Outline:

1. Types of spatial data

2. What is Geostatistics?

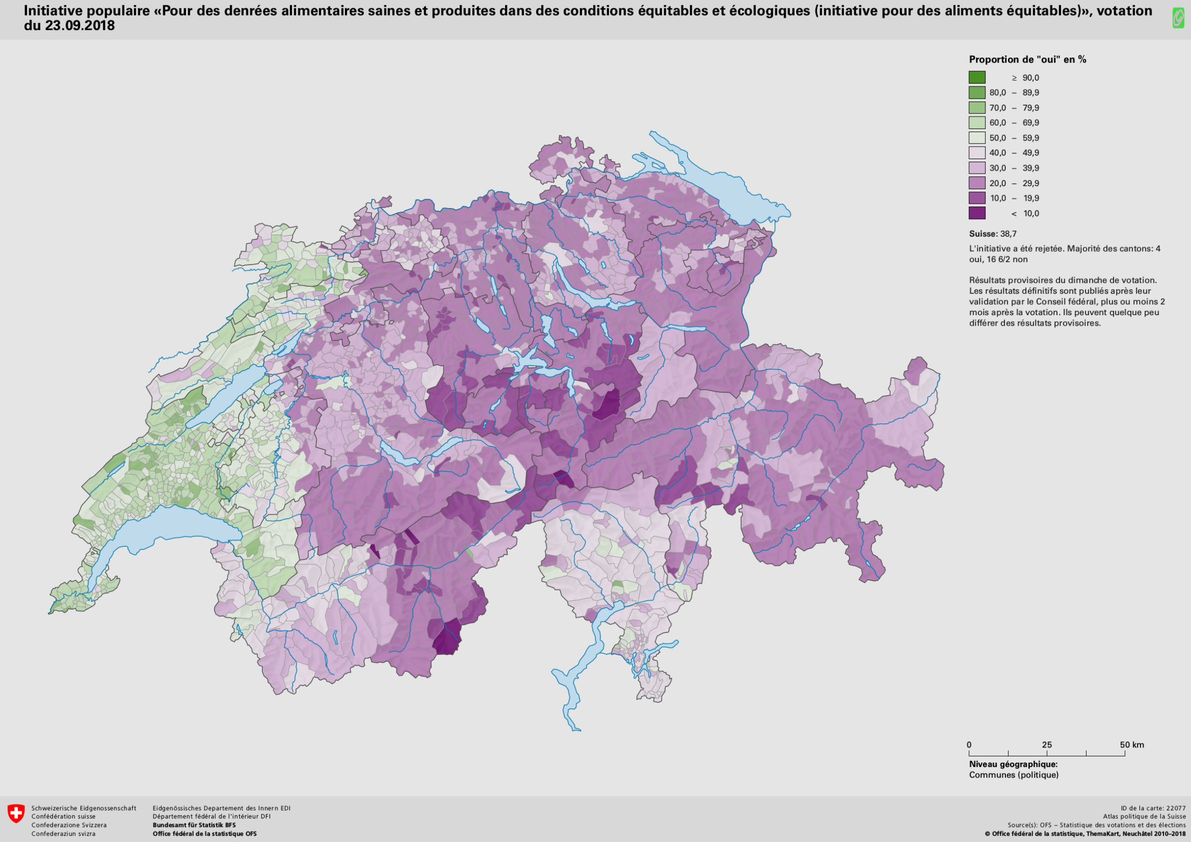
3. Basic statistics reminder...

4. Going into space

Different types of spatial data

1. Lattice data

Domain is fixed and countable, data on grid (regular or not).
Spatial coordinates are not continuous but discrete (one community)



Initiative populaire «Pour des denrées alimentaires saines et produites dans des conditions équitables et écologiques (initiative pour des aliments équitables)», votation du 23.09.2018

Suisse: 38,7

L'initiative a été rejetée. Majorité des cantons: 4 oui, 16 6/2 non

Résultats provisoires du dimanche de votation. Les résultats définitifs sont publiés après leur validation par le Conseil fédéral, plus ou moins 2 mois après la votation. Ils peuvent quelque peu différer des résultats provisoires.

Niveau géographique:
Communes (politique)

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

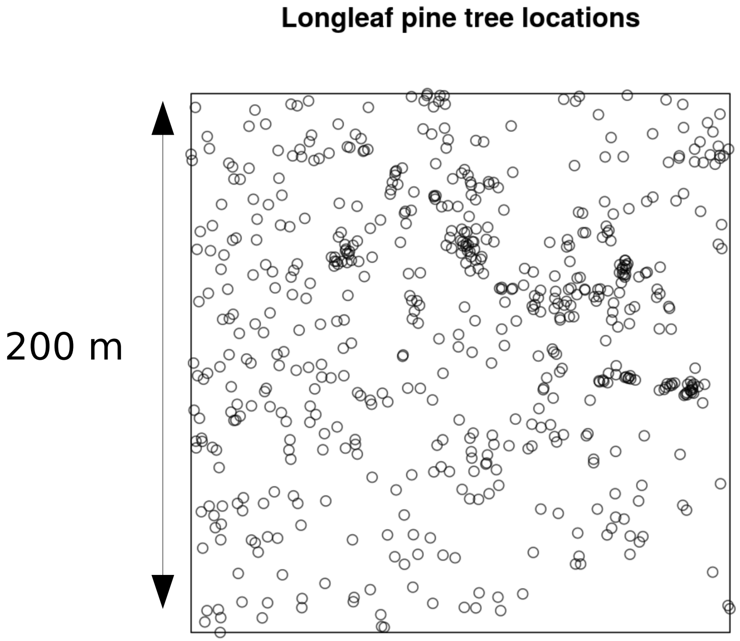
Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Bundesamt für Statistik BFS
Office fédéral de la statistique OFS

ID de la carte: 22077
Atlas politique de la Suisse
Source(s): OFS – Statistique des votations et des élections
© Office fédéral de la statistique, ThemaKart, Neuchâtel 2010–2018

