EGS 2405- Geostatistics

Basic Concepts of Geostatistics

Lecturer: Mr H. Kipkulei, hkipkulei@jkuat.ac.ke
Technologist:TBC

Outline

- Basic Concepts of Geostatistics
- Probability Theory Review
- Spatial Analysis
- Experimental Variogram
- Variogram Modelling
- Geostatistical Estimation (Kriging, CoKriging, Collocated CoKriging, Cross Validation, Block Kriging, Indicator Kriging, Simple Kriging)
- Geostatistical Simulation (Unconditional, Conditional)

Introduction

Geostatistics

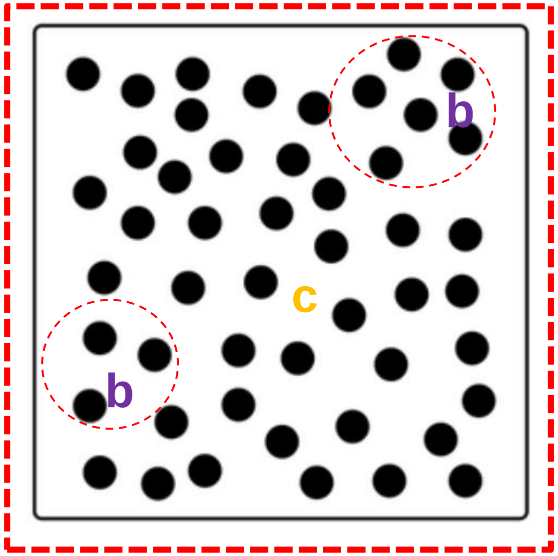
Daniel Krige in the 1950’s

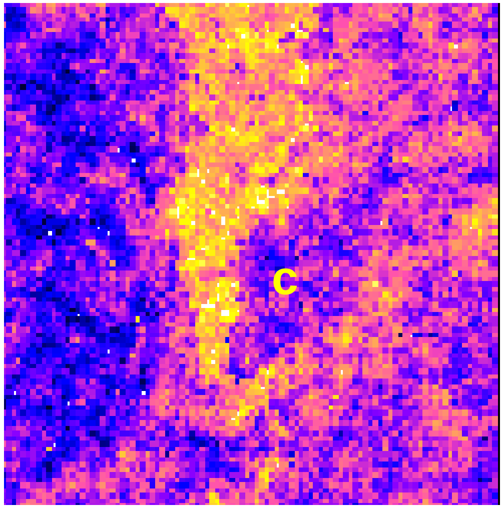
mining and petroleum industries

Georges Matheron in the 1960’s

Introduction

a





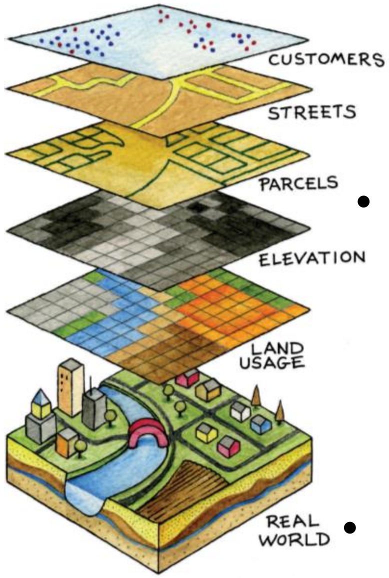
a

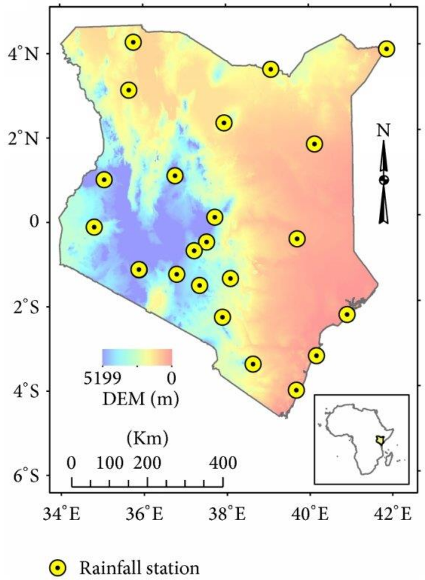
b

GS studies spatial/temporal phenomena capitalizes on spatial relationships to model possible values of variable(s) at unobserved locations

Why Geostatistics

- The environment **continuous**, but we can only measure properties at finite locations.
- Models **variability** better



