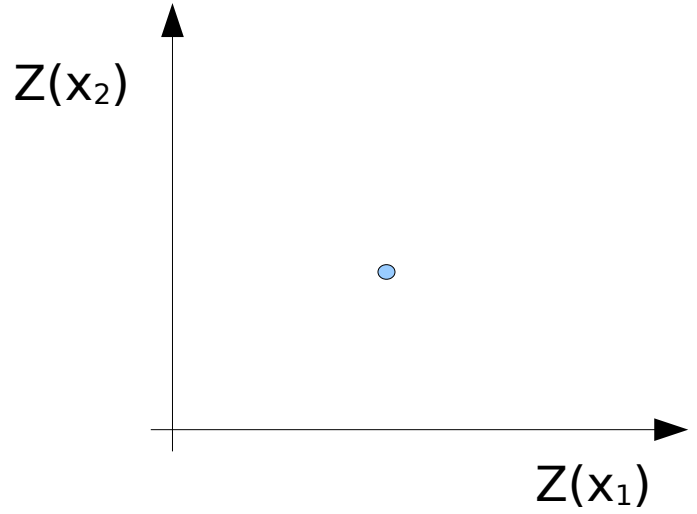
1) They do not explicitly take into account the spatial structure of the data.

2) They do not provide any information on the error associated with interpolated values.

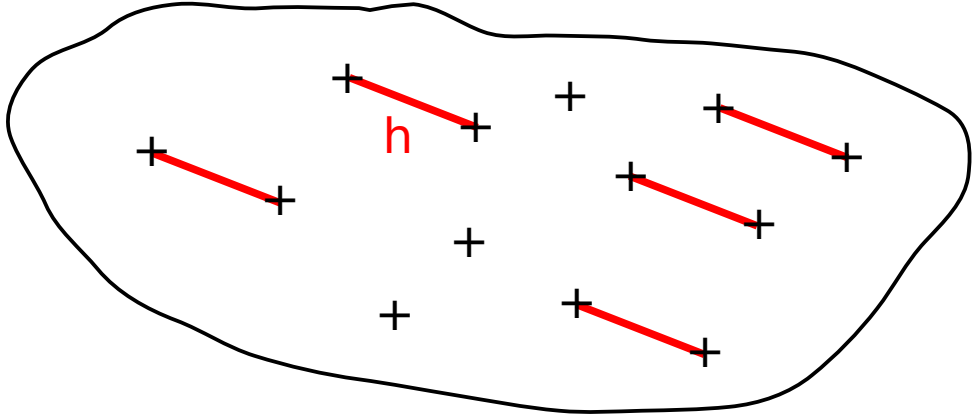
A different approach based on stochasticity has been proposed to cope with these limitations of deterministic interpolation techniques.

We will first see how to characterize the structure (i.e., the “similarity” between neighboring points) in spatial data.

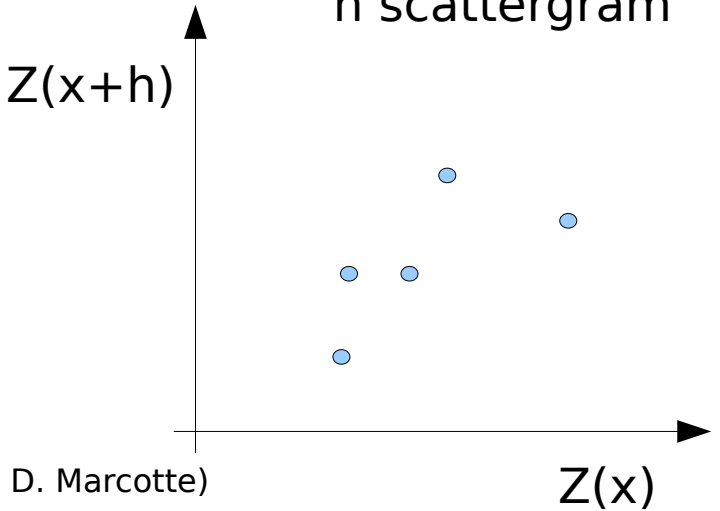




Stationarity

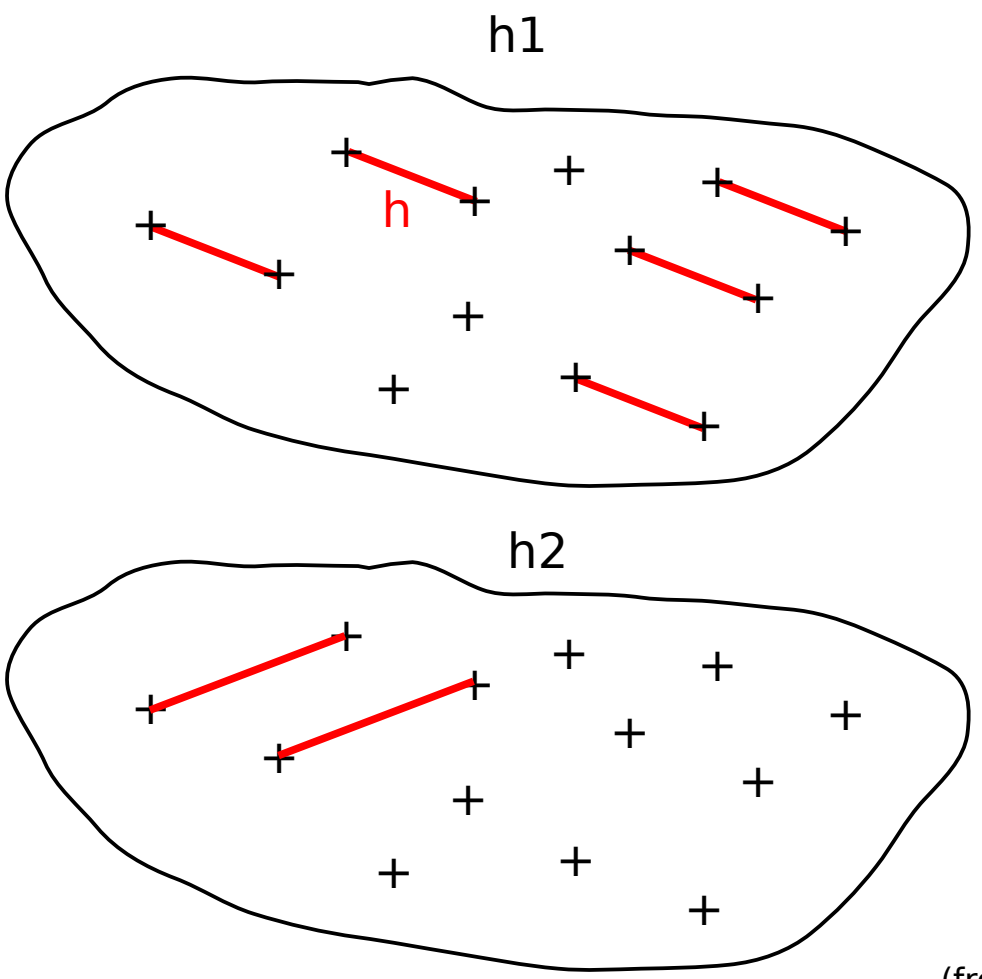


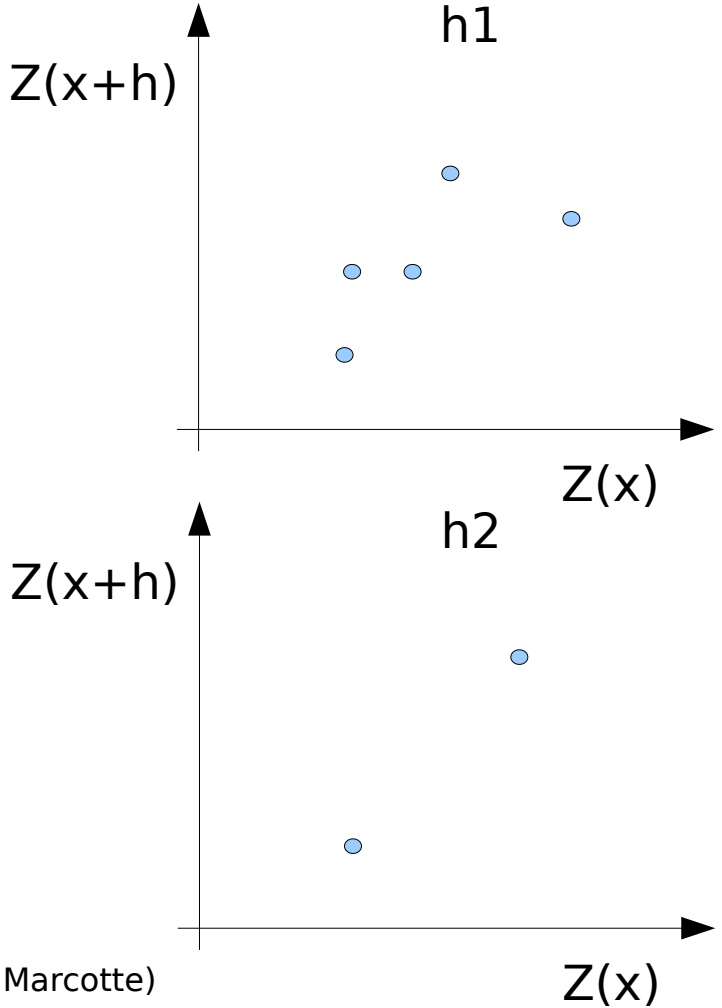
“h scattergram”



(from D. Marcotte)

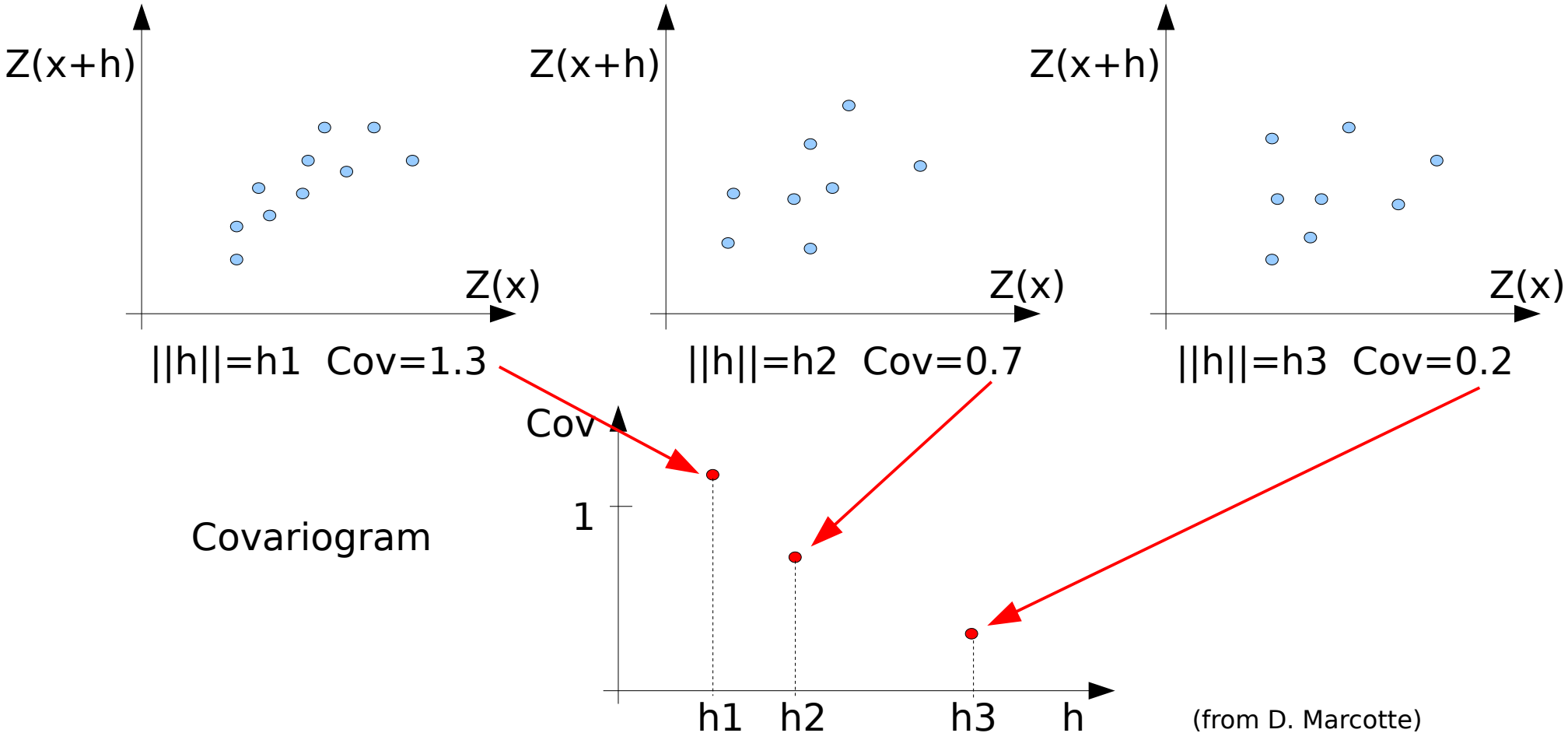
h can vary in length and direction





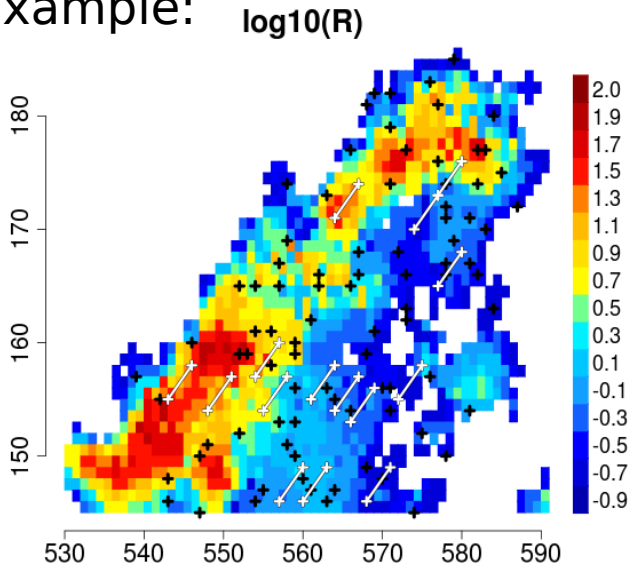
(from D. Marcotte)

h scattergram too much information → summarize these graphs
Use the covariance between Z(x+h) and Z(x): correlation + variability.