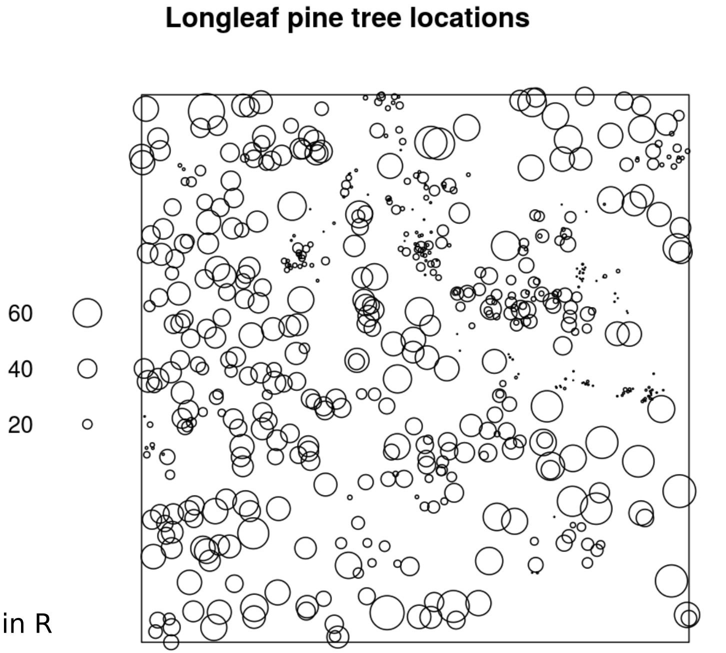
2. Point patterns

Domain is random.
Only location is of interest → unmarked point patterns
Location + attribute are of interest → marked point patterns

Ex of unmarked point process:
location of trees



Ex of marked point process:
location and diameter of trees

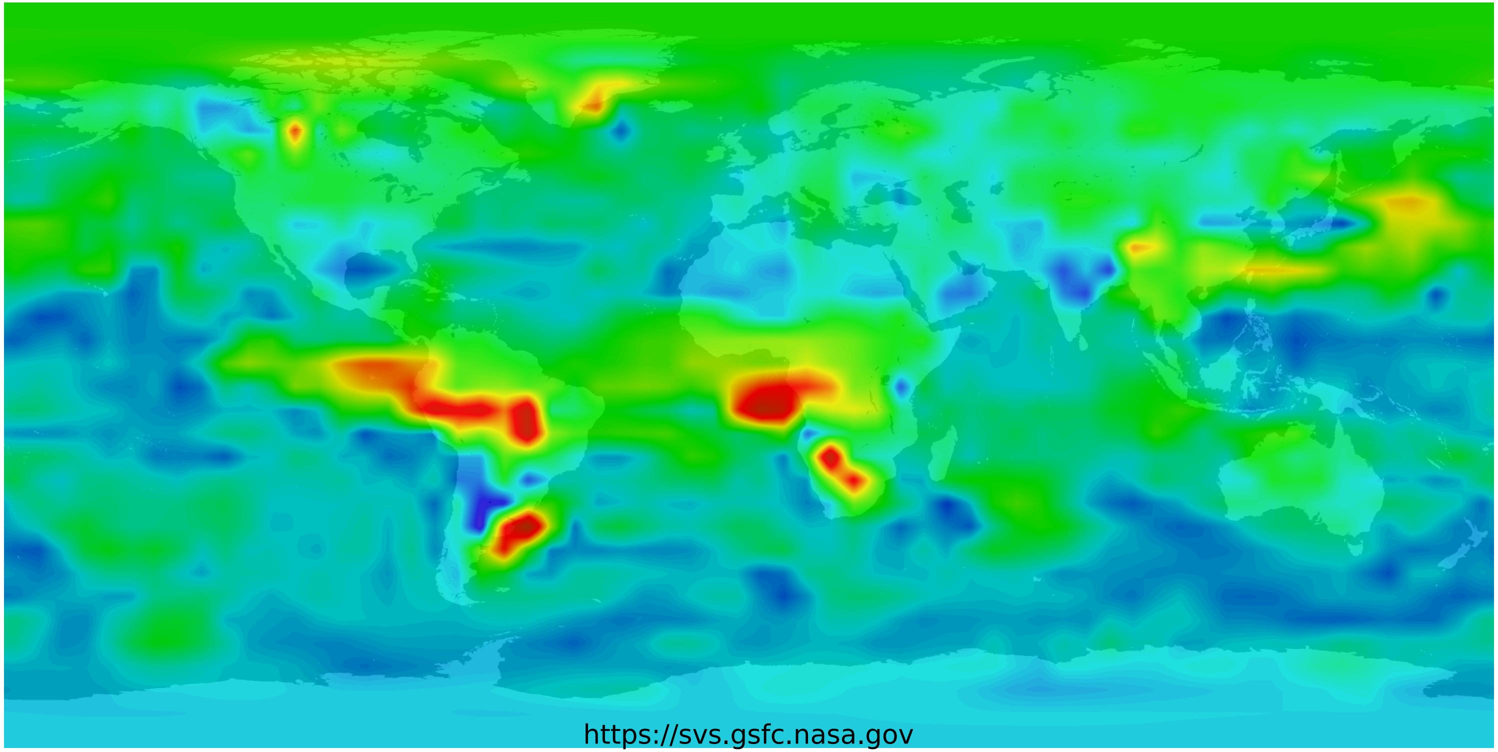


Longleaf data in
spatstat package in R

3. Geostatistical data

Domain (over which variable is analyzed) is fixed.
Space coordinates and variable vary continuously.