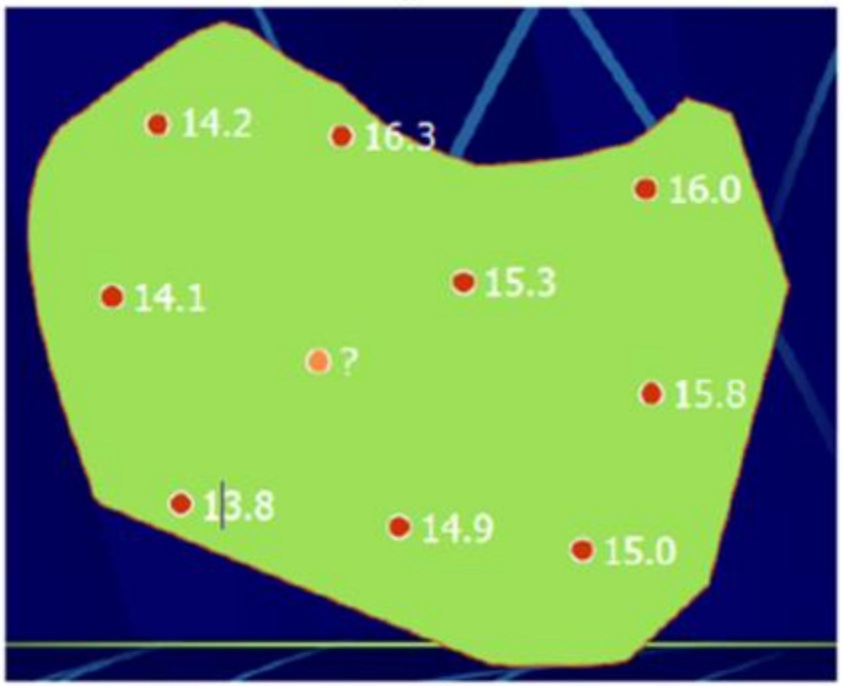


Source: Charles Onyutha

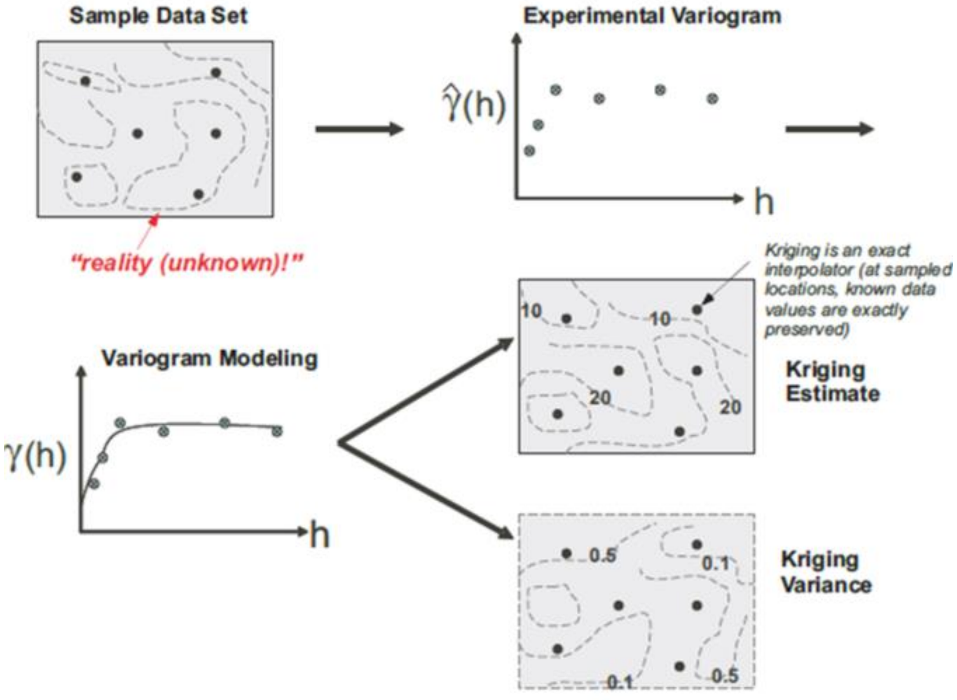
- Provides a framework to integrate **hard** and soft data

- Geostatistical methodologies are **repeatable/replicable**

Spatial prediction



(a) Spatial prediction using sampled points.



(b) Geostatistical estimation workflow.

Geostatistics versus classical statistics

Classical statistics	Geostatistics
Based on linear sum of data, all of whom carry the same weight	Rely on spatial models
Requires no assumptions about the nature of the variable itself	Assumes that the variable is random and the outcome of one or more random processes
Assumes independent observations	Observations are dependent (location is important)
Spatial correlation not included/Location is irrelevant	Considers both distance and spatial correlation
Correlation estimated from a scatter plot	Spatial Correlation modelled by semivariogram

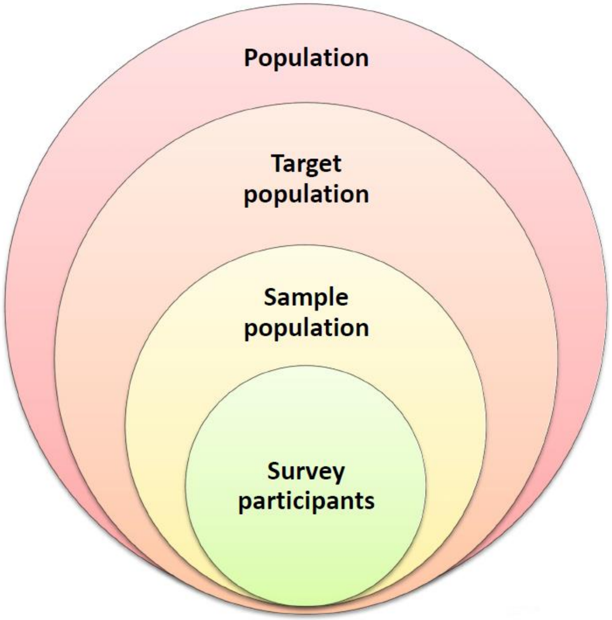
Geostatistical mapping processes

1. Design the **sampling** and **data processing**

Sampling: Selection of a subset of observation from a larger population

- Time and financial cost usually don’t allow to survey the whole (target) population
- Draw a **sample** for your analysis

But: Extrapolations (generalizations) are only valid if the sample is **representative** for the target population!



Sampling process

1. Define the target population
2. Identify the sampling frame
3. Choose the sampling method
4. Determine the sample size
5. Contact interviewees

Source: Shao and Zhou, 2007

Categories of sampling

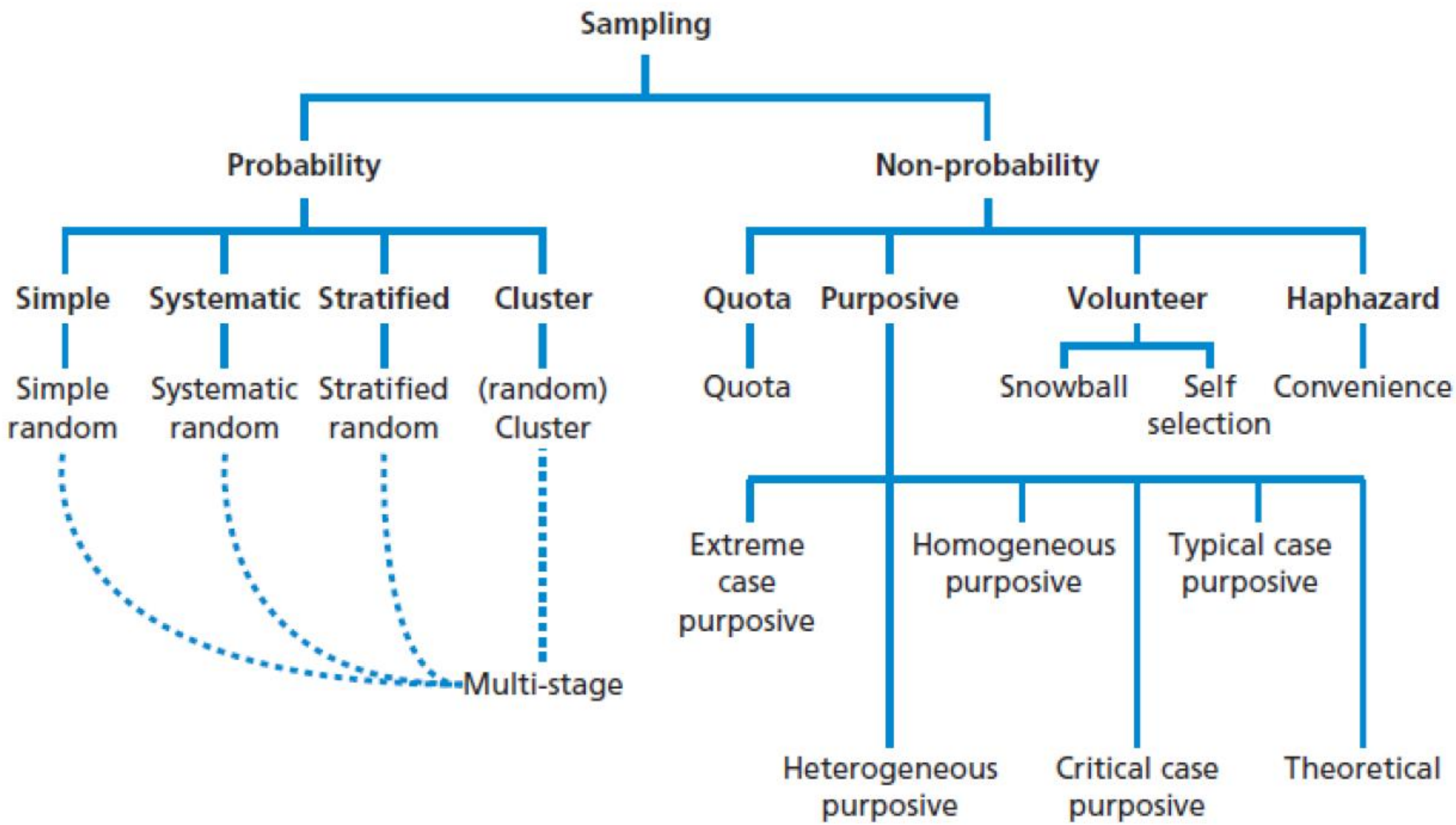


Illustration: https://research-methodology.net/

Probability vs Non-probability sampling

Probability sampling

- Every member of a population has a (known) chance of being selected.
- The probabilities of selection are based on overall population characteristics