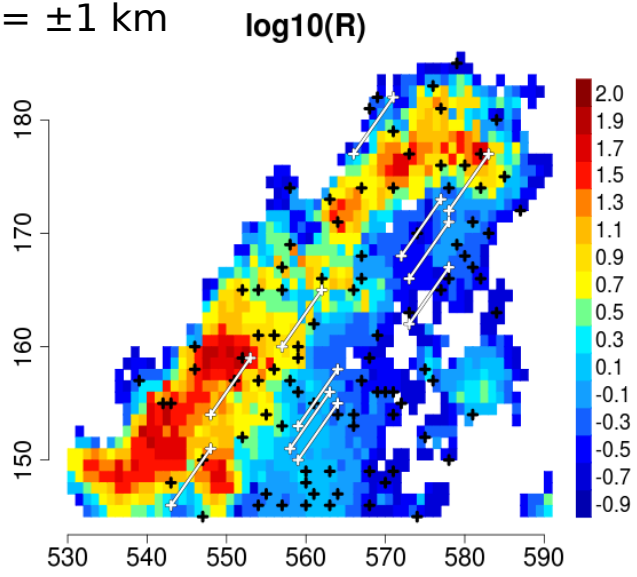


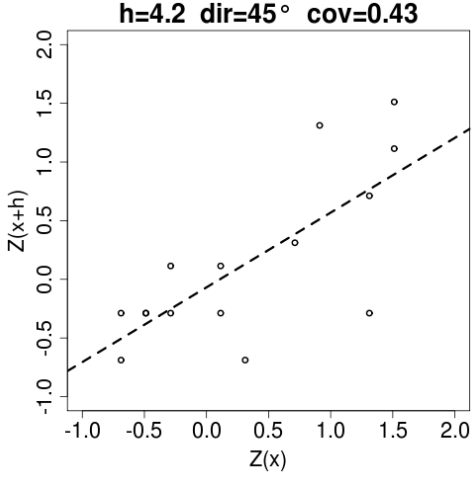
(from D. Marcotte)

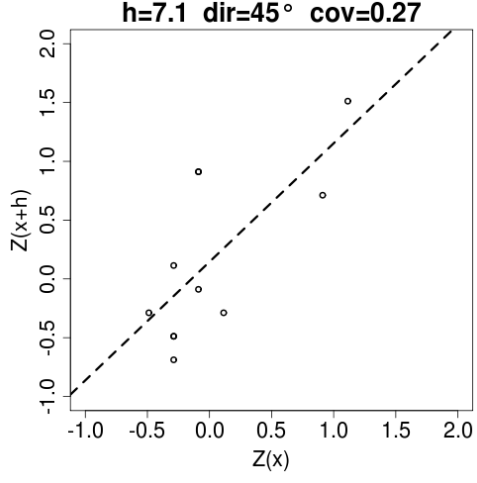
Example:



tol. distance = ±1 km







Covariance is sensitive to sampling effects, in particular because the mean is explicitly involved.

This issue is especially important in case of mono-realization RF.

Another tool has been proposed by Matheron to cope with this difficulty: the (semi-)variogram

It is defined as half of the variance of the increments of a RF:

$$\gamma(h) = \frac{1}{2}\mathrm{Var}\left\{Z(x+h) - Z(x)\right\}$$

As the increments are involved, the variogram is not sensitive to the uncertainty affecting the mean (of Z) estimated from the sample, while the covariance is.

For γ to be defined, RF does not need to be SRF but only an intrinsic random function (IRF).

1st order moment:

$$\mathrm{E}[Z(x_1+h)-Z(x_1)]=\mathrm{E}[Z(x_2+h)-Z(x_2)] \quad \forall(x_1,x_2)\in D^2$$

$$\Rightarrow \mathrm{E}[Z(x+h)-Z(x)]=\alpha h \quad , \ \alpha=\mathrm{cst}$$

1st-order stationarity of the increments implies a linear drift in the RF!

In the following, we assume that the linear drift has been corrected, hence

$$\mathrm{E}[Z(x_1)]=\mathrm{E}[Z(x_2)] \quad \forall(x_1,x_2)\in D^2$$

2nd order moment:

$$\mathrm{Var}\left[Z(x_1+h)-Z(x_1)\right]=\mathrm{Var}\left[Z(x_2+h)-Z(x_2)\right]\forall(x_1,x_2)\in D^2$$

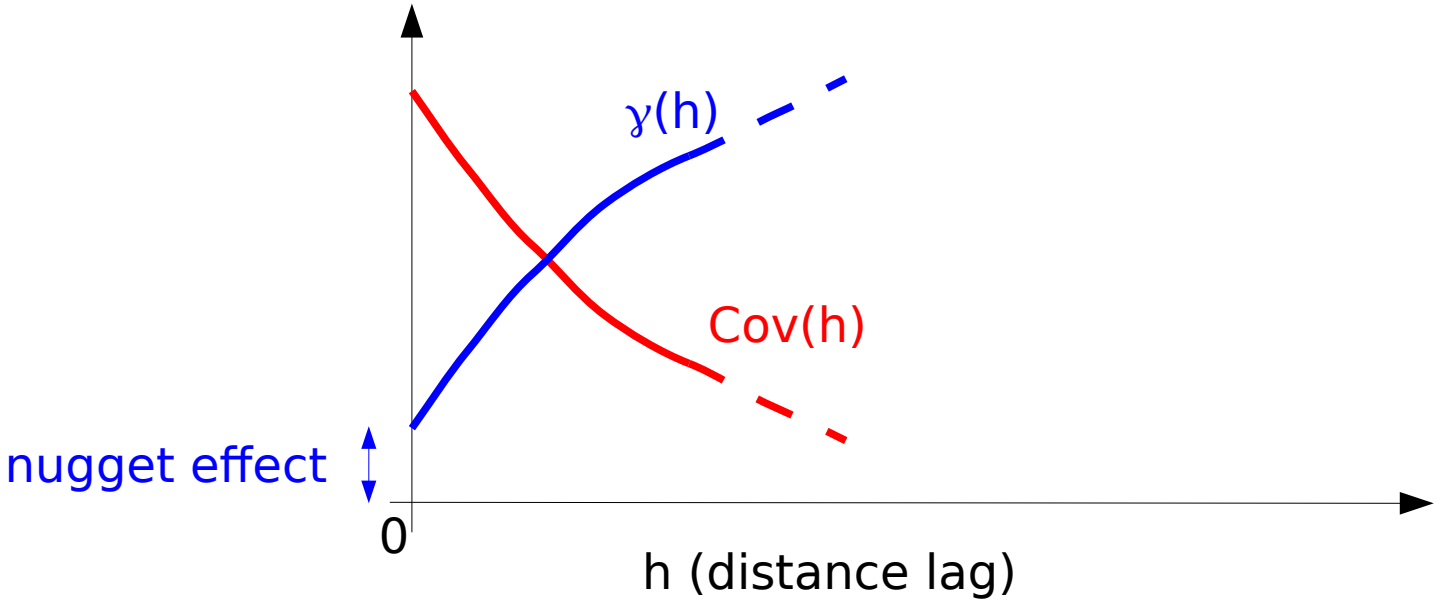
From previous assumption:

$$\begin{aligned}\gamma(h) &= \frac{1}{2}\mathrm{Var}\left[Z(x+h)-Z(x)\right]\\ &= \frac{1}{2}\mathrm{E}\left[\left(Z(x+h)-Z(x)\right)^2\right]\end{aligned}$$

Because the class of IRF includes the class of SRF, the variogram is a more general tool than the covariance.

If Z is a SRF, the link between the covariance and the variogram is

$$\mathrm{Cov}[Z(x+h),Z(x)]=\mathrm{Var}[Z]-\gamma(h)$$

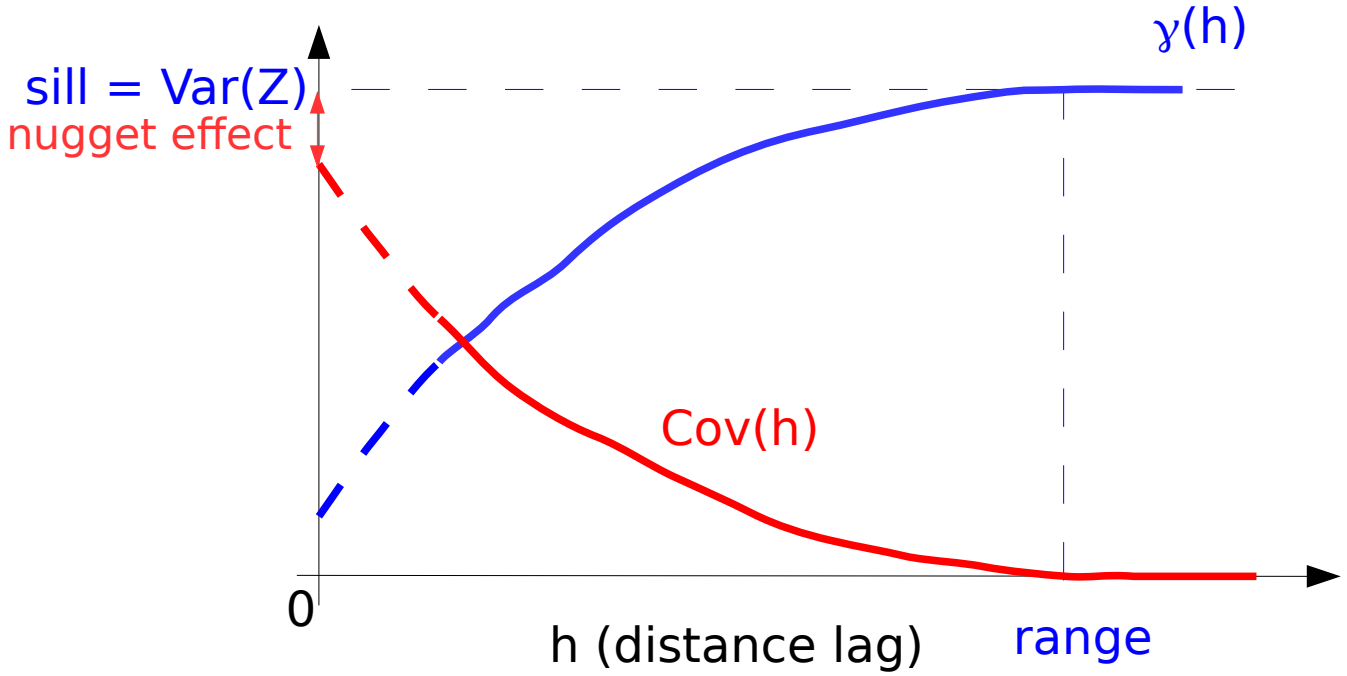


Physical interpretation of the variogram at “short” distance lags

Variogram characterized by:

- nugget effect (name from mining): possible discontinuity at h=0, can be null.

- Slope: rate of increase of γ reflects dissimilarity at increasing distance lags.



Physical interpretation of the variogram at “long” distance lags

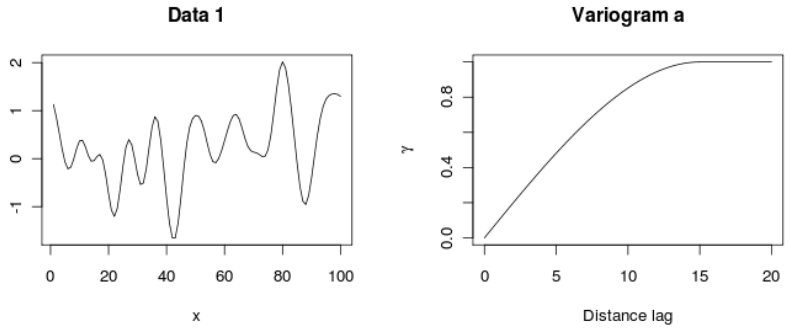
If Z 2nd order stationary:

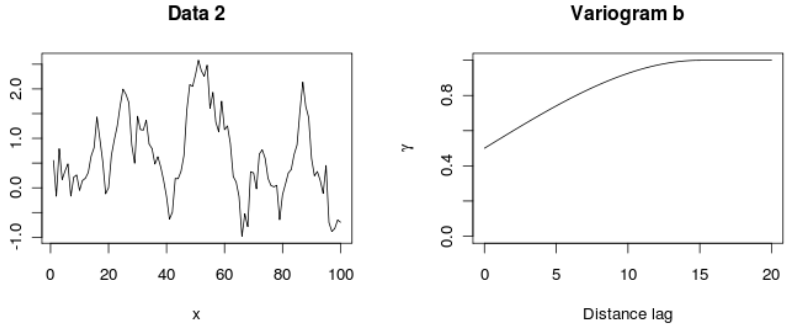
- Range: decorrelation distance → area of influence (if it exists).

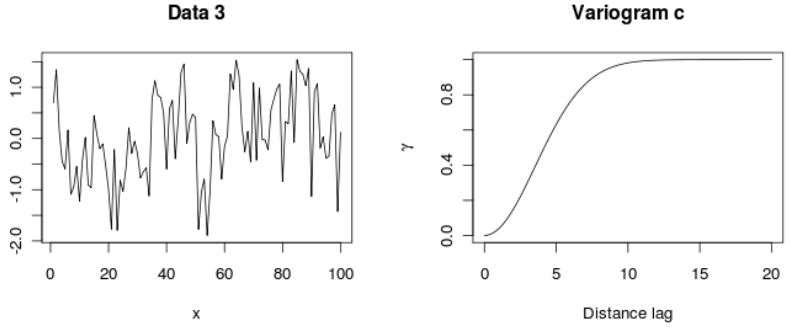
- Sill: related to the variance of the RF (if it exists).

The variogram quantifies the spatial structure (variability + continuity) of the studied RF

Which vario corresponds to which time series?

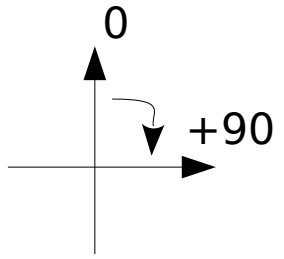


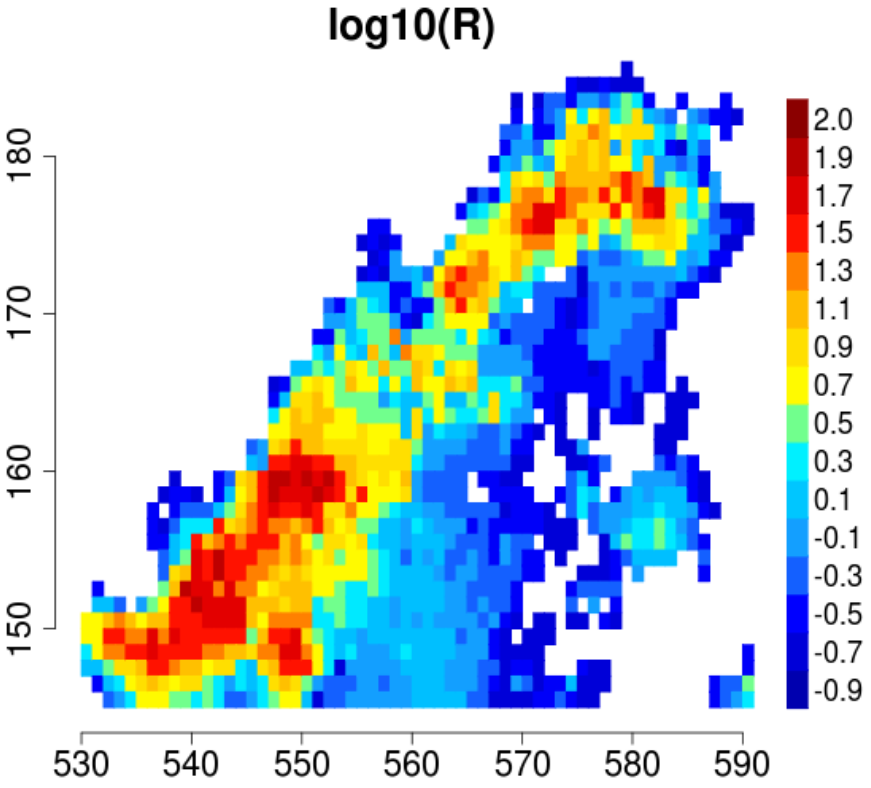


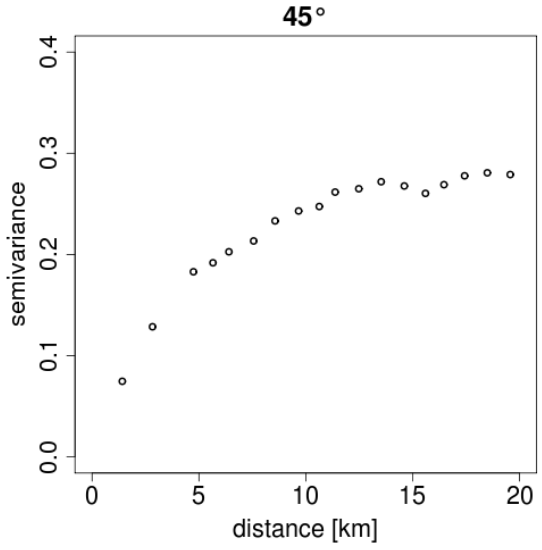


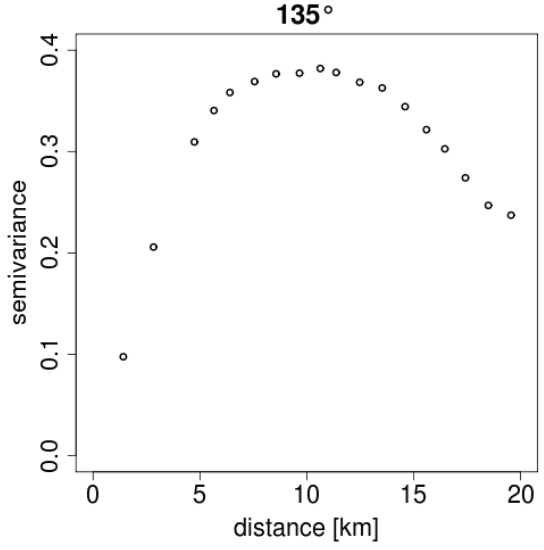
In 2D (or more), in addition to quantification of variability, variogram also a useful tool to analyze anisotropy.