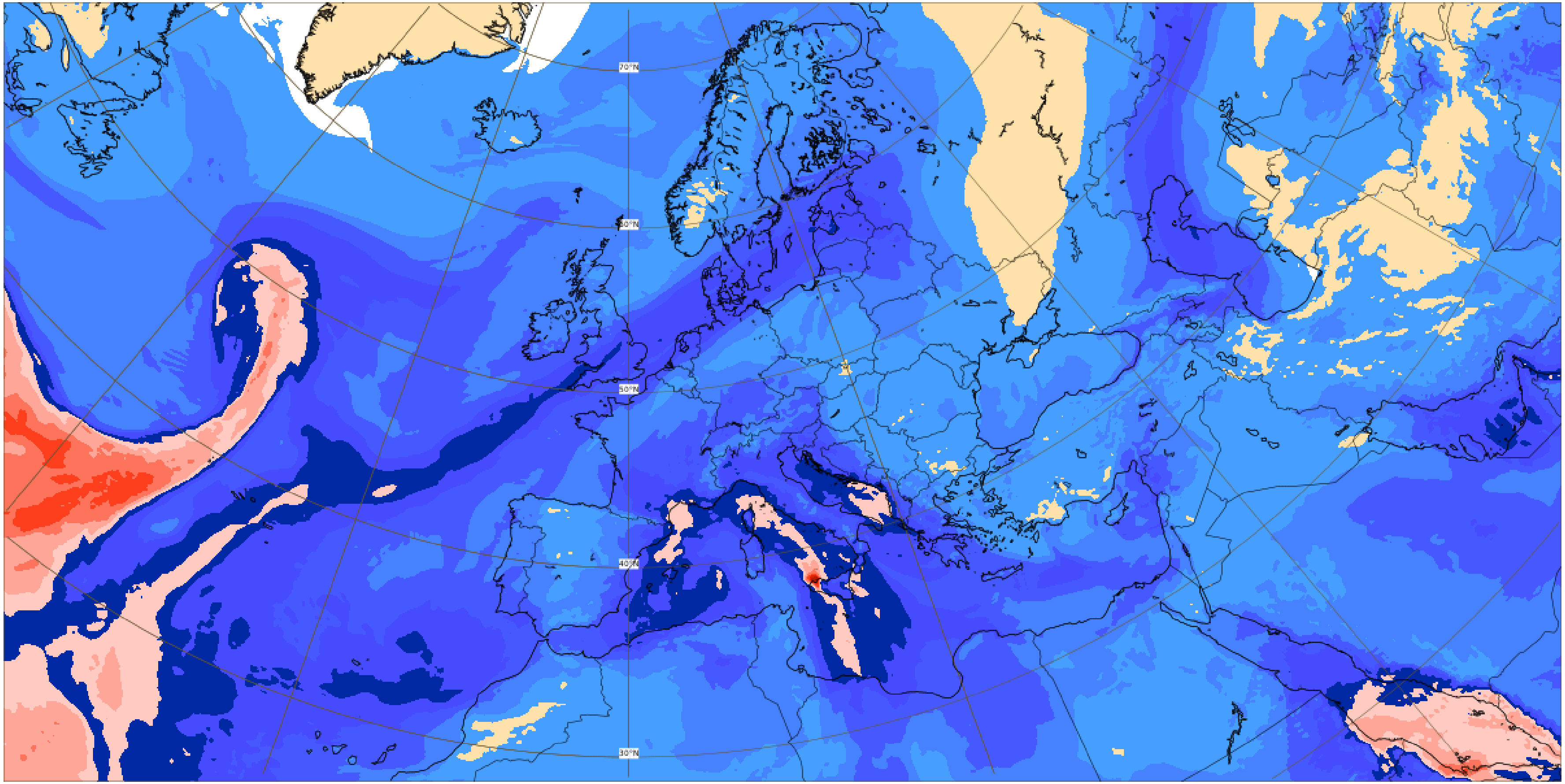
Global map of CO concentration as seen by Terra/MOPITT



https://svs.gsfc.nasa.gov

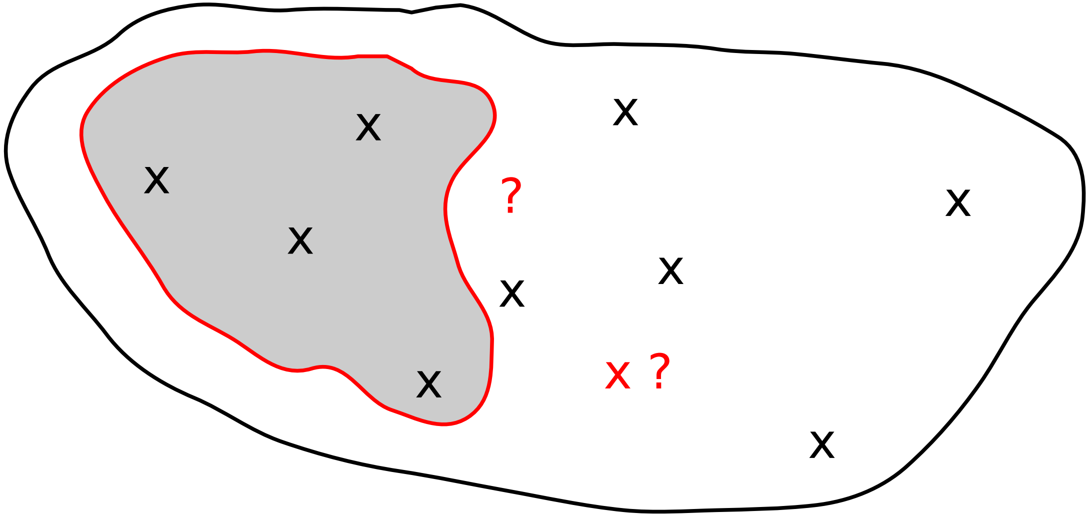
Phenomenon with STRUCTURE → essential to take this structure into account!

Ex: vertically integrated total water column (ECMWF)



https://charts.ecmwf.int/catalogue/packages/opencharts/

What is Geostatistics?



Main questions:

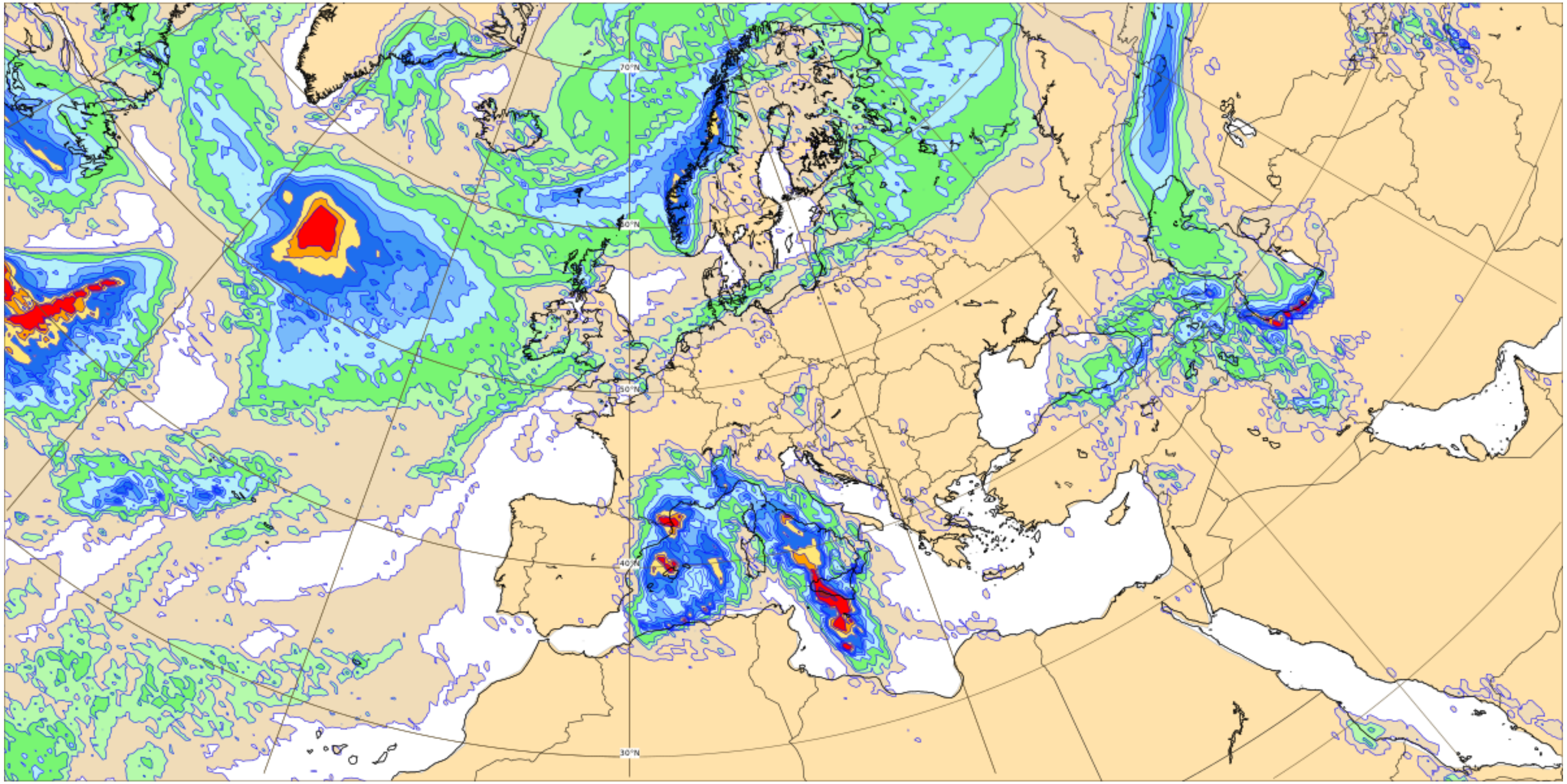
- What are the values where no measurements (interpolation / mean estimation)?
- What is the error associated with these estimates?