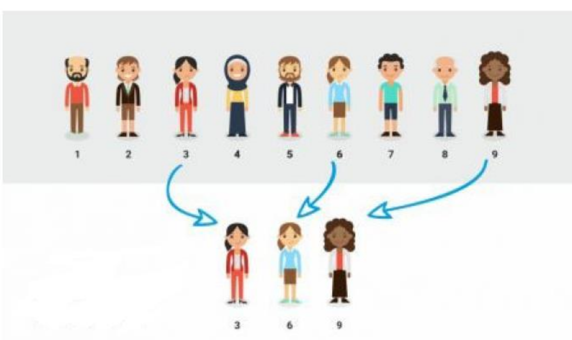
Non-probability sampling

- The odds of unit being selected into the sample cannot be calculated.
- Representativeness of the sample (i.e. sampling quality) relies on the subjective judgement of the researcher.

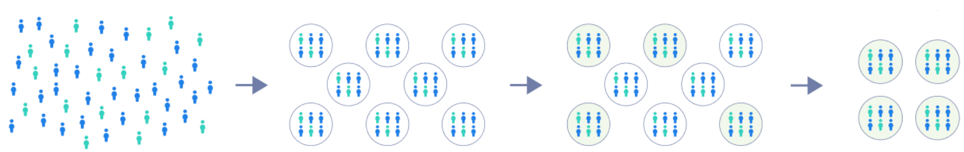
Types of Probability sampling

Simple random sampling

Stratified random sampling



Systematic sampling



Clustered sampling

Unequal probability sample

Non-probability sampling

Sampling method where each unit's probability of being sampled is unknown.

- Resulting sample often not representative for population
- Mainly used for marketing purposes
- In research mainly useful to pretest questionnaires

<u>Types:</u>

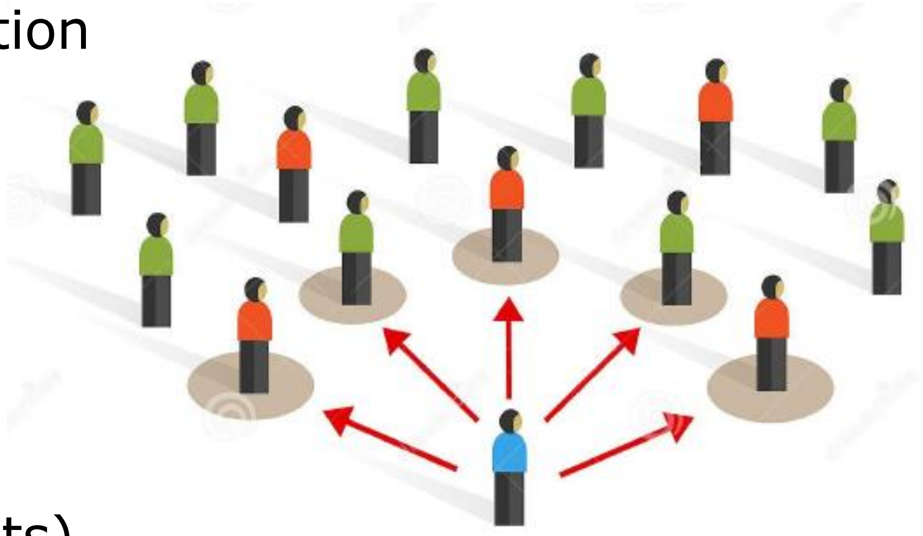
- Quota sampling
- Opportunity (convenience) sampling
- Snowball sampling

Quota sampling

- Population is classified according to certain characteristics (e.g. gender), similar to stratified sampling.
- Quotas are assigned to each group (e.g. 10 men and 10 women)
- Selection is based on accessibility instead of a randomization process.
- Representativeness is not secured
- Example: Customer survey at supermarket

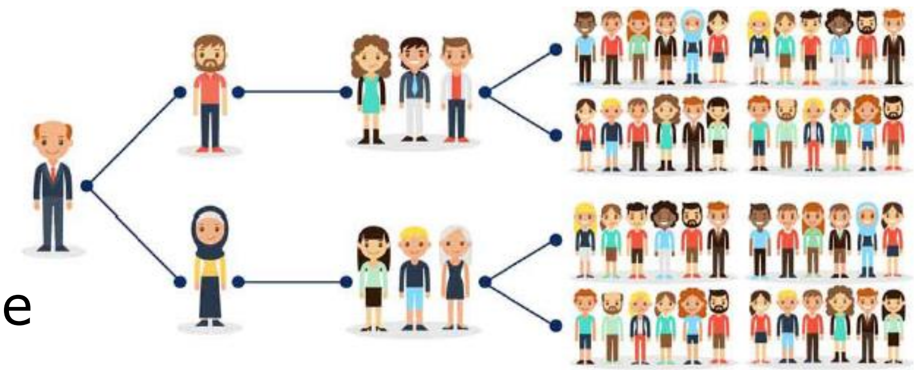
Opportunity (convenience) sampling

- Selection of interview partners based on **accessibility** only
- No randomization or stratification
- Relative less costs and organizational challenges
- Example: Pretesting of questionnaire, theoretical experiments (e.g. with students)