→ directional variograms reveal difference in structure / direction.







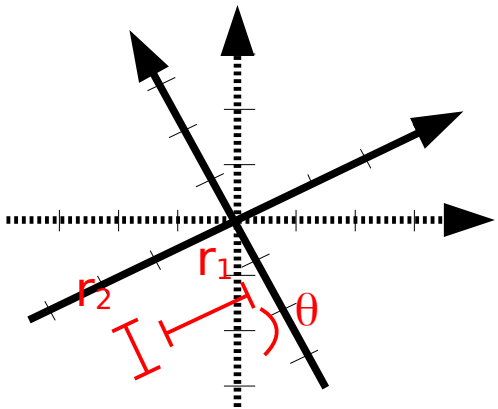


Geometric anisotropy

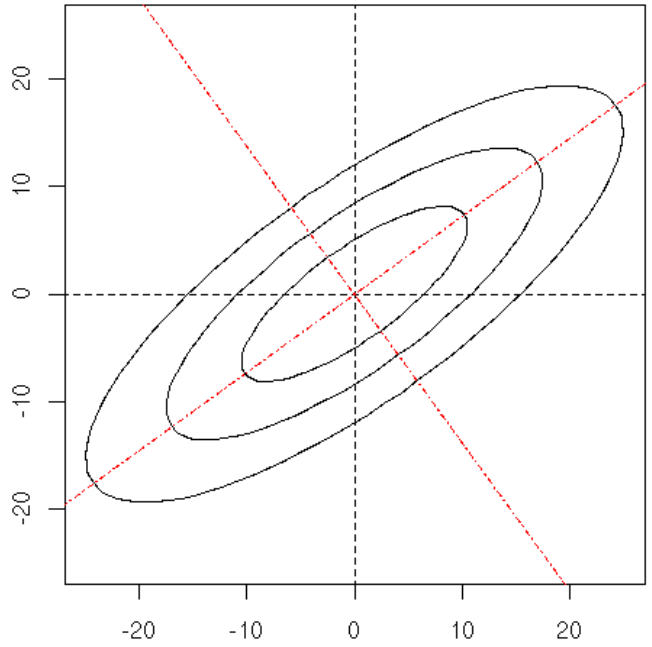
Through a linear change of coordinates, variation of Z becomes isotropic.

Transformation matrix

$$\mathbf{A} = \begin{pmatrix} r_1 & 0 \\ 0 & r_2 \end{pmatrix} \begin{pmatrix} \cos\theta_0 & \sin\theta_0 \\ -\sin\theta_0 & \cos\theta_0 \end{pmatrix}$$



Variogram exhibits elliptic isovalue contours.



$$\gamma(\mathbf{h}) = \gamma_0(|\mathbf{Ah}|) \quad \text{where} \quad \mathbf{h} = \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}$$

In this sense, the variogram is a relevant tool for the structural analysis of a spatial process (structure, anisotropy, variability).

The variogram is a 2nd order moment. But it is not enough to completely describe a RF and its variability.

2 RF can have identical 1st and 2nd order moments and still be very different.

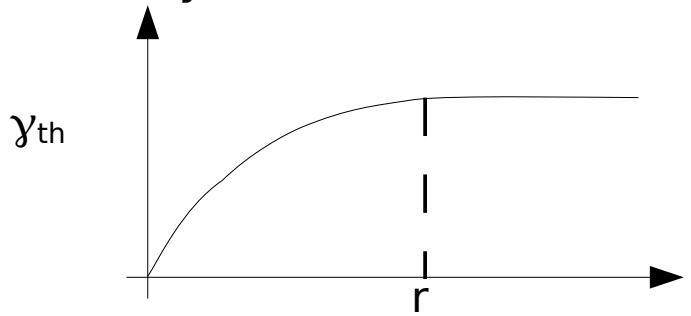
"The variogram is the corner stone of geostatistics, and it is therefore vital to estimate it and model it correctly."
Geostatistics for Environmental Scientists, R. Webster and M. Oliver.

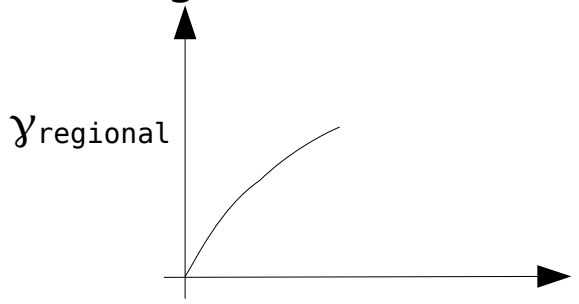
Previous definitions describe the variogram of the random function, called theoretical variogram. It corresponds to the population of possible values.

We consider a regionalized variable (i.e., realization of the studied RF over a given finite domain)
→ variogram over this particular region = regional variogram.

Usually regional variogram ≠ theoretical variogram!

Ex. of stationary RF but domain D smaller than range of theoretical variogram





Focus on RF over studied domain → focus on regional variogram to analyze and interpolate regionalized variable on this domain.

In practice, only access to a sample of 1 (or more) realization(s) of RF over studied domain.

Expectation is estimated as arithmetic mean: $$\mathrm{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

→ sample or experimental variogram (assuming IRF), also called Matheron estimator

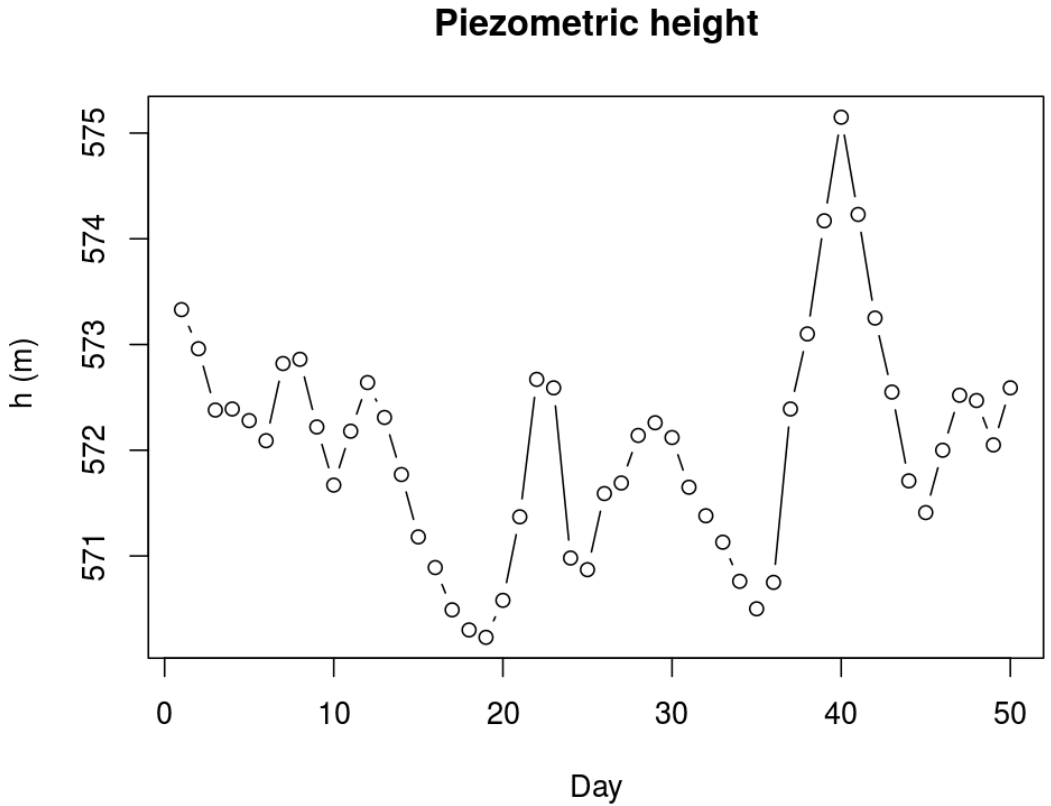
$$\widehat{\gamma}(h) = \frac{1}{2n_h}\sum_{i,j\in S_h}[Z(x_i) - Z(x_j)]^2$$

To obtain representative variogram estimates at each distance lag

$$n_h \geq 20 - 30$$

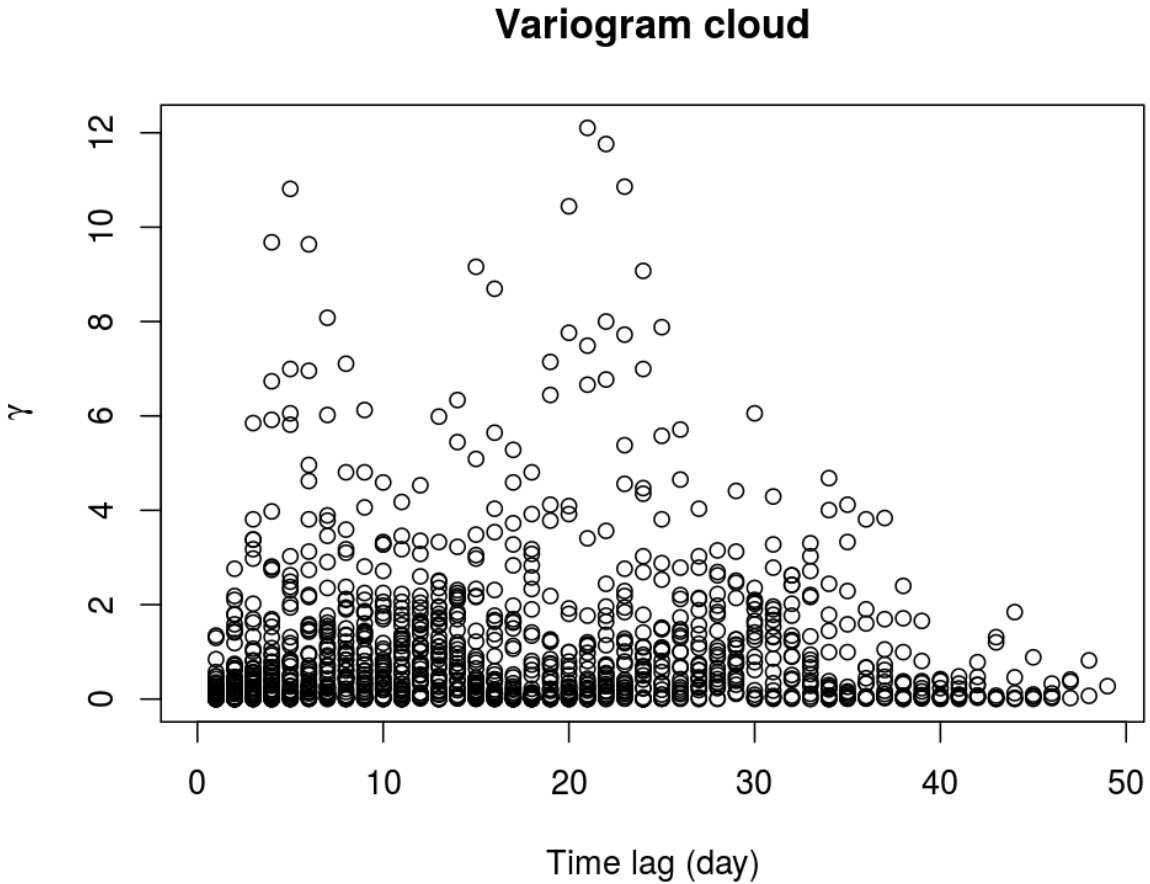
Example of 1D data:

Piezometric height from a well on a waste site (Bioley – Orjulaz, CH) during 50 days.

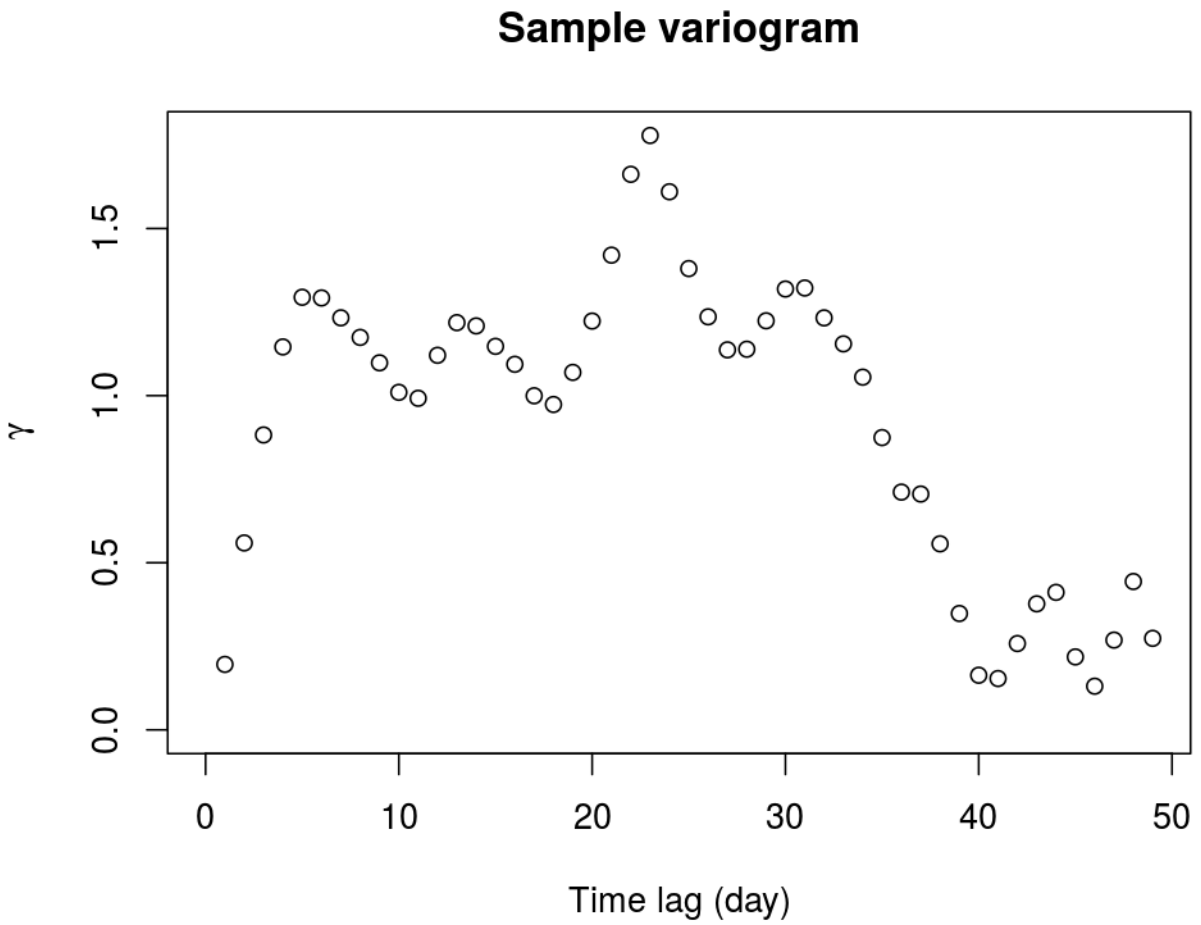


γ values can be estimated at time lags = k time resolution (1 day)