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

Different types of applications: Structural analysis (1)

What does an observation tell us about neighboring points?
Are variations similar in all directions or is there anisotropy?
Is there a trend? Are there characteristic scales?

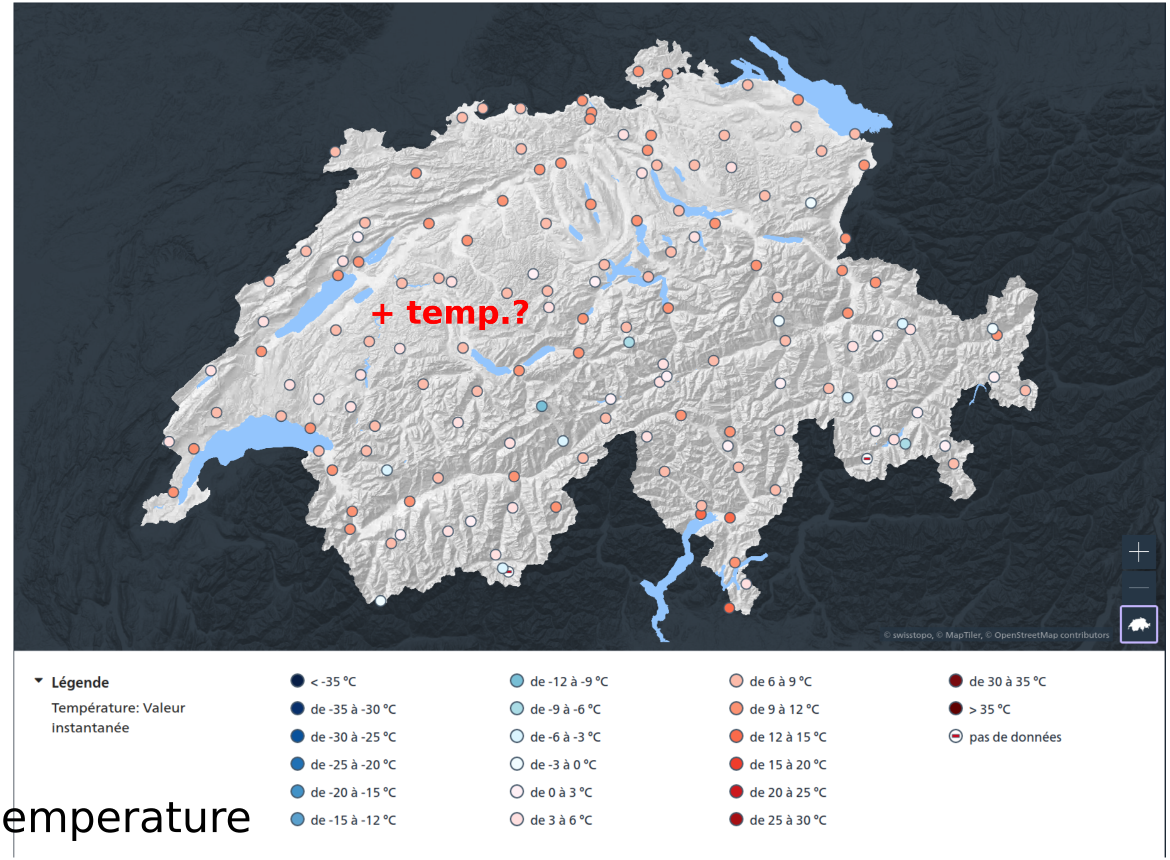
Ex: 24h-precip (ECMWF)



Variogram will be a key function to answer these questions

Different types of applications: Interpolation (2)

What is the value among observed points?
What is the uncertainty associated with the interpolated values?



Légende

Température: Valeur instantanée

- ● < -35 °C
- ● de -35 à -30 °C
- ● de -30 à -25 °C
- ● de -25 à -20 °C
- ● de -20 à -15 °C
- ● de -15 à -12 °C
- ● de -12 à -9 °C
- ● de -9 à -6 °C
- ○ de -6 à -3 °C
- ○ de -3 à 0 °C
- ○ de 0 à 3 °C
- ● de 3 à 6 °C
- ● de 6 à 9 °C
- ● de 9 à 12 °C
- ● de 12 à 15 °C
- ● de 15 à 20 °C
- ● de 20 à 25 °C
- ● de 25 à 30 °C
- ● de 30 à 35 °C
- ● > 35 °C
- ⊖ pas de données