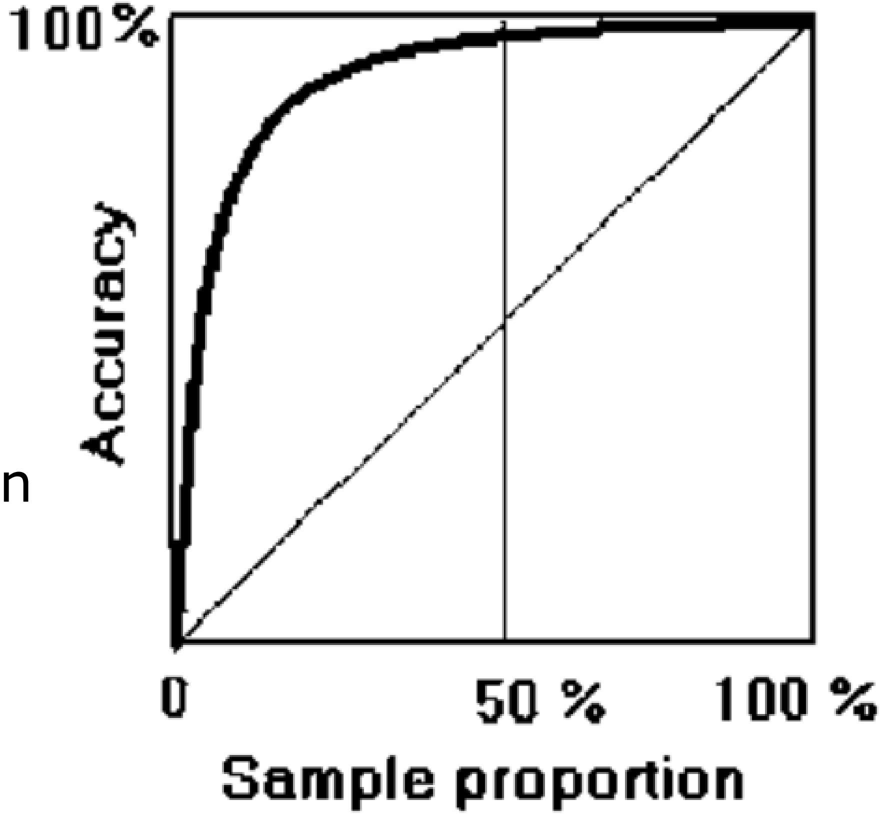
Snowball sampling

- Useful if relevant group makes up small proportion in the population
- Researcher may take interview from one key person from the company and that person provides contacts of other people in the company to speak with
- More qualitative information about organizational structure is required



What to consider when drawing a sample

- Small sampling error
- Costs for data collection are minimized
- Systematic bias is controlled
- Spread/Variability of population
- Practicality
- Information already known



Accuracy growth relative to sample size

Systematic bias

- It should be minimized by adopting a proper study design
- The causes of the bias include;
  - ❖ **Inappropriate** sampling frame Under coverage
- **Defective** measuring device

Properties of a good sample design

- Result in a truly representative sample
- Lead to only a small sampling error
- Be cost effective
- Be one that controls systematic bias
- Be one such that the results of the sample study can be applied for the population with a reasonable degree of confidence.

Geostatistical mapping processes

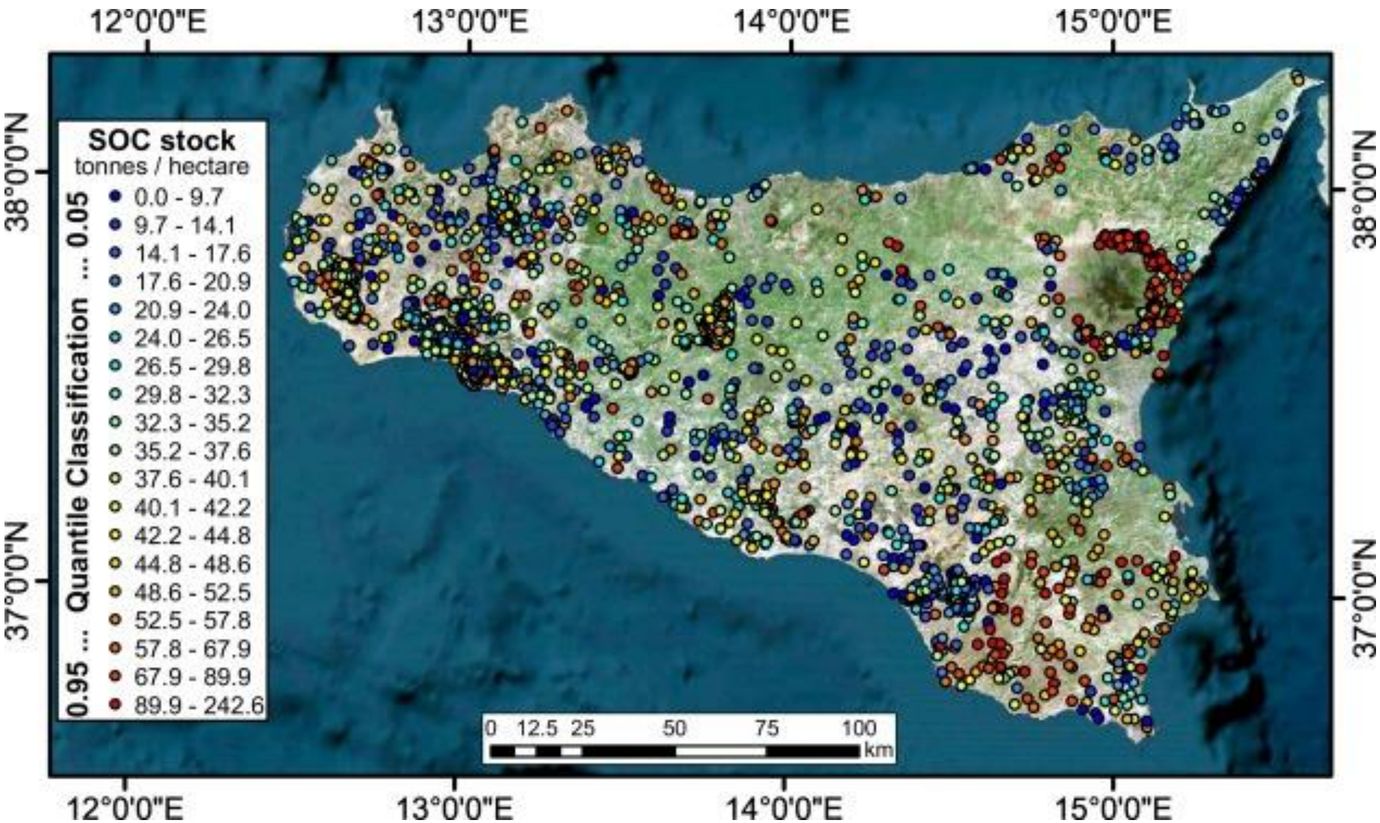
2. **Collect** field data and do **laboratory** analysis