→ variogram cloud: all estimates of γ from data

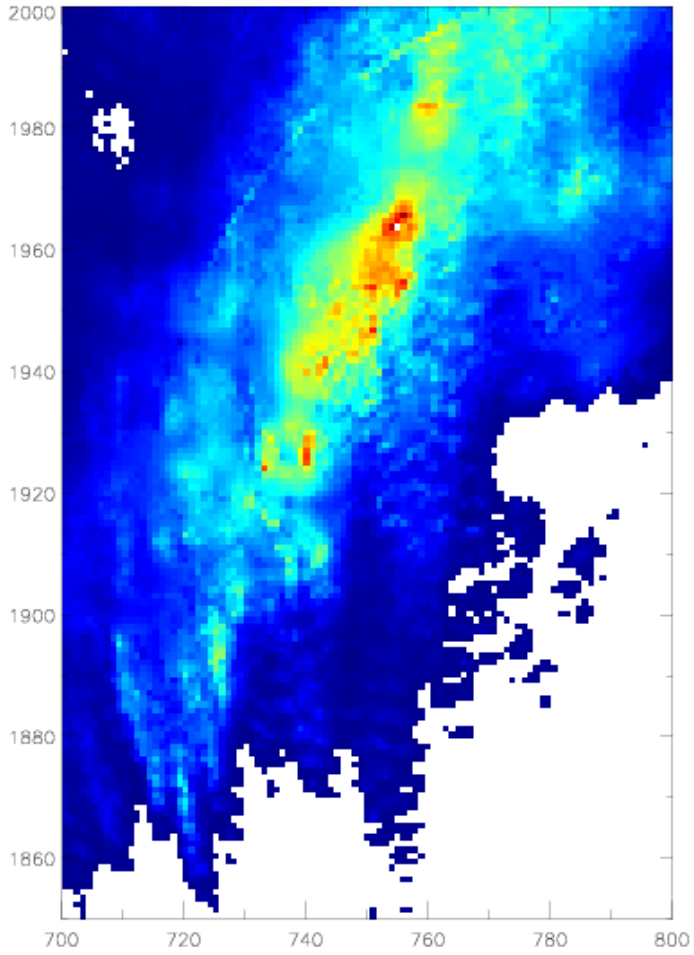


Sample variogram: mean of γ values per time lag classes (ex: 1 day)

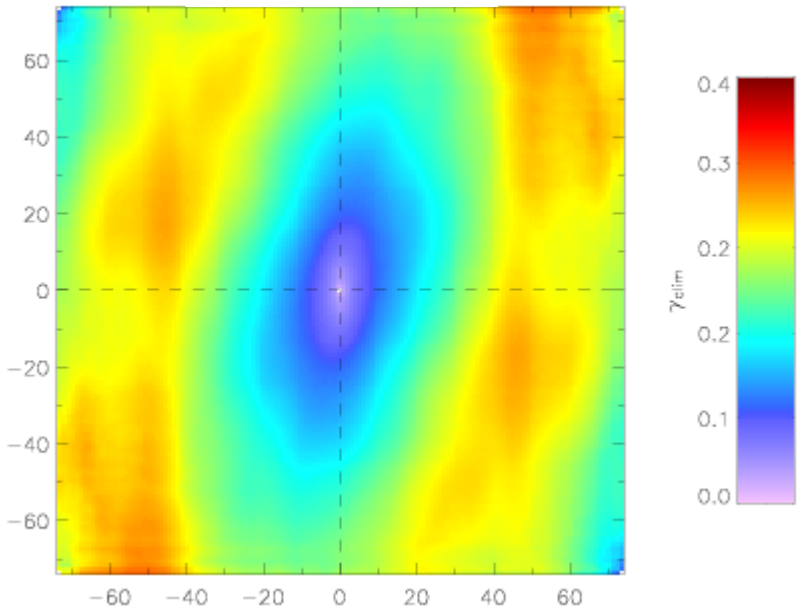


Example in 2D: 2h rainfall amount

Rainfall intensity

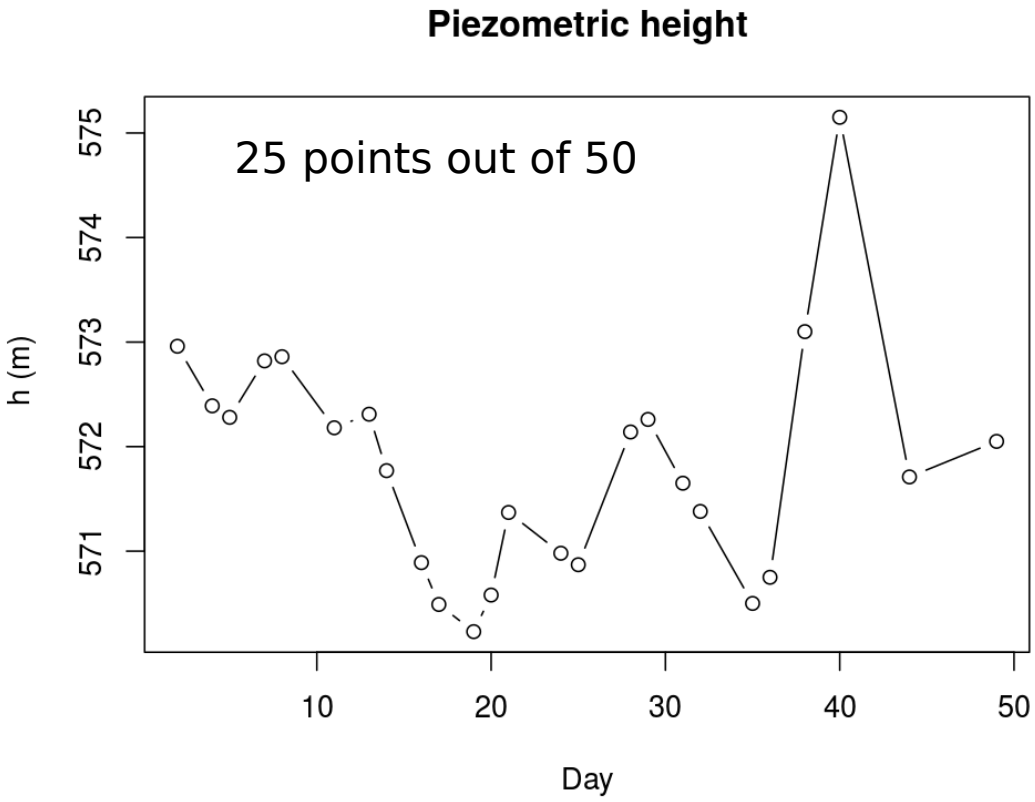


2D variogram



When measurements irregularly sampled
When only a few measurements

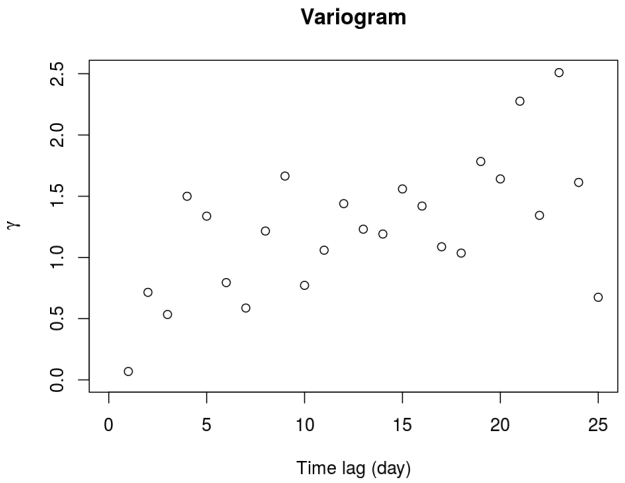
→ too few γ estimates per h

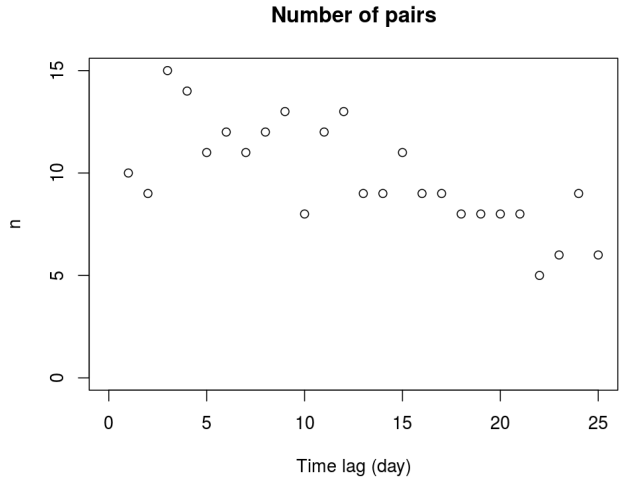


→ tolerance on distance lag h in order to increase number of points per class.

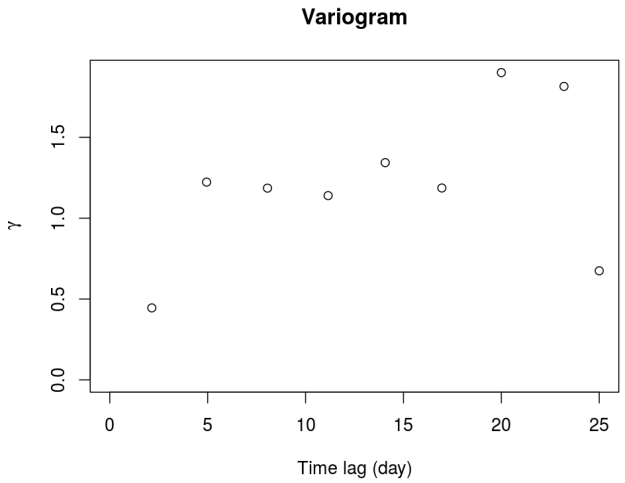
→ more representative variogram estimates.

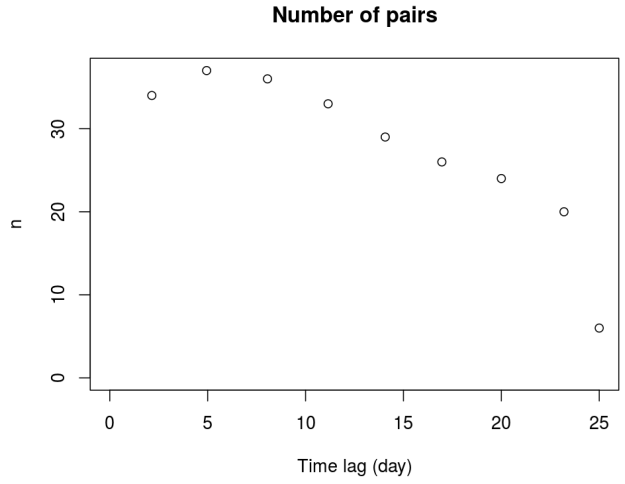
Class width = 1 day





Class width = 3 days

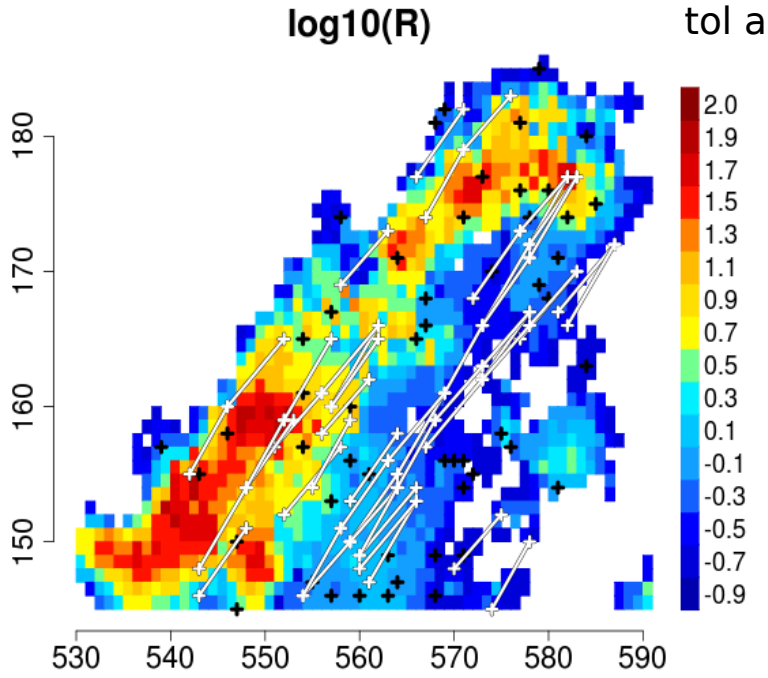


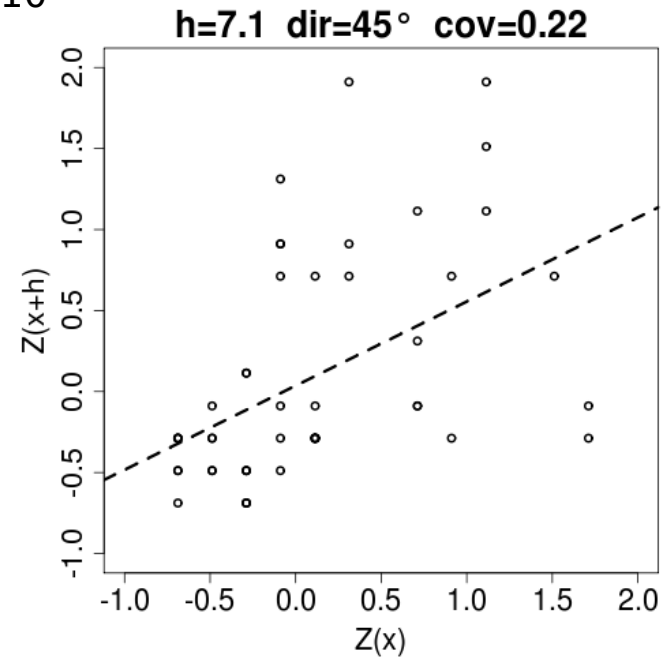


In 2D (or more)

Similar approach with class of h per distance and direction.

tol. distance = ±1 km
tol angle = ±10 °





If enough measurements → sample variogram per class of direction.
→ analysis of possible anisotropy.

Questions

- Why is the sample variogram usually different from the theoretical one?
- What is the minimum number of values per class to obtain a reliable estimate of the sample variogram?
- How can you detect anisotropy with the variogram?
- For irregular sampling, how can you increase the number of values per class?