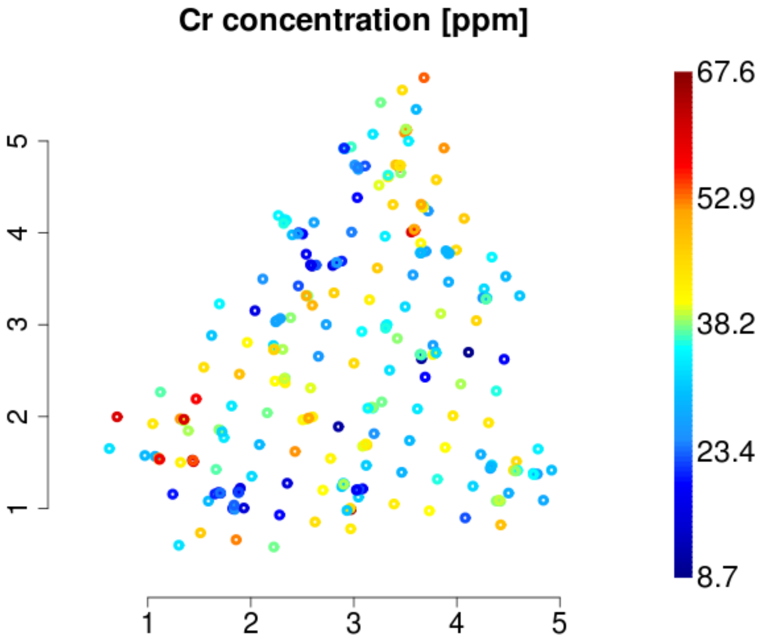
Ex : temperature

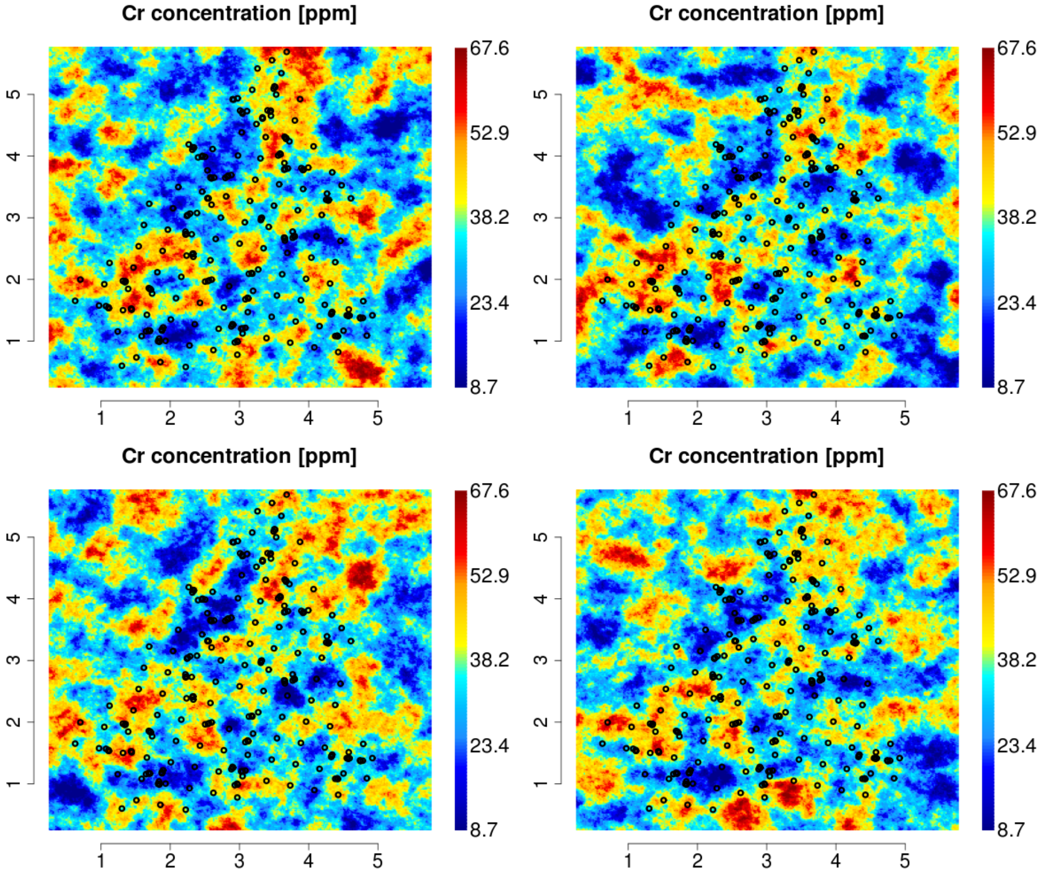
Kriging is a possible approach to answer these questions.

Different types of applications: Simulation (3)

How to overcome smoothing effects of interpolation?
How to take into account measurements at given points?

Ex: conditional simulation





A brief history...

- 1930s – Kolmogorov developed structure function to study turbulence.

- 1950s – Krige: empirical approach for mining applications.

- 1960s – Matheron (mining), Matern (forestry) and Gandin (meteorology) developed an ensemble of tools that eventually formed geostatistics.

- Still an active research domain (simulation, multivariate, non-stationary,...).

Geostatistics has been applied to a variety of domains:
mining, hydrology, forestry, geophysics, biology...

References

Books:

- “Geostatistics for natural resources evaluation”, 1997, Goovaerts, Oxford University Press, Applied Geostatistics Series.
- “Geostatistics: Modeling Spatial Uncertainty”, 2nd Ed., 2012, Chilès and Delfiner, Wiley Series in Probability and Statistics, 699 p.
- “Statistics for Spatial Data”, 1993, Cressie, Wiley Series in Probability and Mathematical Statistics, 900 p.
- “Geostatistics for Environmental Scientists”, 2007, Webster and Oliver, Wiley, Statistics in Practice, 271 p.

On-line:

- A Practical Primer on Geostatistics – USGS: https://pubs.er.usgs.gov/publication/ofr20091103
- https://www.coursehero.com/file/26842503/Bohling-2007-Introduction-to-Geostatisticspdf/

Exercises

Exercises associated with each chapter.

Based on open-source software R.

Not evaluated.

Solutions the next week.

Feel free to contact the assistants if you have questions!

Probability density function (PDF):

A random variable can be fully described by its associated PDF

If X is a discrete RV:

$$f_X(x) = \begin{cases} P[X = x_i] & \text{if x = x}_\text{i}\text{ , i=1..n} \\ 0 & \text{if x} \neq \text{x}_\text{i} \end{cases}$$

$$\sum_i f_X(x_i) = 1$$

If X is a continuous RV:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

$$P[x < X(\omega) \leq (x + \mathrm{dx})] = f_X(x)\mathrm{dx}$$

$$\int_A f_X(x)\,dx = 1$$

Descriptors of a RV

Expected value of a discrete RV:

$$\mathrm{E}[X] = \sum_i f_X(x_i) x_i$$

Expected value of a function g of X:

$$\mathrm{E}[g(X)] = \sum_i f_X(x_i) g(x_i)$$

Expected value of a continuous RV:

$$\mathrm{E}[X] = \int_A x f_X(x)\, dx$$

Expected value of a function g of X:

$$\mathrm{E}[g(X)] = \int_A g(x) f_X(x)\, dx$$

Expectation is a linear operator:

$$\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$$

Descriptors of a RV

Raw moments of a RV

$$\mu'_n(X) = \mathrm{E}[X^n]$$

Central moments of a RV

$$\mu_1(X) = \mu'_1(X)$$

$$\mu_n(X) = \mathrm{E}\left[(X - E[X])^n\right]$$

$$\mu_n(X) = \sum_{k=0}^{n} (-1)^{n-k} \tbinom{n}{k} \mu'_k \mu_1^{n-k}$$

Binomial series identity:

$$(a+b)^n = \sum_{k=0}^{n} \tbinom{n}{k} a^k b^{n-k} \qquad \tbinom{n}{k} = \frac{n!}{k!(n-k)!}$$