Source: BRACED

Source. Geplus

Geostatistical mapping processes

3. **Analyse** the point's data and estimate the model



SOC

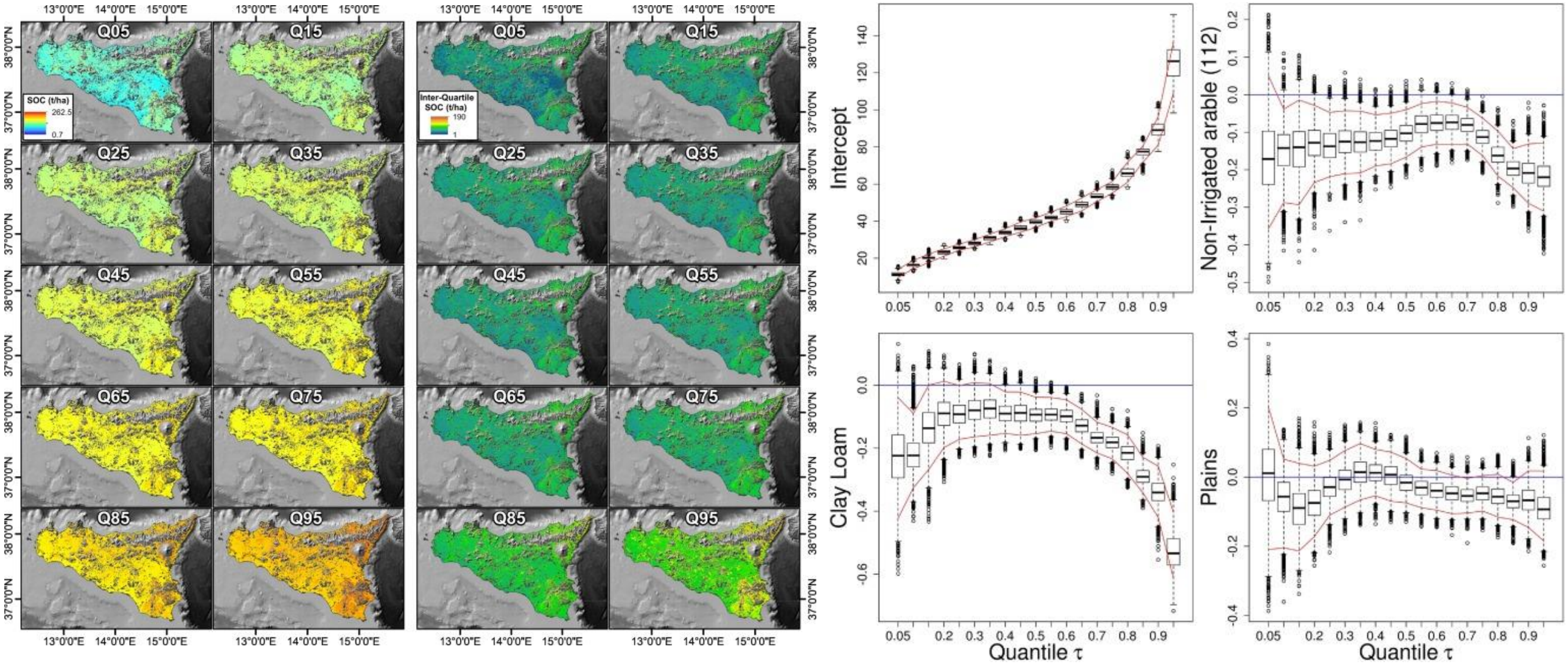
&

Predictors

https://doi.org/10.1016/j.geoderma.2017.12.011.

Geostatistical mapping processes

4. **Implement** the model and **evaluate** its performance



Source: Mohammad et., al

Geostatistical mapping processes

5. **Produce** and **distribute** the output geoformation

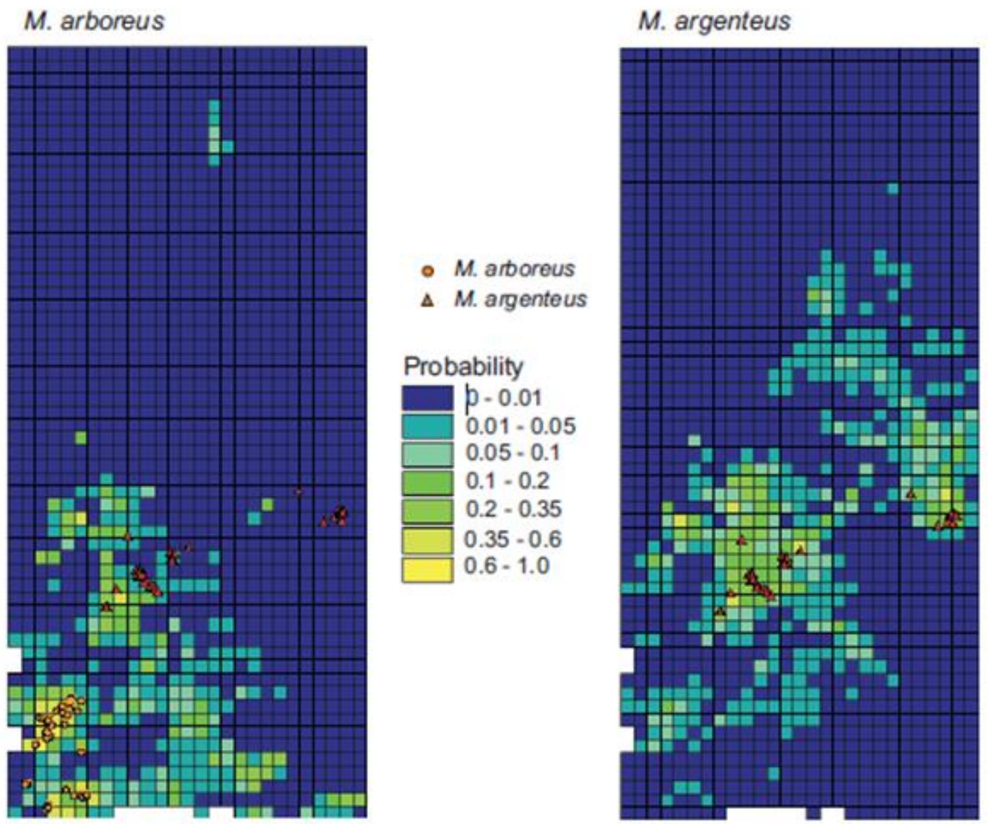
Questions of interest for every Geostatistician

- How does a variable vary in space-time?
- What controls its variation in space-time?
- Where to locate samples to describe its spatial variability?
- How many samples are needed to represent its spatial variability?
- What is a value of a variable at some new location/time?
- What is the uncertainty of the estimated values?

Environmental variables

- Environmental variables are **quantitative** or **descriptive** measures of different environmental features.
- They belong to different domains, ranging from
  - ❖ **Biology** (distribution of species and biodiversity measures),
  - ❖ **Soil science** (soil properties and types),
  - ❖ **Vegetation science** (plant species and communities, land cover types)
  - ❖ **Climatology** (climatic variables at surface and beneath/above), hydrology (water quantities and conditions) e.t.c

Environmental variables (Geostatistical mapping example)