Sampling effects

Sample variogram = mean of γ values at given distance lags.

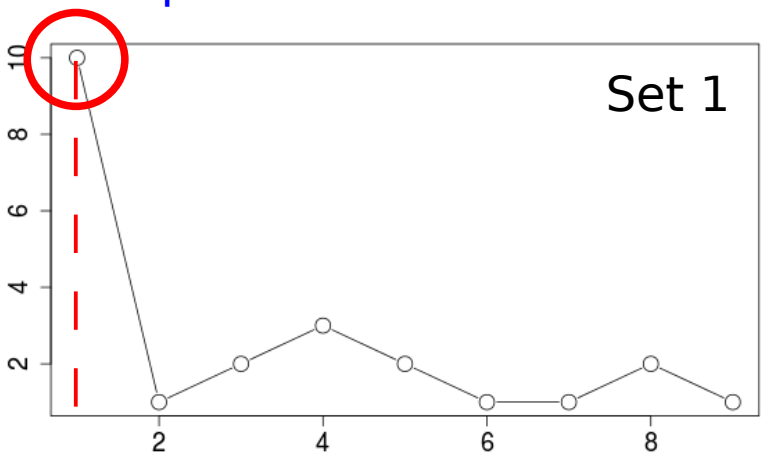
$$\widehat{\gamma}(h) = \frac{1}{2n_h} \sum_{i,j \in S_h} [Z(x_i) - Z(x_j)]^2$$

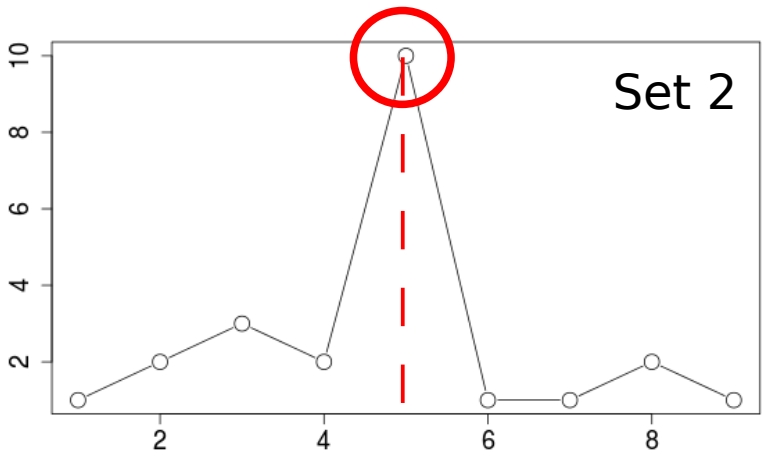
Arithmetic mean is not very representative of distribution of γ values for small samples or when extreme values in the data set.

→ sample variogram sensitive to
- "outliers": very large or very small values compared to the others
- position of these outliers within the studied domain:
  - impact is ≠: vario ↗ if outlier on the edge, ↘ if in the middle.
  - influence will be larger if outlier in the middle vs on the edge.

Variogram cloud and sample variogram

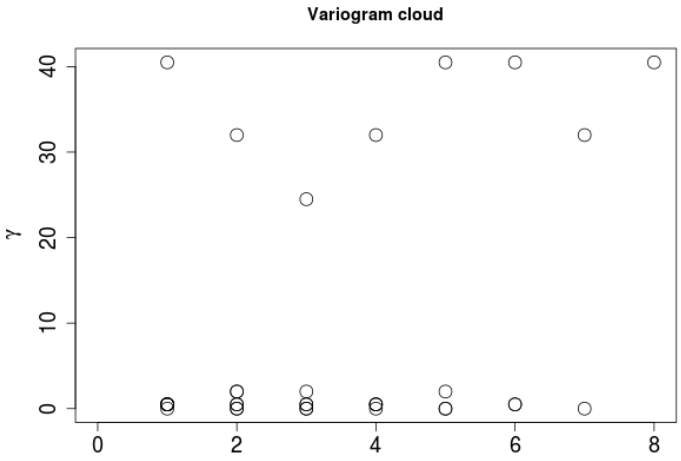
Example:

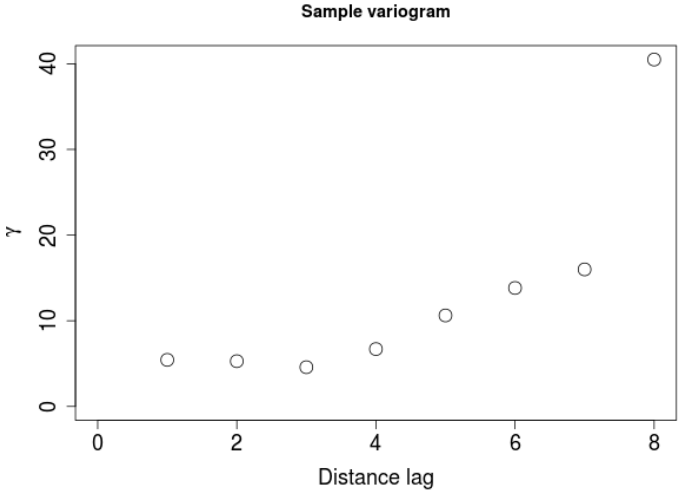




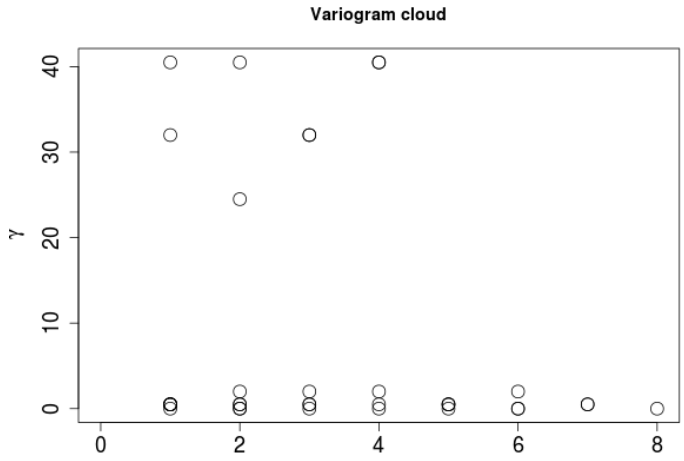
Variogram cloud and sample variogram

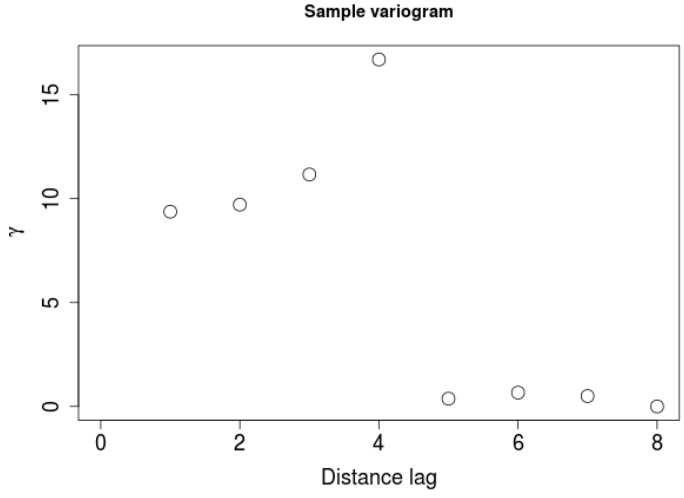
Set 1





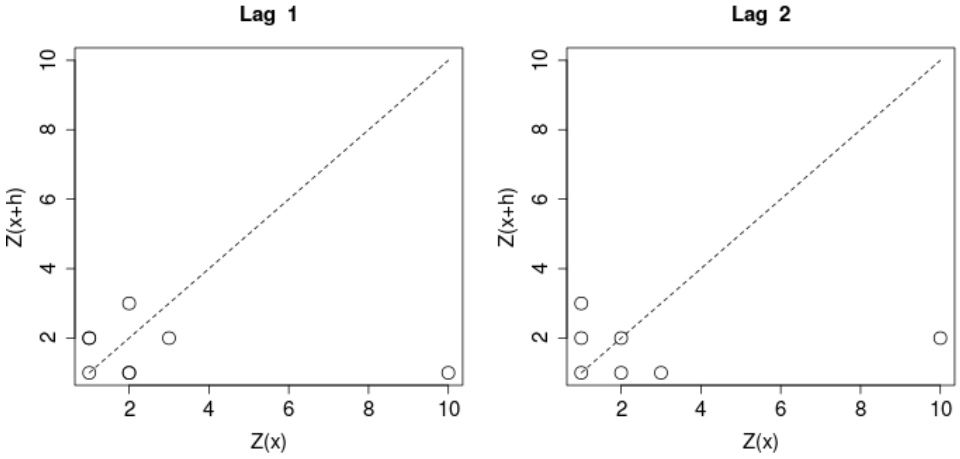
Set 2

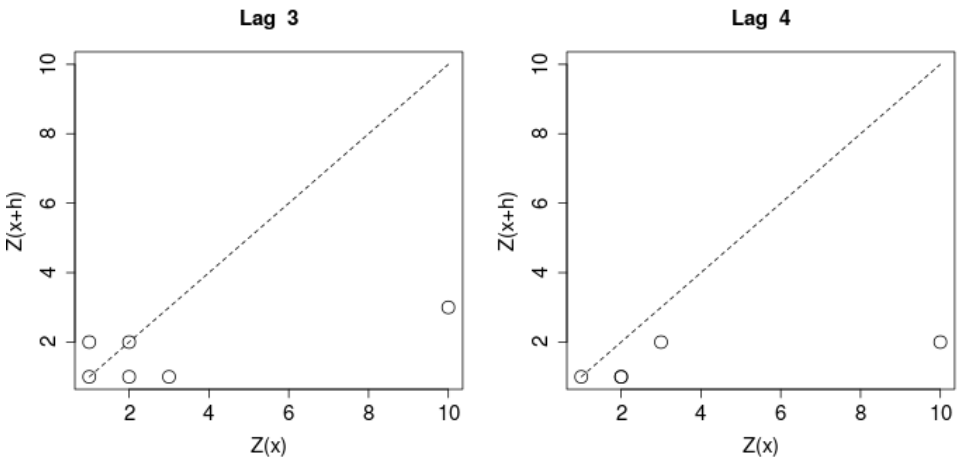




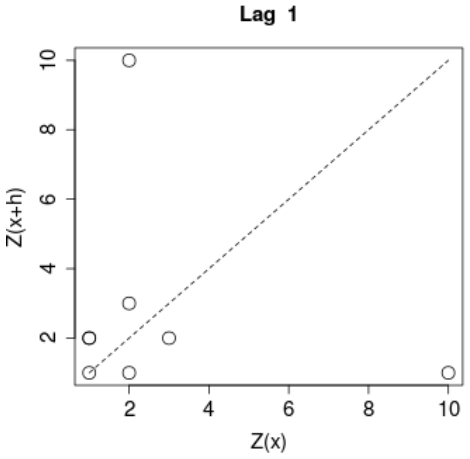
h-scattergram

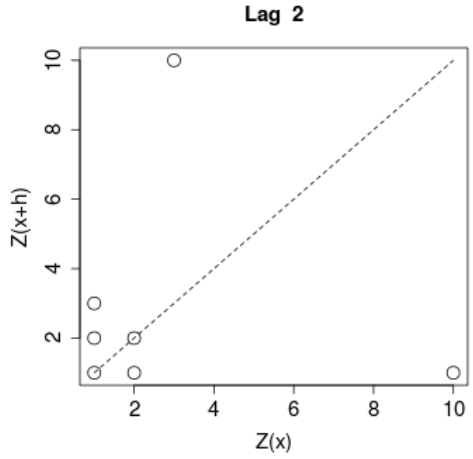
Set 1

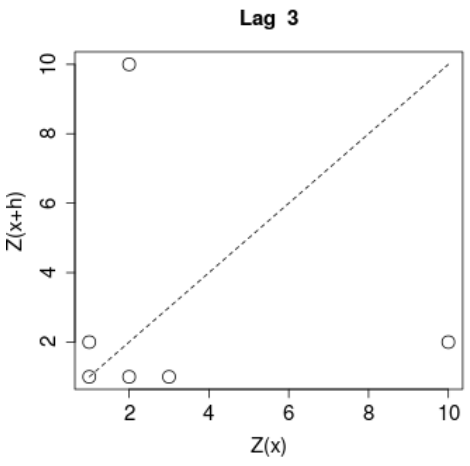


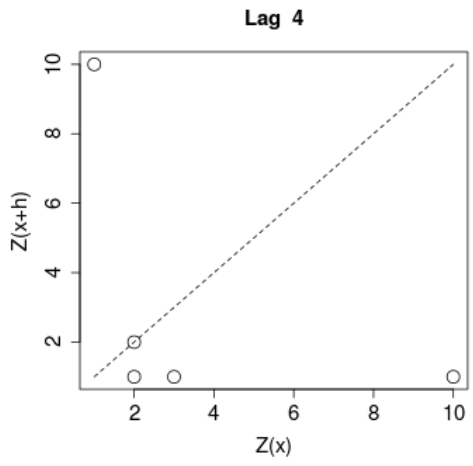


Set 2



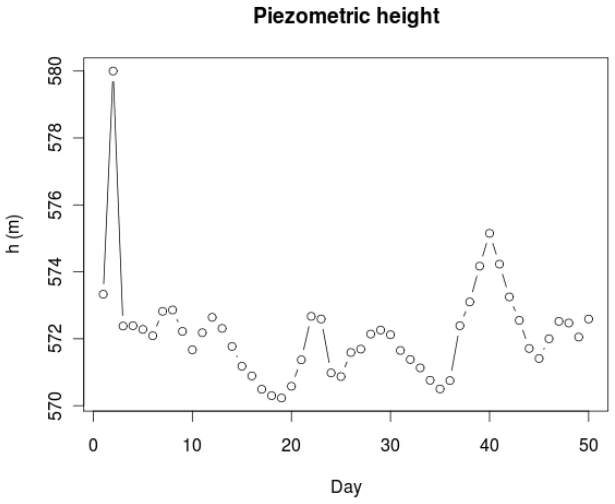


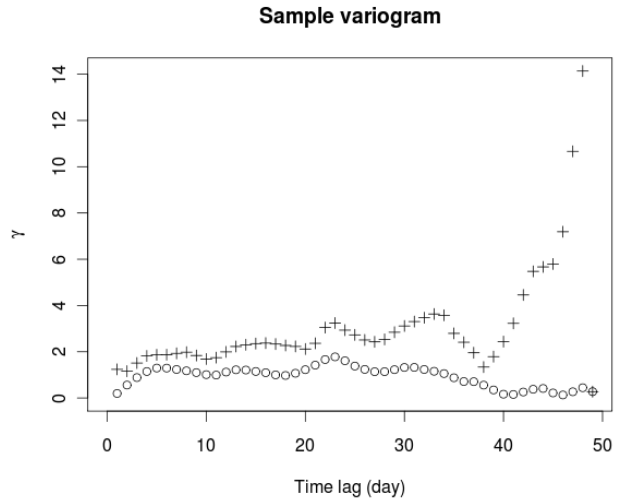




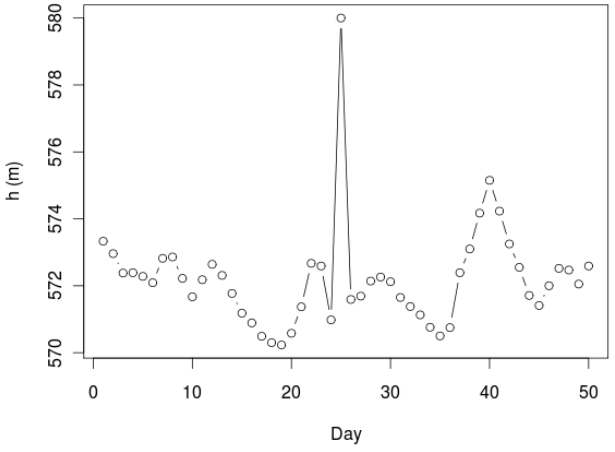
Example with piezometric height data

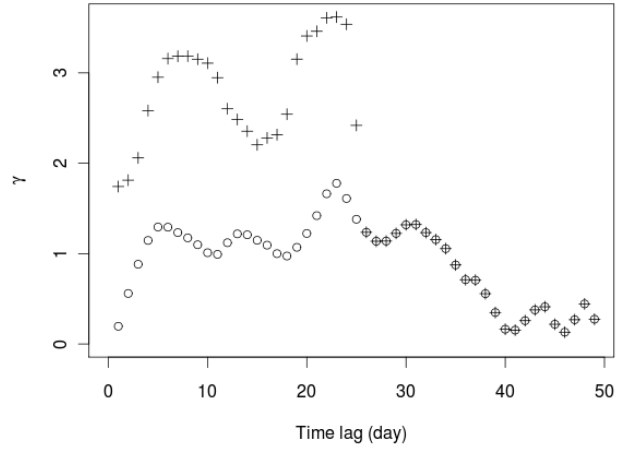
Set 1



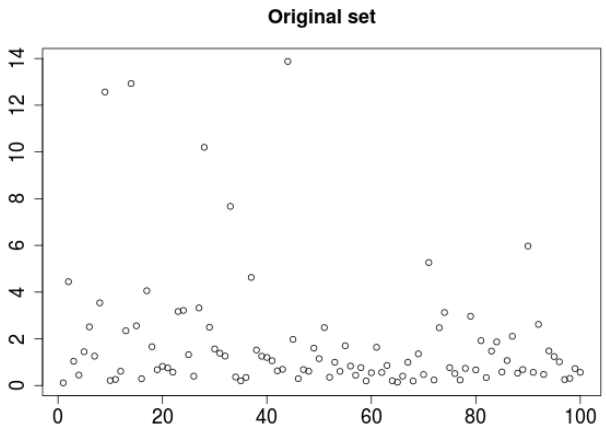


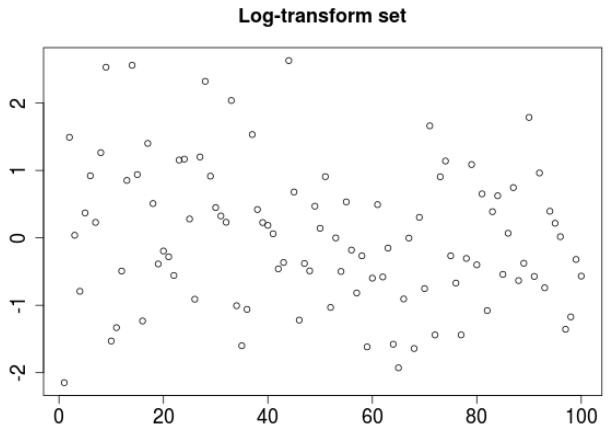
Set 2

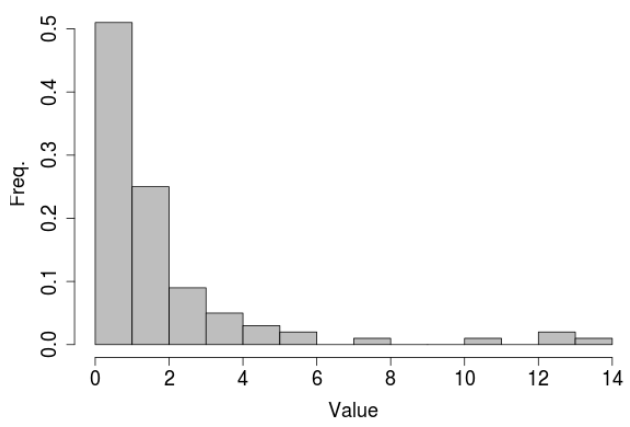


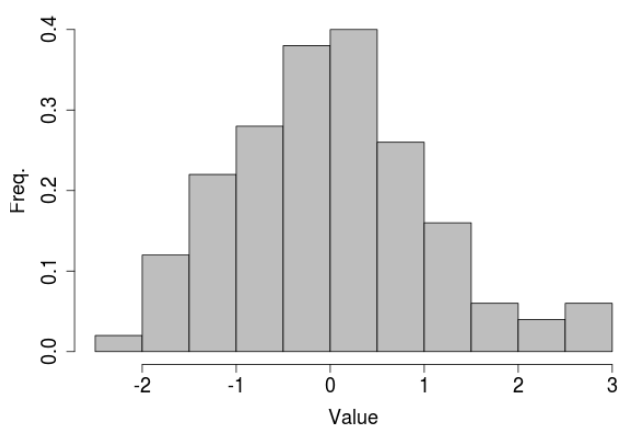


In case of heavy tailed or asymmetrical distribution of Z values, a possible way out is to transform Z values to get a more “regular” distribution. (log-transform, square-root transform, anamorphosis etc...)

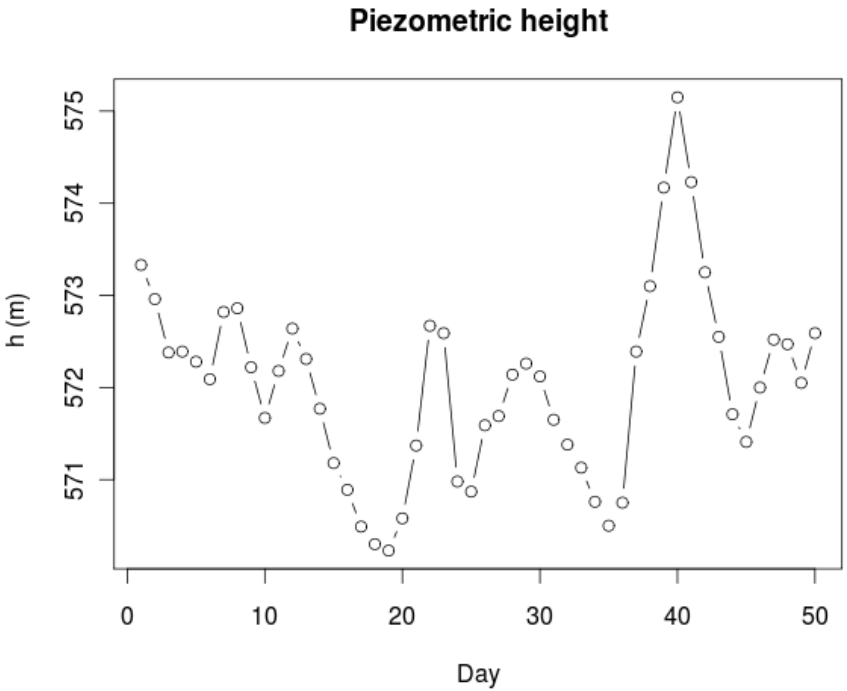


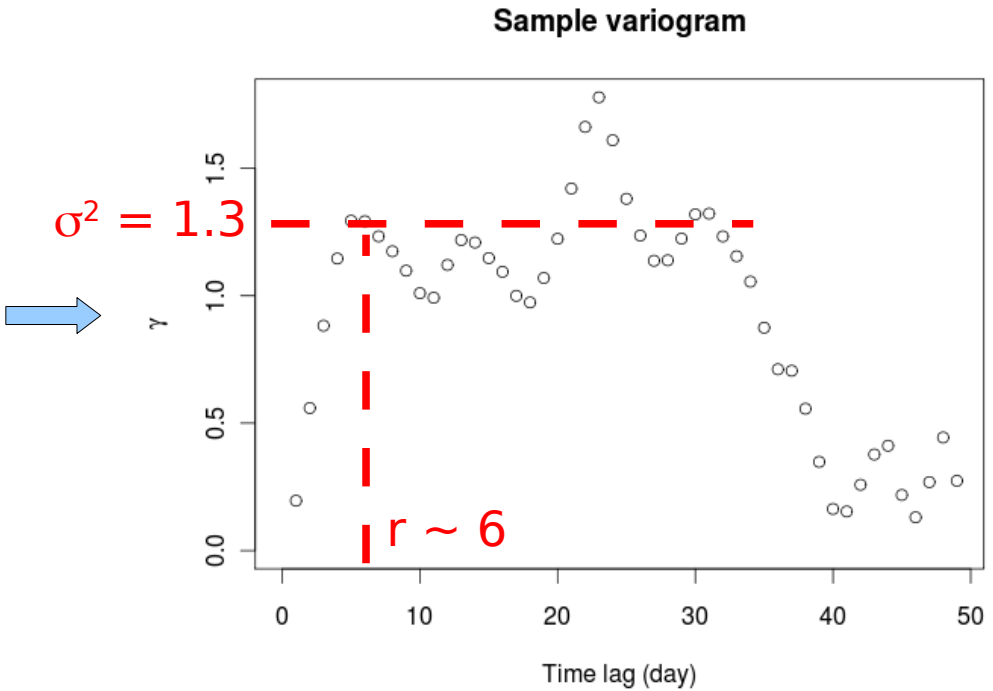






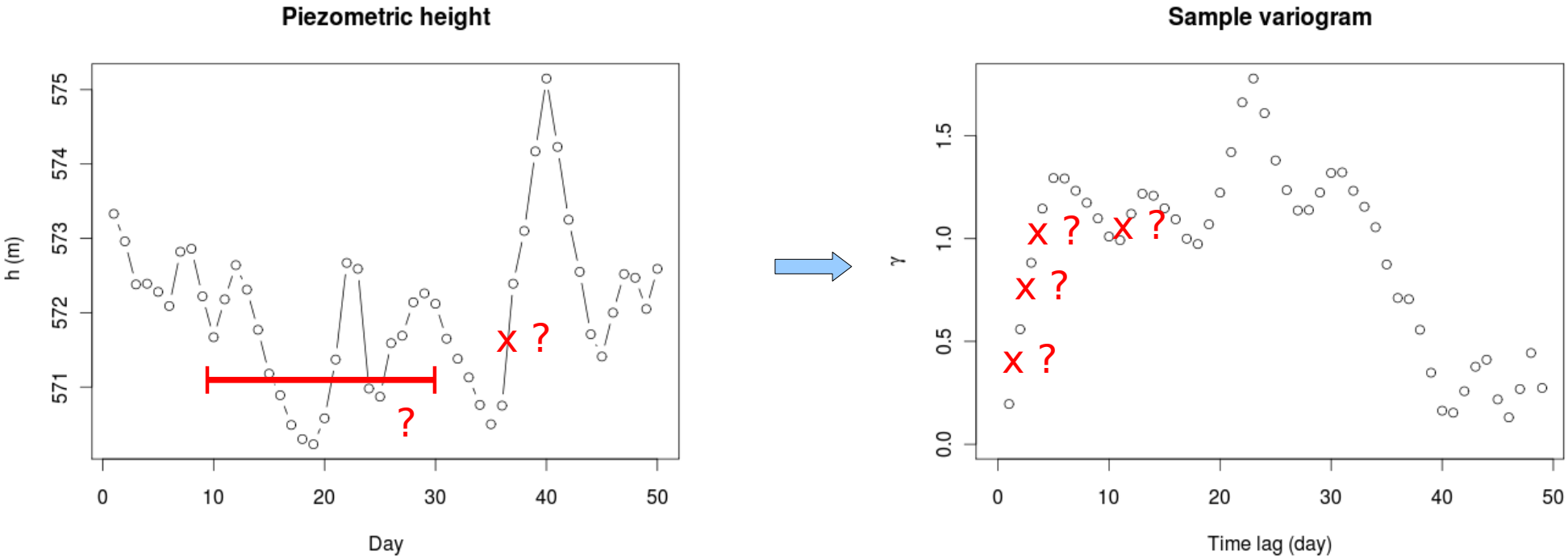
Water level at Bioley-Orjulaz





Sample variogram → structural analysis:

Range ~ 6 days
Sill ~ 1.3 m²
No nugget effect



If we are interested in the estimation of mean value over a domain or in interpolating measured data points, we must know the variogram at every distance lags + fulfill math properties (see chapter on kriging).

→ Need to fit an appropriate model to the sample variogram.

Mathematical properties of acceptable covariance/variogram models:

All variances calculated from model must be positive.

If 2nd-order stationary RF:

Y = linear combination of Z $$Y = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

Variance of Y must be > 0 $$\mathrm{Var}[Y] = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{C}(x_i - x_j)$$

C(h) = cov. of RF

→ C() must be positive definite

If intrinsic RF:
(variance is defined only for increments because *Var[Z] does not exist*)

Allowable linear combinations: $\sum_{i=1}^{n} \lambda_i = 0$

Variance of Y must be > 0 $\mathrm{Var}[Y] = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i - x_j)$

→ $\gamma()$ must be conditionally negative definite

$\gamma(h)$ = vario. of RF

Important property: if γ_1 and γ_2 are valid variogram models and given (a,b)>0, then $\gamma = a\gamma_1 + b\gamma_2$ is a valid model.