Descriptors of the center of a distribution

(using mainly 1st order moments)

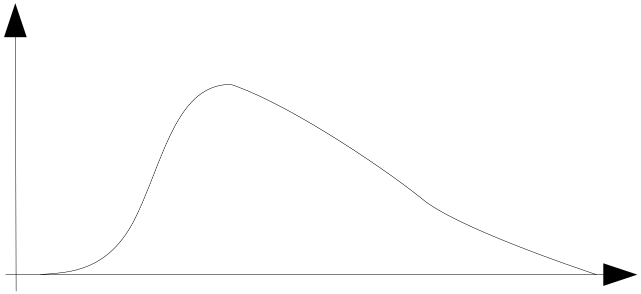
(1) Mean $\mu(X) = \mathrm{E}[X]$ Mean = x value of the center of gravity of PDF

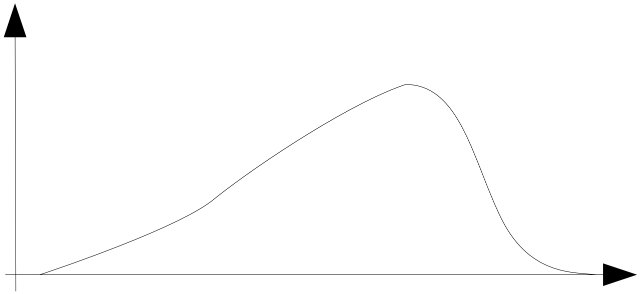
(2) Median : 50% of the values are below the median. Median is more robust than mean.

X in {xi,..,x_n} if n odd median = $x_{(n+1)/2}$

if n even median = 1/2($x_{n/2}$ + $x_{1+n/2}$)

(3) Mode : most probable value = value at which f_X is max (if any).





Descriptors of the “dispersion” of a distribution

(moment order > 1)

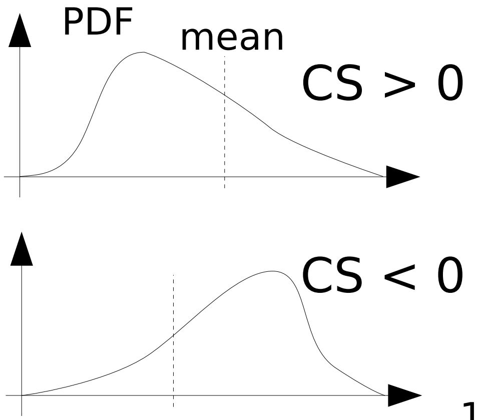
Variance
$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

Standard deviation
$$\sigma[X] = \sqrt{\mathrm{Var}[X]} = \sqrt{\mathrm{E}[(X - \mathrm{E}[X])^2]}$$

Coefficient of variation
$$CV = \frac{\sigma}{\mu_1}$$

Quantifies variability normalized by the mean

Coefficient of skewness
$$CS = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$



Considering 2 RV X and Y, one possible way to characterize their relationship is to calculate their covariance or their correlation coefficient.

Covariance:
$$\begin{aligned}\mathrm{Cov}[X,Y] &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY] - \mathrm{E}[X]\mathrm{E}[Y]\end{aligned}$$

Correlation coefficient:
$$\rho_{XY} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$$
(ρ = Cov for RV with std = 1)

Both coefficients measures the linear relationship between X and Y.

Correlation coef. is better in this sense because it is dimensionless and normalized, so it eases comparison.

Properties of variance

$$\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X]$$

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$$

useful
in geostat

$$\mathrm{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}[X_i, X_j]$$

$$\mathrm{Var}\left[\int_A cZ(x)\,dx\right] = c^2 \int_A \int_A \mathrm{Cov}[Z(x), Z(y)]\,dx\,dy$$