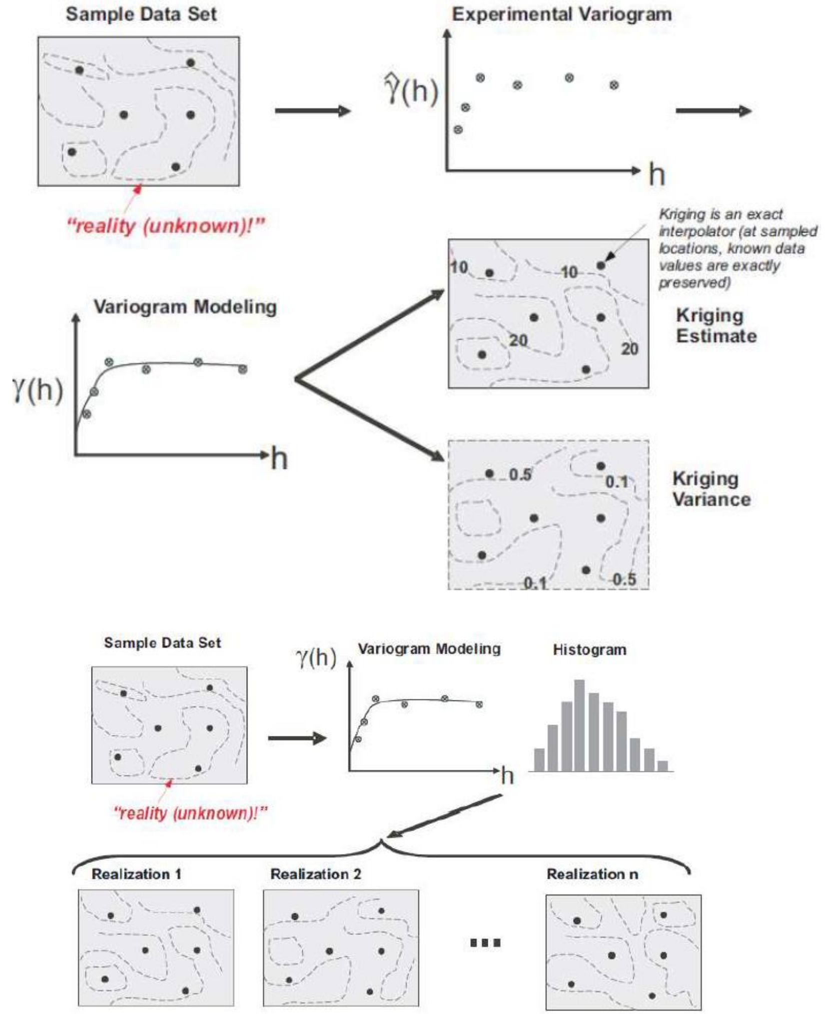
Geostatistical mapping of occurrence of sister (plant) species.
Latimer et al. (2004).

Goal of Geostatistics

predict the possible spatial distribution

Estimation

Simulation



Estimation versus Simulation

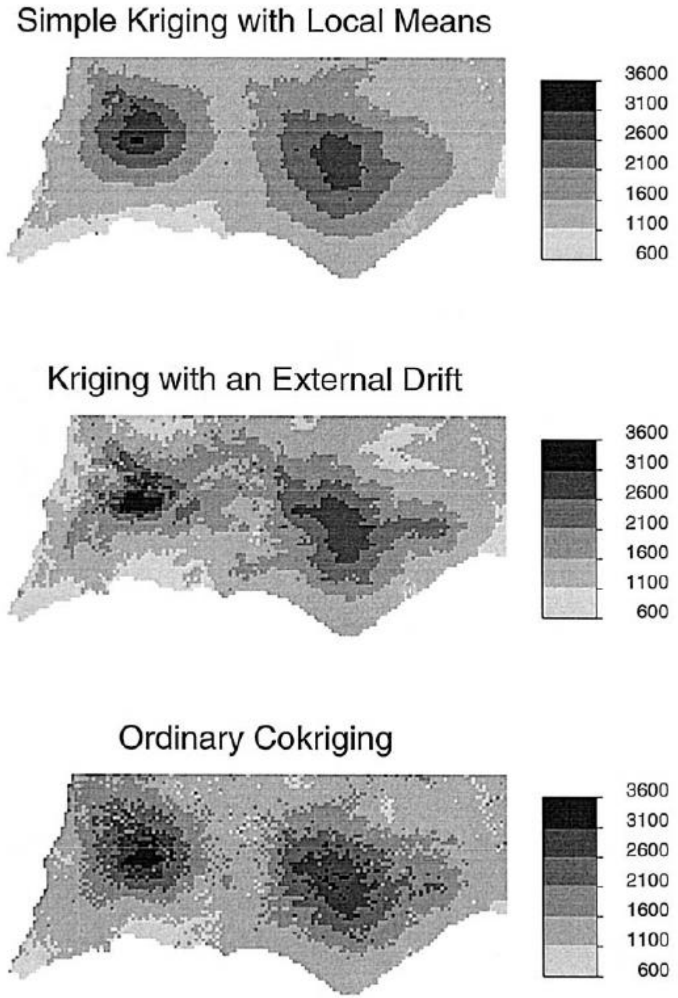
- **Estimation** - single, statistically “best” estimate (map) of the spatial occurrence is produced.
- Based on both the sample data and on a model (**variogram**) determined as most accurately representing the spatial
- Map produced by the **kriging** technique.
- **Simulation**, many equally likely maps of the property distribution are produced, using the same model
- Differences between the alternative maps provide a measure of quantifying the **uncertainty**, an option not available with kriging estimation

Geostatistics vs simple interpolation

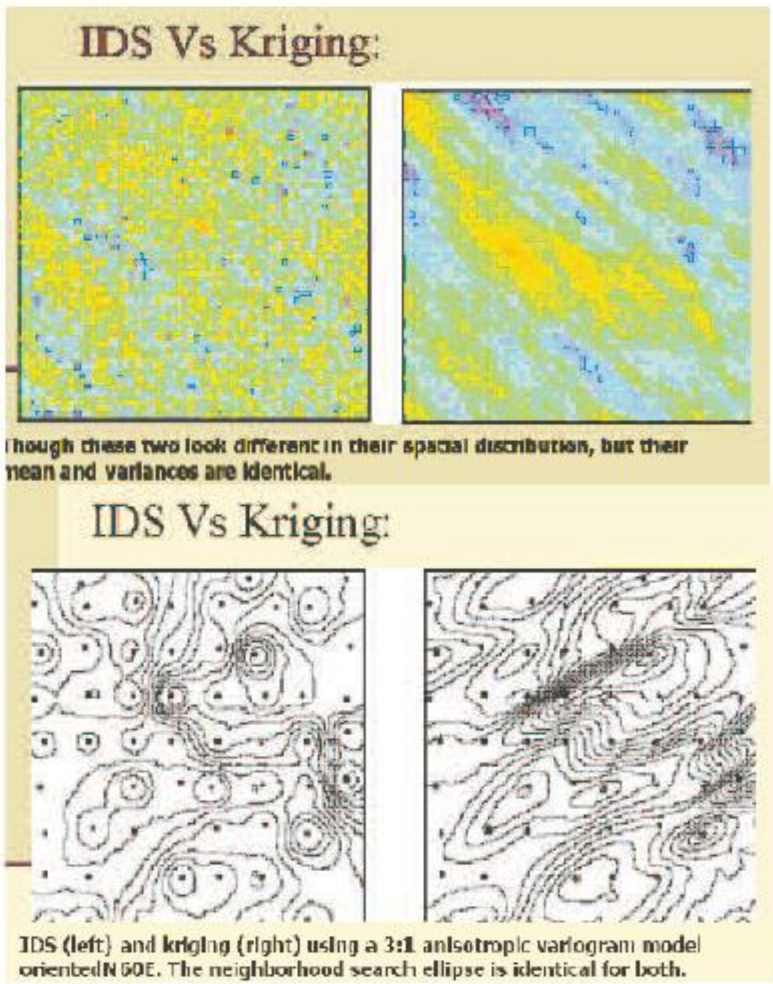
Geostatistics

- Interpolation based on statistical relations between the value and distance
- Using the example of altitude, to compute the value of altitude between two points, you could take into consideration the Earth curvature