Proof

$$Y = \sum_{i=1}^{n} \lambda_i Z(x_i) = \sum_{i=1}^{n} \lambda_i Z_i \quad \text{with} \quad \sum_{i=1}^{n} \lambda_i = 0$$

$$\begin{aligned}
\mathrm{Var}[Y] &= \mathrm{Var}\left[\sum_{i=1}^{n} \lambda_i Z_i\right] = \mathrm{Var}\left[\sum_{i=1}^{n} \lambda_i (Z_i - Z_\delta)\right] \\
&= E\left[\left(\sum_{i=1}^{n} \lambda_i (Z_i - Z_\delta)\right)^2\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j E\left[(Z_i - Z_\delta)(Z_j - Z_\delta)\right]
\end{aligned}$$

$$\begin{aligned}
\gamma_{ij} &= \frac{1}{2} E\left[(Z_i - Z_j)^2\right] = \frac{1}{2} E\left[(Z_i - Z_\gamma - (Z_j - Z_\gamma))^2\right] \\
&= \frac{1}{2} E\left[(Z_i - Z_\gamma)^2 + (Z_j - Z_\gamma)^2 - 2(Z_i - Z_\gamma)(Z_j - Z_\gamma)\right] \\
&= \gamma_{i\delta} + \gamma_{j\delta} - E\left[(Z_i - Z_\gamma)(Z_j - Z_\gamma)\right]
\end{aligned}$$

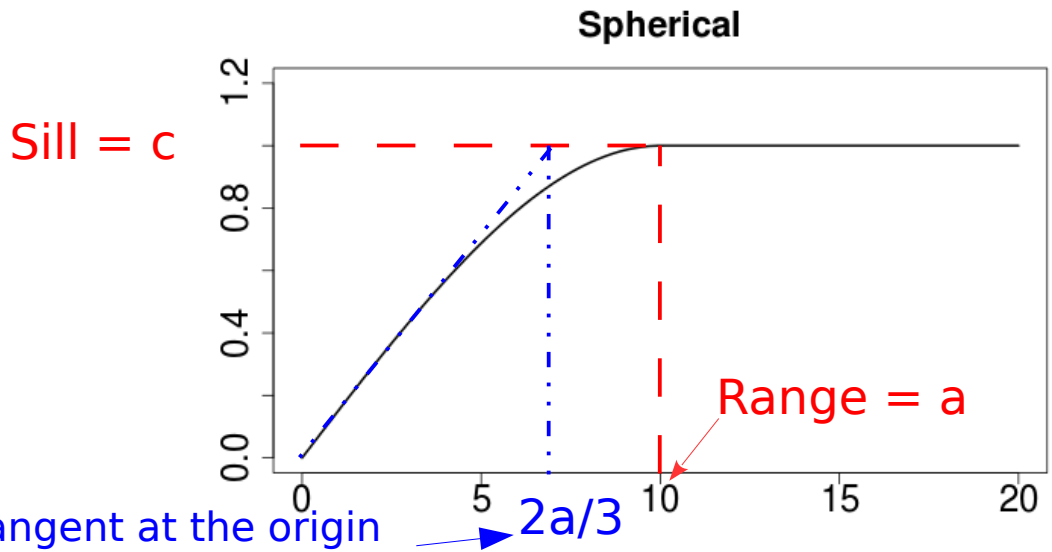
$$\Rightarrow \mathrm{Var}[Y] = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma_{ij}$$

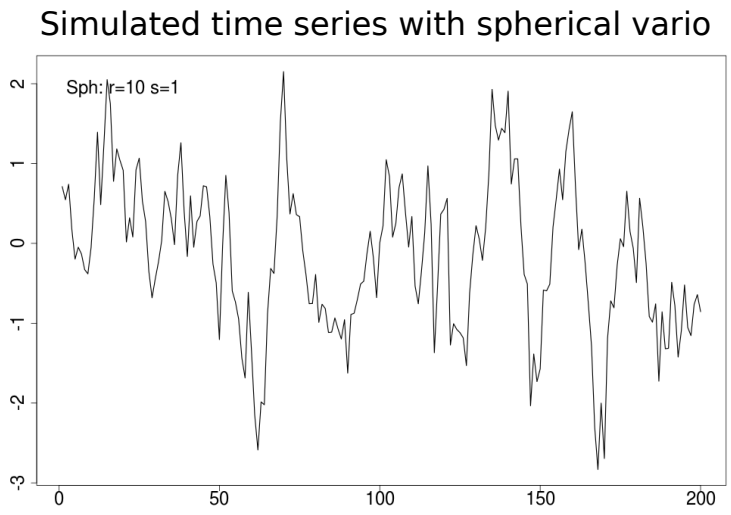
Examples of commonly used isotropic models

1. Spherical model

$$\gamma(h) = \begin{cases} c\left[\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right] & \forall\ h \leq a \\ c & \forall\ h > a \end{cases}$$

Explicit range
Related to overlapping volume of 2 spheres



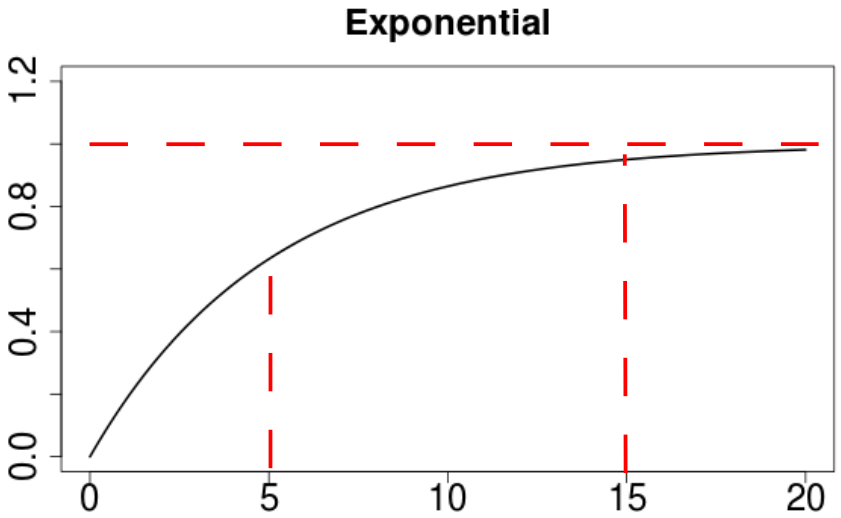


2. Exponential model

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

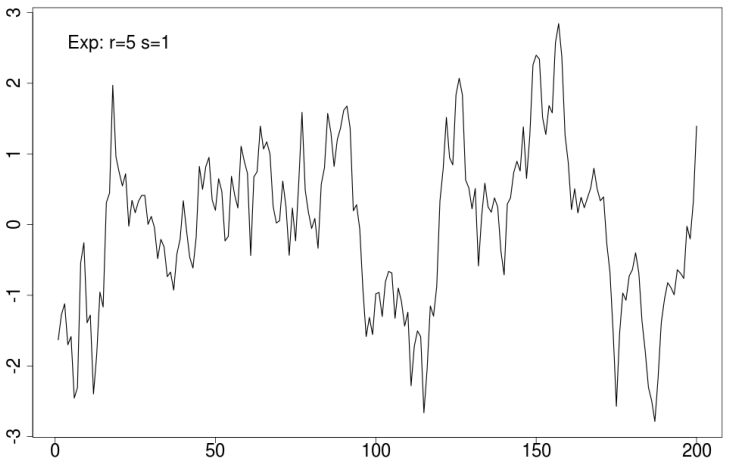
No explicit range
pseudo-range: r / γ(r) ~0.95 c → r ~ 3a

Sill = c



a Pseudo-range ~ 3a

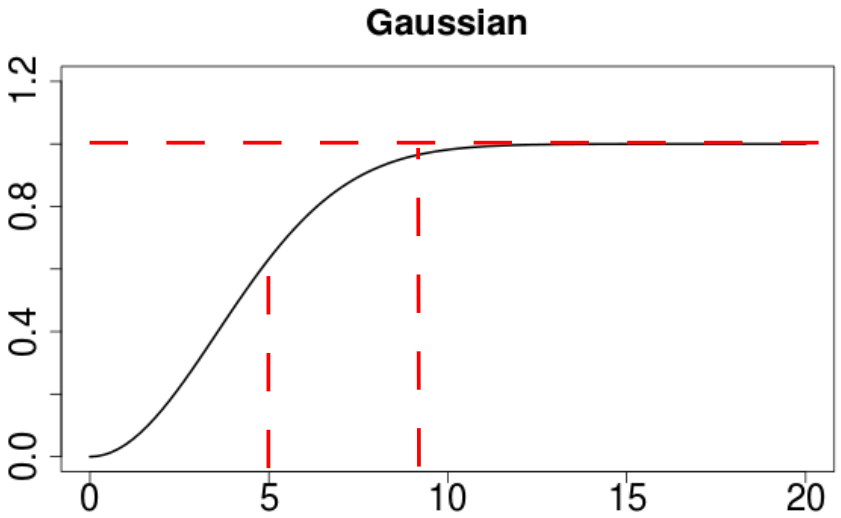
Simulated time series with exponential vario



3. Gaussian model $$\gamma(h) = c\left[1 - e^{-\frac{h^2}{a^2}}\right]$$

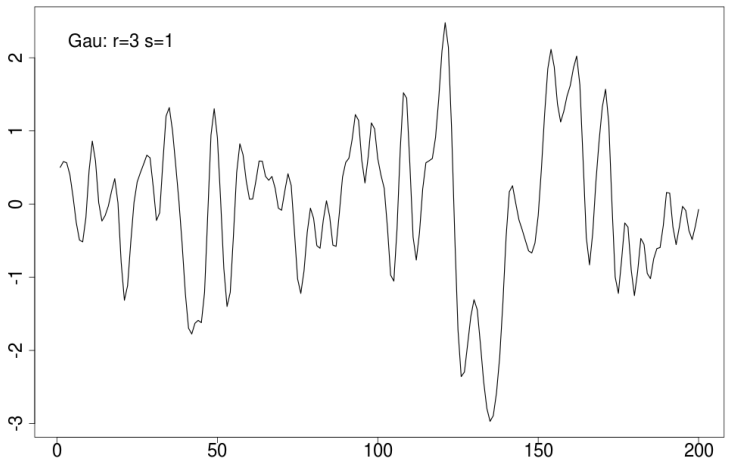
No explicit range
pseudo-range: r / γ(r) ~0.95 c → r ~1.7a

Sill = c



a 1.7a ~ pseudo-range

Simulated time series with Gaussian vario

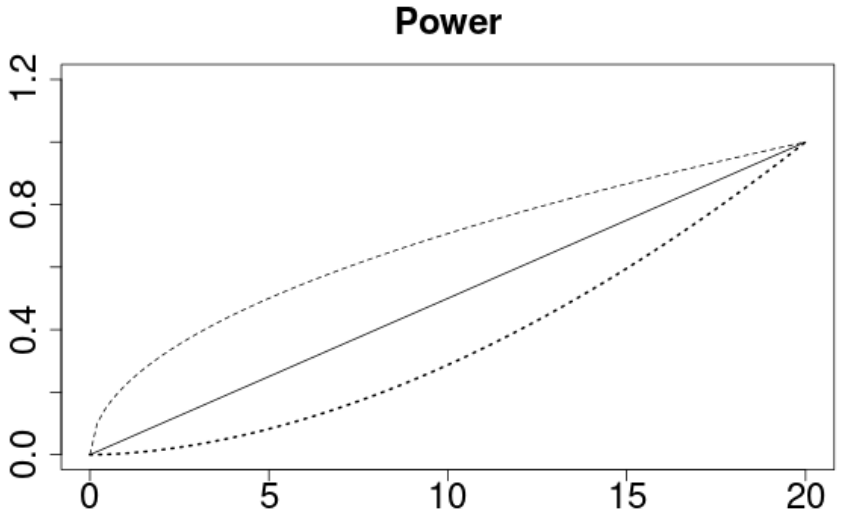


4. Power model

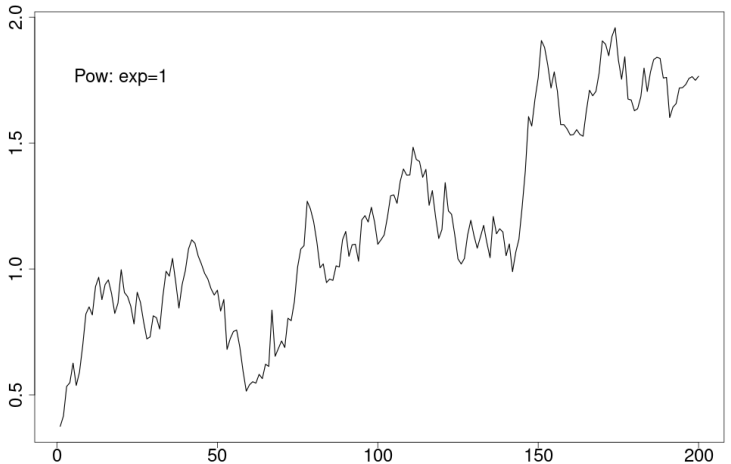
$$\gamma(h) = \alpha h^{\beta} \quad 0 < \beta < 2$$

No range / sill

β = 0.5 ; 1 ; 1.8
α = 1/(20^b)



Simulated time series with power vario

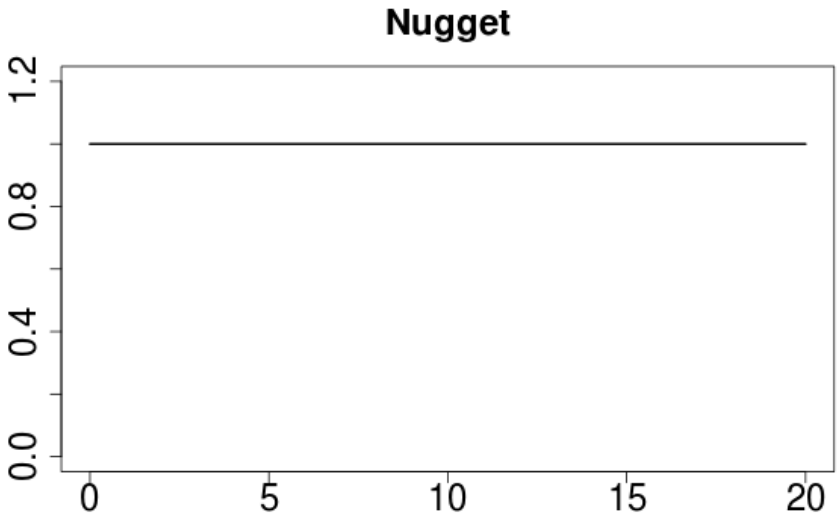


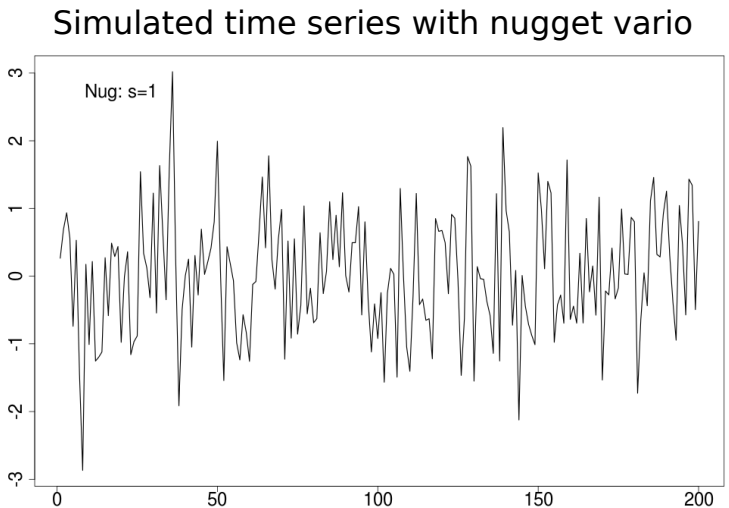
5. Pure nugget model

$$\gamma(h) = \begin{cases} c_0 & h > 0 \\ 0 & h = 0 \end{cases}$$

No structure!

c_0

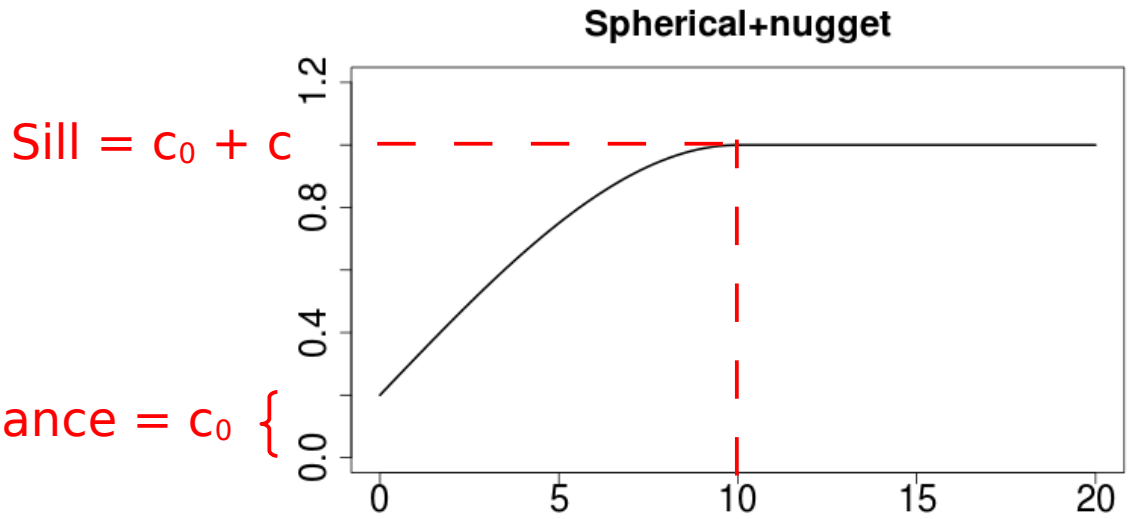




Combination of models

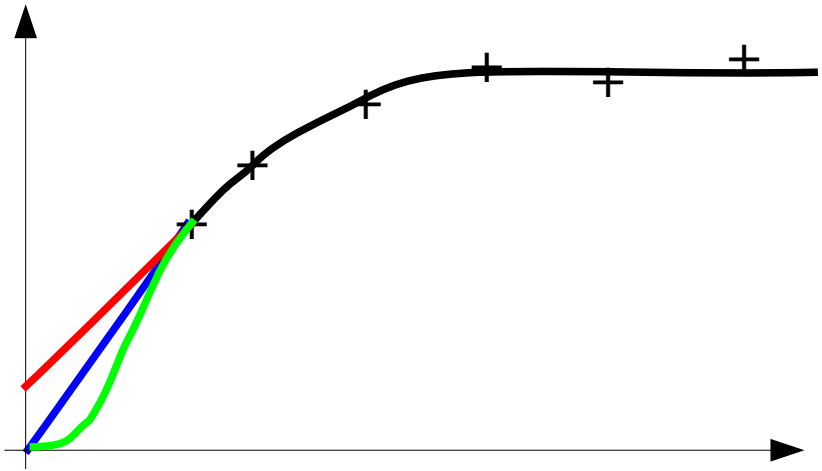
$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c\left[\dfrac{3h}{2a} - \dfrac{1}{2}\left(\dfrac{h}{a}\right)^3\right] & 0 < h < a \\ c_0 + c & h \geq a \end{cases}$$

Spherical + nugget



Fitting a model to a sample variogram is tricky and involves some arbitrary choices that can be based on “physical” knowledge of the studied process. Hence, the fitting procedure generally involves the practitioner rather than being entirely automatic.