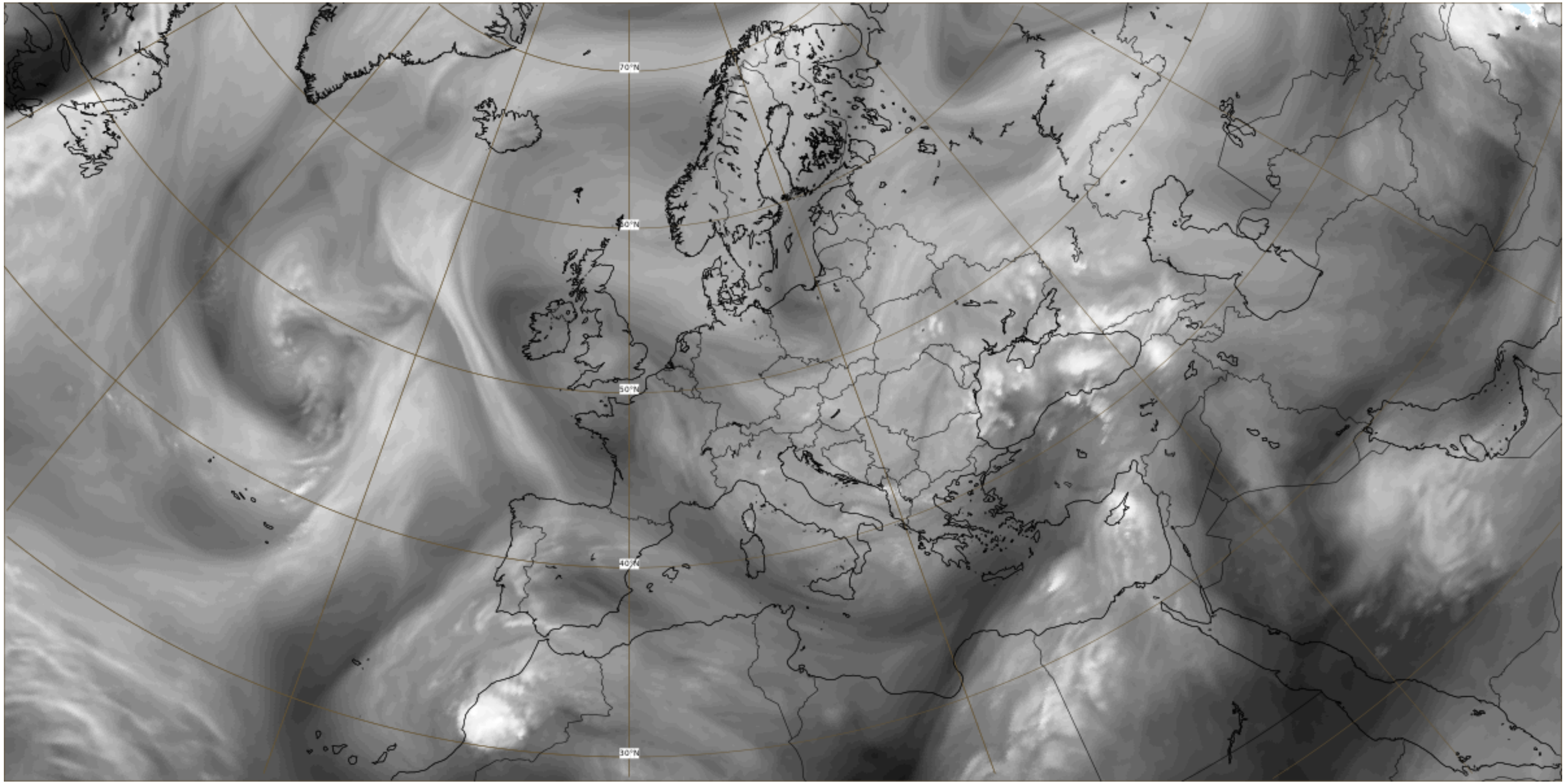
Questions

1. What is the PDF of a random variable?

2. How can one describe the “center” of a distribution? And its “dispersion”?

3. Considering 2RV X and Y, what does Cov[X,Y] quantify?

Phenomenon with random component as well as spatial coherence



Water vapor simulated by ECMWF IFS model

Matheron introduced the concept of “regionalized variable”:
spatial correlation together with high irregularity of detail.

Random function

Generalization of the concept of RV.

Random function Z → Z(w,x) = V (ReV = realization of a RF)

V: random variable
w: realization
x: vector of space coordinates over domain D

If dim(x) = 1 → stochastic process
If dim(x) > 1 → random field

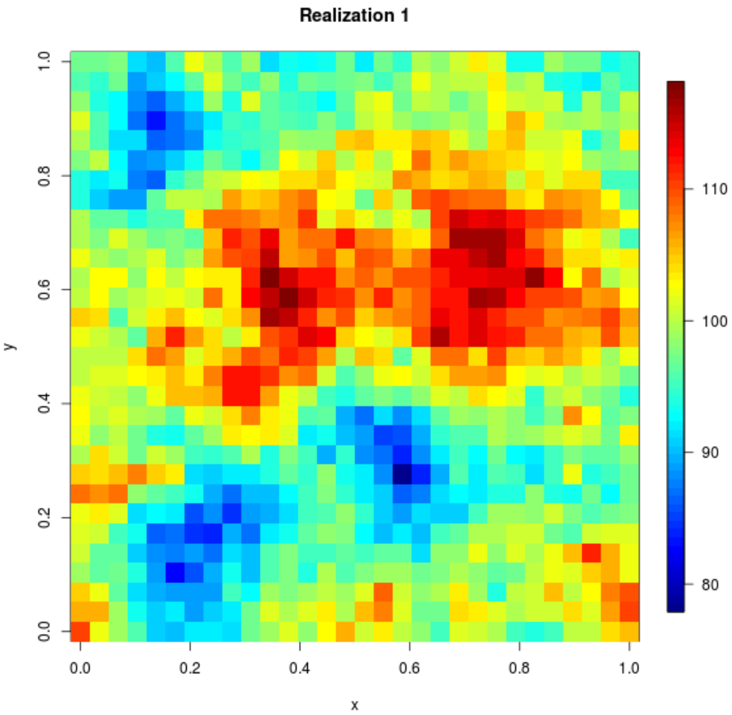
Random function (RF) can be described by its spatial distribution.

Moments are defined at given x, as for a RV:

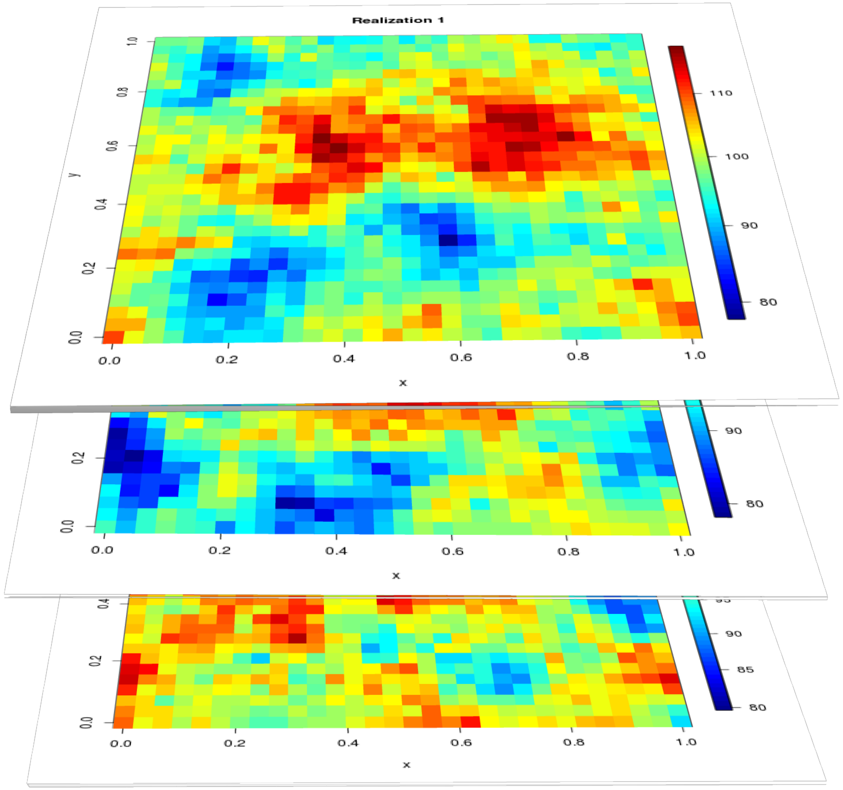
$$\mathrm{E}\left[Z(x)\right] = \mathrm{E}[V] \qquad M_n\left[Z(x)\right] = M_n[V]$$

Ex of random function

Mono-realization



Multi-realization



Stationary random functions

Strict stationarity = invariance of joint distribution by any translation.
Weak stationarity (order n) = moments up to order n of RF independent of location x within the domain D.

$$M_l\left[Z(x_2)\right] = M_l\left[Z(x_1)\right]\ \forall (x_1, x_2) \in D^2 \text{ and } \forall l = 1..n$$

Second-order stationary random functions

First 2 moments of RF independent of location x within the domain D.