Source: Goovaerts



Geostatistics vs simple interpolation



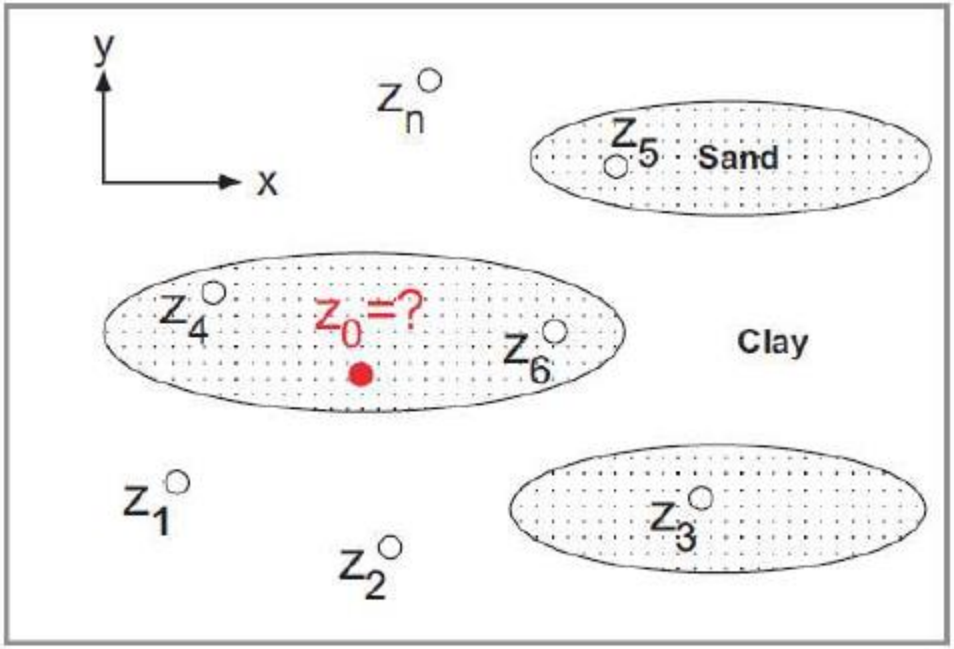
Are the figures similar?

Are there commonalities?

Related trends?

Simple Interpolation vs Geostatistics

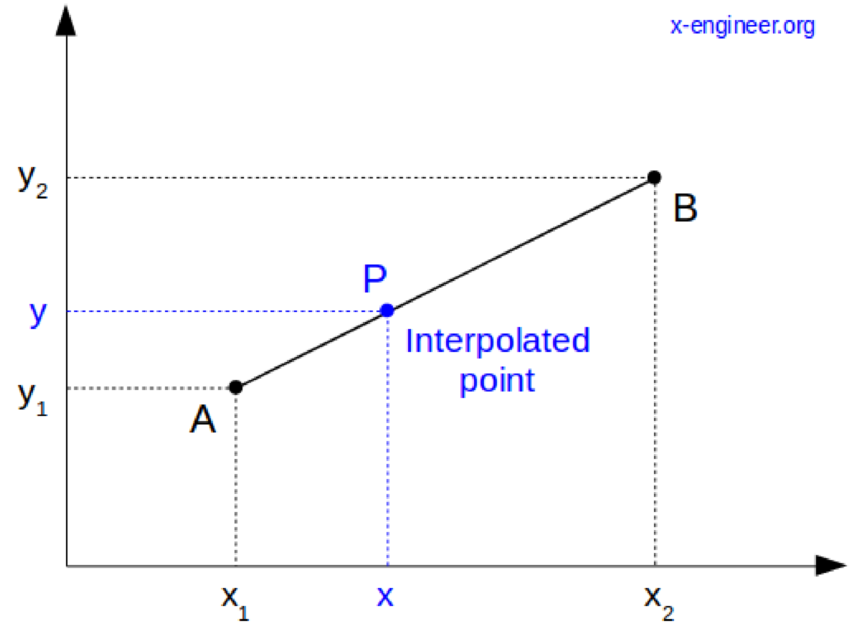
- We wish to estimate a property at an unsampled location Zo
- For example, we know **permeability** at n sampled locations; we wish to estimate the permeability at an unsampled location, Zo



Why not just use simple interpolation?

Simple Interpolation

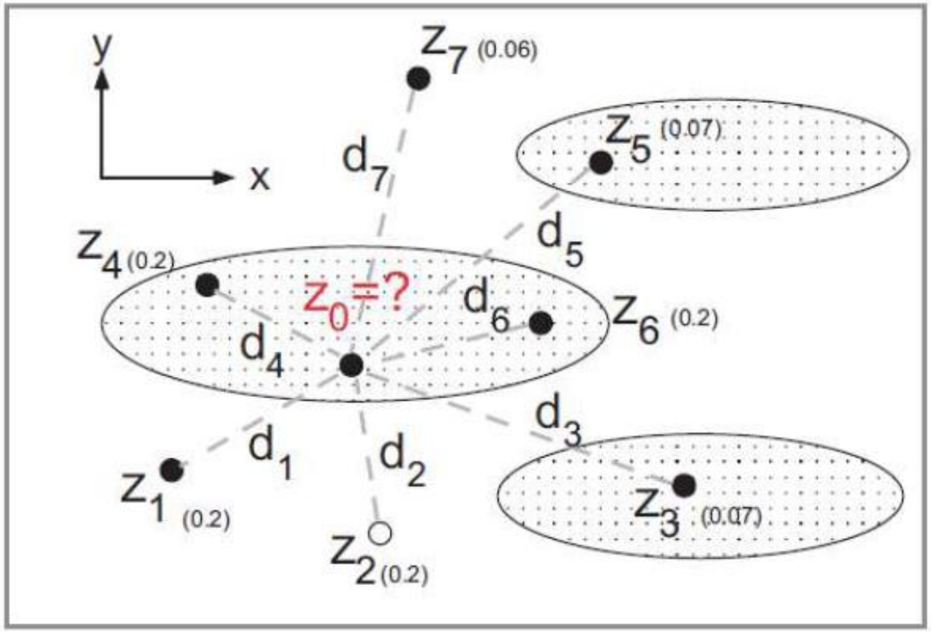
- A **mathematic interpolation** based only on the **value** and **distance**.
- For example, to compute the value of altitude between two points, you can apply a simple method of linear interpolation.
- **Ignores** the information provided by neighboring stations



Simple Interpolation

- SI yields the **largest** prediction errors in most situations
- It assumes that the outcome values are **independent** from one another