Fitting the behavior near the origin



1. Parabolic: very regular structure.

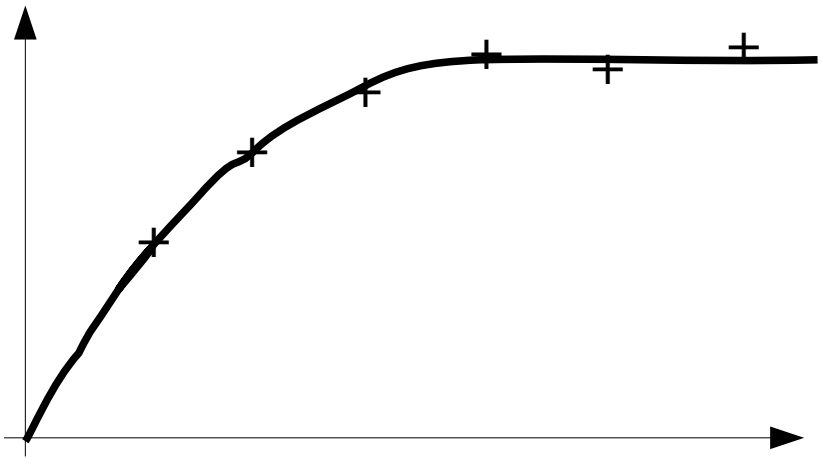
2. Linear: regular structure.

3. Discontinuous: nugget effect + measurement errors.

The practitioner has to choose a type of behavior at the origin, depending on the knowledge he/she has on the process.

This choice has consequences on the subsequent interpolation...

Fitting the behaviour at medium and long distance lags



Characteristics that needs to be correctly reproduced:

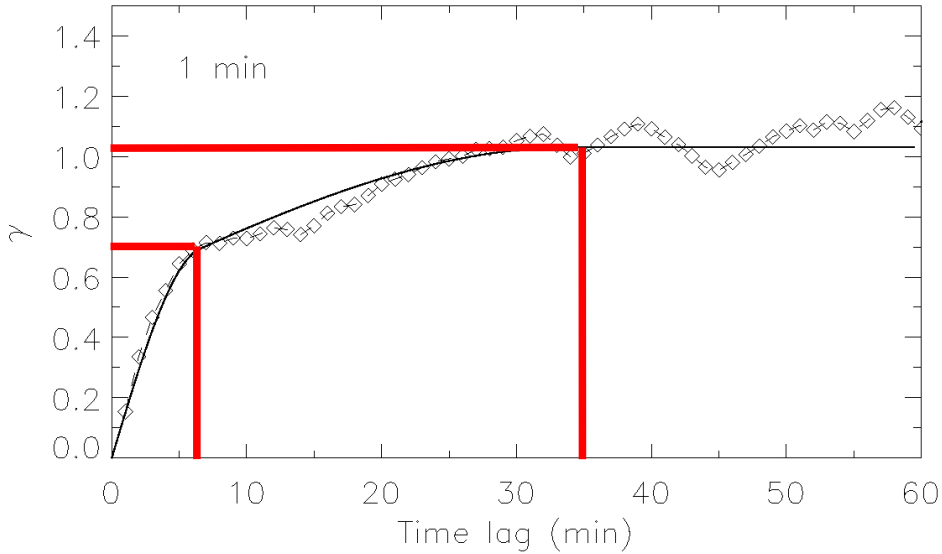
- Slope at the origin: evaluated from short distance lags.

- Sill: level at which sample variogram stabilizes.

- Range: variogram reaches the sill
(spherical model: tangent at origin intersects sill at 2/3 of the range)

Nested structure

ex: time series of rain rate
(1-min resolution)



2 structures appear in the sample variogram.

Model fitted = sum of 2 spherical models:

- 1st structure with range ~6 min and sill ~ 0.7 → convective cells.

- 2nd structure with range ~ 35 min and sill ~1 → precipitating system.

Questions

- Why is it necessary to fit a variogram model to the sample variogram?

- Which variogram models (among those listed) have an explicit range?

- Demonstrate that the pseudo-range of the exponential model is 3a.

Automatic fitting procedures

Main difficulties in fitting a varioram model on a sample variogram:
1. Uncertainty associated with sample variogram (generally not known).
2. Most models are non-linear in one or more parameters.
3. Possible anisotropy.

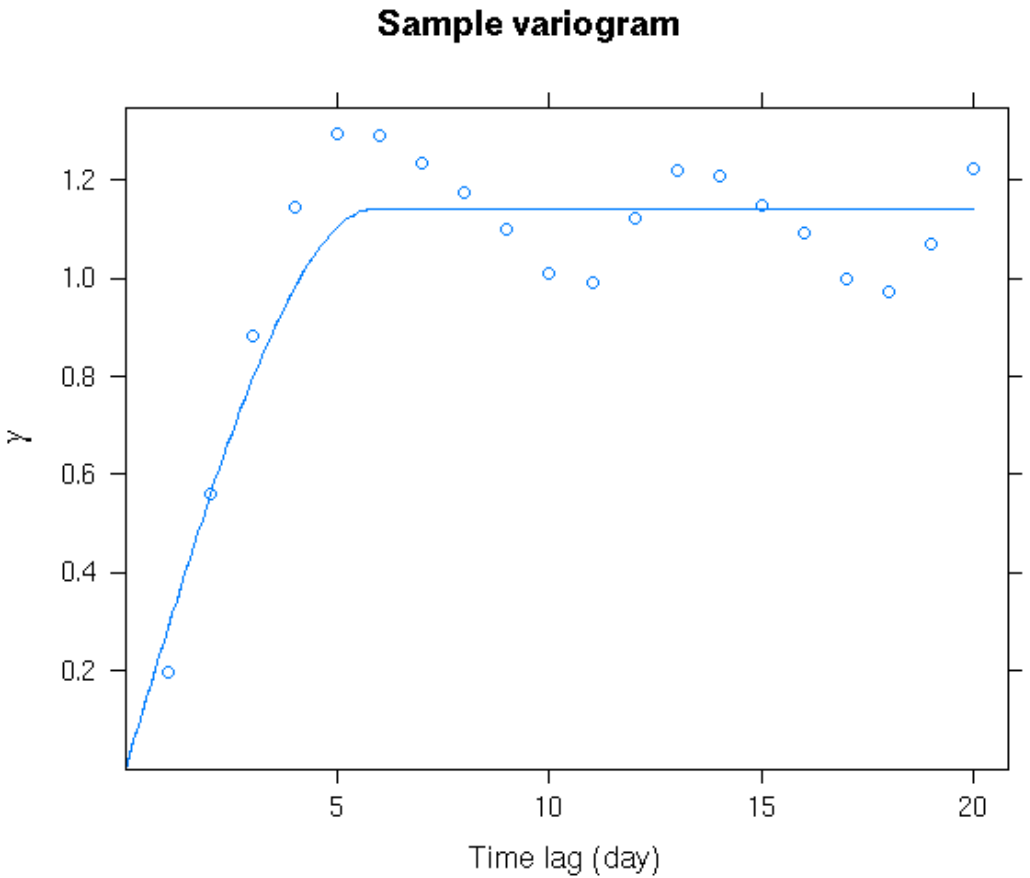
1 and 2 → equal weight for all sample variogram estimates is not relevant
→ weighted least square approach.

$$Q(\mathbf{p}) = \sum_{i=1}^{n_h} w_i \left[\hat{\gamma}(h_i) - \gamma(h_i, \mathbf{p})\right]^2$$

p vector (n_px1) of model parameters to be estimated so that Q(**p**) is min.
n_h number of distance lags used to estimate $\hat{\gamma}$
h_i distance lag of i^{th} class.
$\hat{\gamma}$ sample variogram.
γ model variogram.
w_i weight at h_i, can take into account number of pairs $n(h_i)$ used to estimate $\hat{\gamma}$ at h_i

$$\omega_i = \frac{n(h_i)}{\gamma(h_i, \mathbf{p})^2}$$

Ex piezometric height from Bioley - Orjulaz



range: 5.9 days

sill: 1.14 m^2

Nugget: 0

Variography in presence of a drift

Intrisic RF Z is sum of zero-mean RF Y and deterministic drift m:

$$Z(x) = Y(x) + m(x)$$ Y and m are unknown!

Residuals: $$R(x) = Z(x) - \hat{m}(x)$$ $\hat{m}$ = estimate of m on sample $Z(x_i)$

Variogram of R can be different from variogram of Y (even if $\hat{m}$ unbiased)

What to do?
- Rigorous approach: IRF-k, beyond the scope of this course!
(interested reader can refer to Chilès and Delfiner, p.238)

- 2D data: if drift strong directionality θ_0 → vario ⊥ θ_0 less affected by drift.
→ estimate of Y vario.

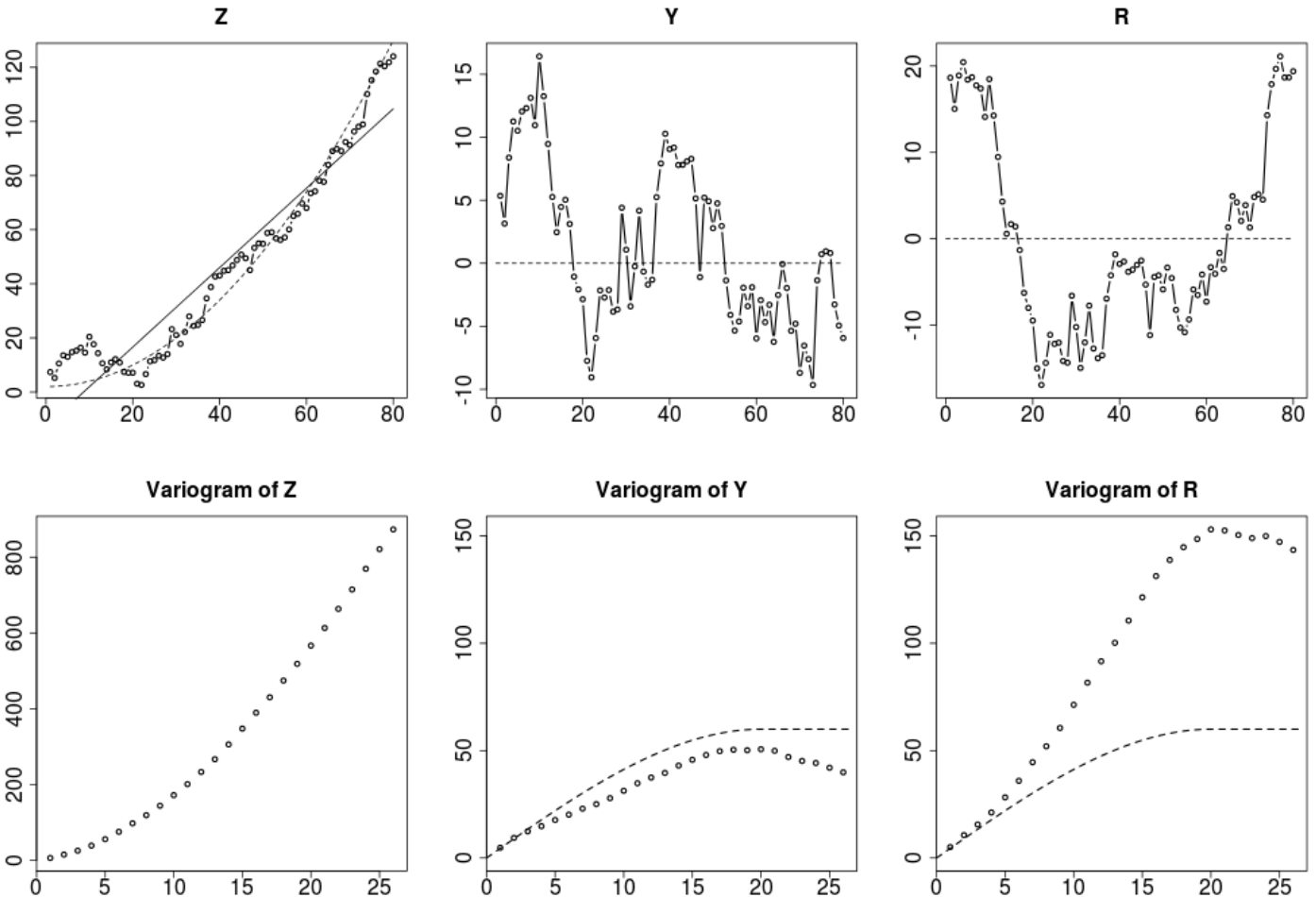
- Bias is negligible at short distance: → estimate $\hat{m}$ (e.g., least squares fitting)
→ assume vario R ~ vario Y (at short dist).

Example

Y simulated from sph. model (mean=0, r=20, s=60).

Z = Y + quadratic drift (~ x^2)

R: residuals from linear drift estimated using least square fitting on Z.



Guidelines

- Careful definition of the domain.
- Careful analysis of data: distribution, outliers?
- Sampling must be more or less homogeneous.
- Estimation and fitting of isotropic variogram:
  n > 20-30 per class to have reliable estimates
  → adjust class width and angular tolerance.
  → general rule: distance lags up to ½ domain size...
- Check anisotropy using directional variograms.
- Emphasis must be on short distance lags for model fitting!

1. Variogram

- Definition, link with covariance

- In 2D (or more), anisotropy

- Physical interpretation:
    1. Nugget: small-scale variability, measurement error
    2. Range: decorrelation distance
    3. Sill: variance of the considered (stationary) random function

2. Estimating the variogram

- Sample or experimental variogram, Matheron estimator (min 20-30 pairs)
- Sampling effects

3. Modeling the variogram

- Mathematical properties (covariance, variogram)
- Variogram models (spherical, exponential, Gaussian, power)
- Fitting a variogram model (manual/automatic, in presence of a drift)