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

$$\mathrm{E}\left[\left(Z(x_2) - \mathrm{E}[Z]\right)^2\right] = \mathrm{E}\left[\left(Z(x_1) - \mathrm{E}[Z]\right)^2\right]$$

$$\mathrm{Cov}\left[Z(x_2), Z(x_1)\right] = C(x_2 - x_1)$$

Isotropy

If Z is a 2^{nd}-order stationary isotropic RF

$$\mathrm{Cov}\left[Z(x_2), Z(x_1)\right] = C(||x_2 - x_1||)$$

Covariance only depends on norm of $(x_2 - x_1)$, not on the direction.

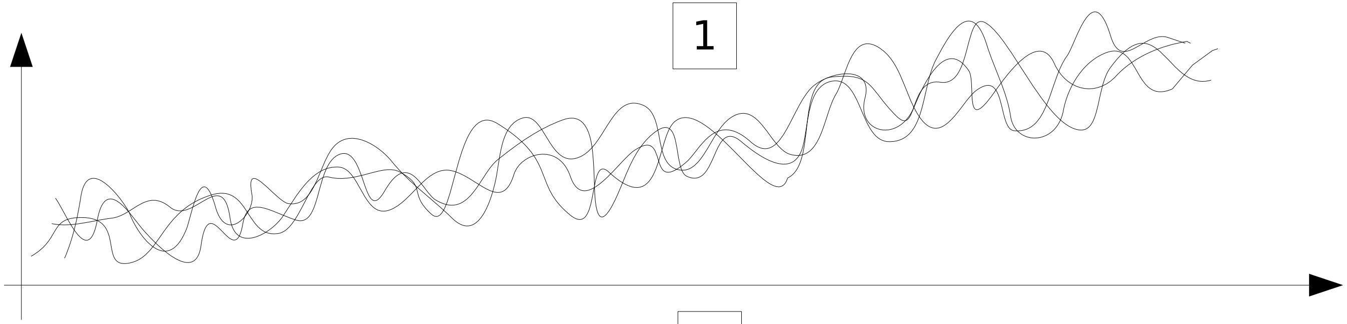
Counter-examples to illustrate these definitions

a) Stationarity in mean and variance, not in covariance

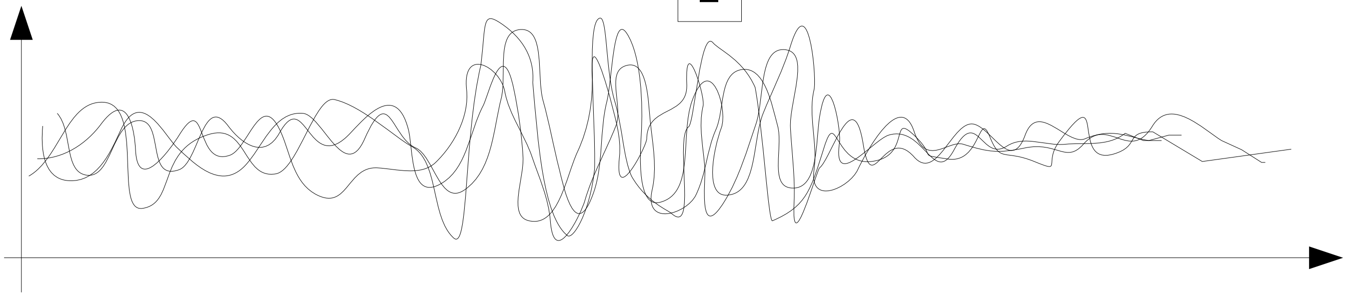
b) Stationarity not in mean, not in variance

c) Stationarity in mean, not in variance

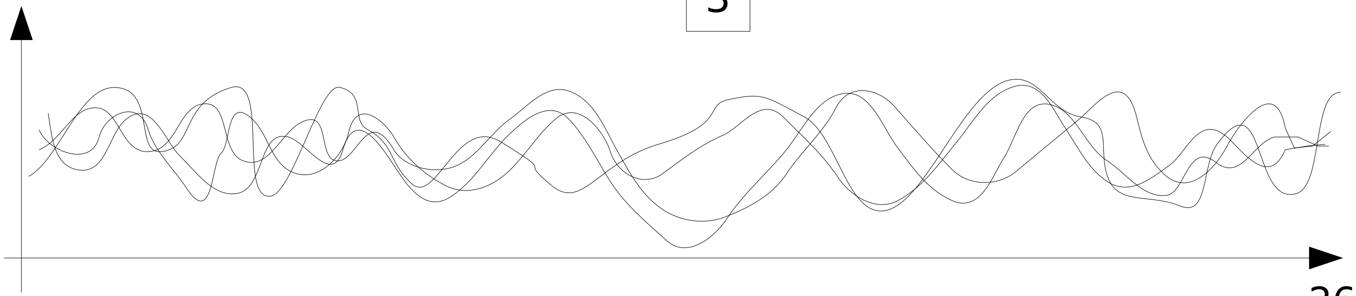
1



2



3



Ergodicity

Definition: if Z is an ergodic stationary RF $$\lim_{|V|\to\infty} \frac{1}{|V|} \int_V Z(\omega, x)\, dx = \mathrm{E}[Z(\omega)]$$

(1st order - mean)

Sufficient condition (1st order): $$\lim_{|V|\to\infty} \frac{1}{|V|^2} \iint_{x_1, x_2 \in V^2} \mathrm{Cov}(Z_{x_1}, Z_{x_2})\, dx_1 dx_2 = 0 \quad (1)$$

Sufficient condition (2nd order): $$\lim_{|V|\to\infty} \frac{1}{|V|^2} \iint_{x_1, x_2 \in V^2} [\mathrm{Cov}(Z_{x_1}, Z_{x_2})]^2\, dx_1 dx_2 = 0 \quad (2)$$

(2nd order - covariance)

Ergodicity

In practice: when only 1 realization of RF, it is a choice...
if domain large enough for (1) to be true → ergo. assumption OK
if domain not large enough for (1) to be true → ergo. assump. not OK

Stationarity ≠ ergodicity: stationarity does not imply ergodicity.
ex: $Z(\omega,x) = A(\omega)$, A being a RV.

Summary of Chap.1

1. What is geostatistics?

- Types of data

- Motivating questions:
  (1) what are the values at locations with no measurements?
  (2) what is the error associated with the estimated values?

- Possible applications:
  (1) structural analysis, (2) interpolation, (3) stochastic simulation

2. Basic spatial statistics concepts

Random variable:
- PDF, expectation, moments, median, mode...
- Variance and covariance

Going into space:
- Random function
- Stationarity, ergodicity