Simple Interpolation using IDW

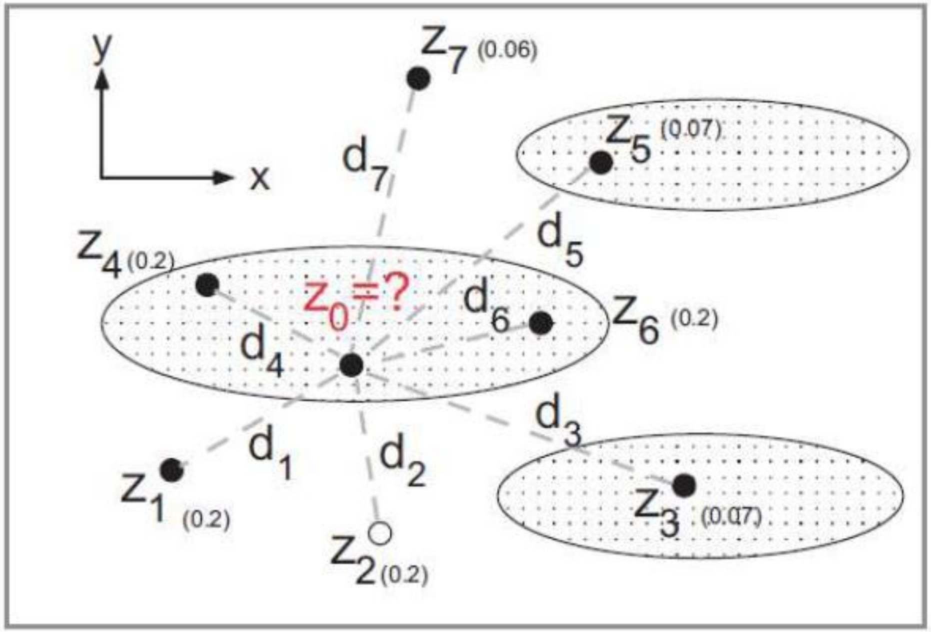


$$z_0 = \sum_{i=1}^{n} w_i z_i \qquad (estimate)$$

$$w_i = \frac{1/d_i}{\sum_{i=1}^{n}(1/d_i)} \qquad (weight)$$

- Linear estimator, i.e., Zo a weighted sum of the n known values (samples). Each weight (W_i) assigned to a known random variable Z_i is determined by the distance of the known data point to the unknown data point

Simple Interpolation using IDW

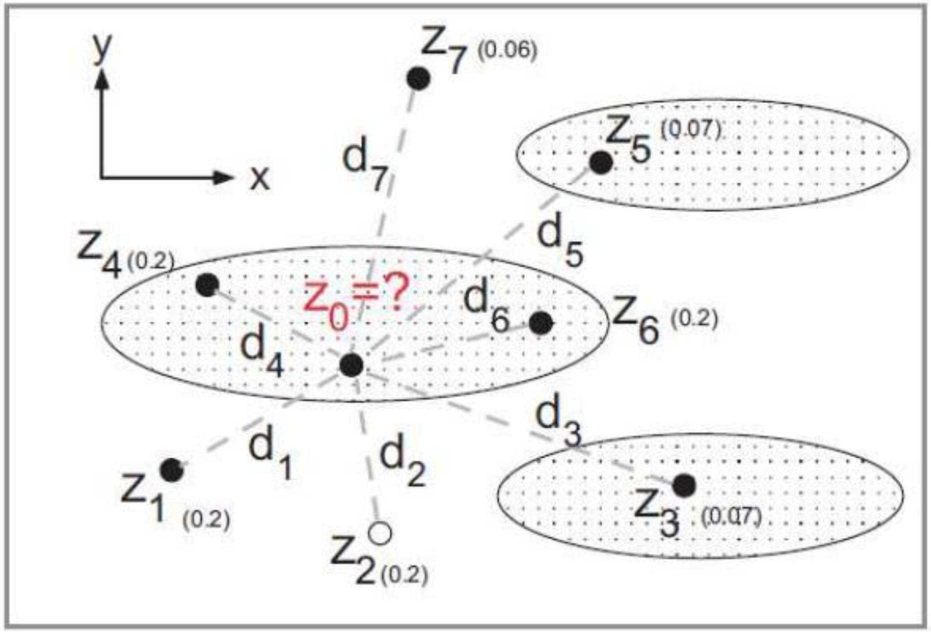


$$z_0 = \sum_{i=1}^{n} w_i z_i \qquad (estimate)$$

$$w_i = \frac{1/d_i}{\sum_{i=1}^{n}(1/d_i)} \qquad (weight)$$

- Linear estimator, i.e., Zo a weighted sum of the n known values (samples). Each weight (W_i) assigned to a known random variable Z_i is determined by the distance of the known data point to the unknown data point

Simple Interpolation using IDW

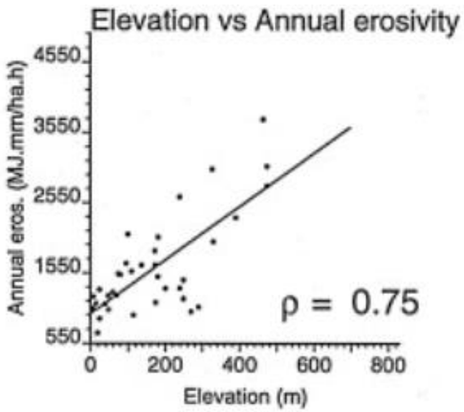


$$z_0 = \sum_{i=1}^{n} w_i z_i \qquad (estimate)$$

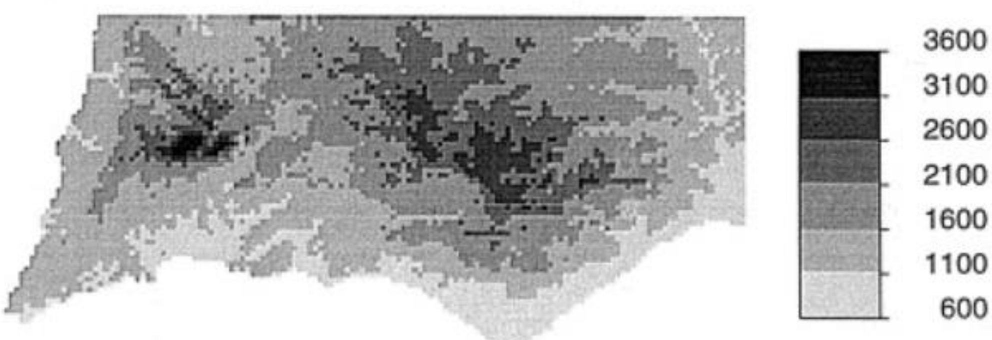
$$w_i = \frac{1/d_i}{\sum_{i=1}^{n}(1/d_i)} \qquad (weight)$$

- Linear estimator, i.e., Zo a weighted sum of the n known values (samples). Each weight (W_i) assigned to a known random variable Z_i is determined by the distance of the known data point to the unknown data point

Example of SI



Annual erosivity



- Models the **relation** between **elevation** and erosivity, e.g., using a **linear** function of the type

$$z(u)=f[y(u)]=a^*_0+a^*_1 y(u).$$

$$Z(u)= 968.2+3.8087\ y(u)$$

Source: Goovaerts

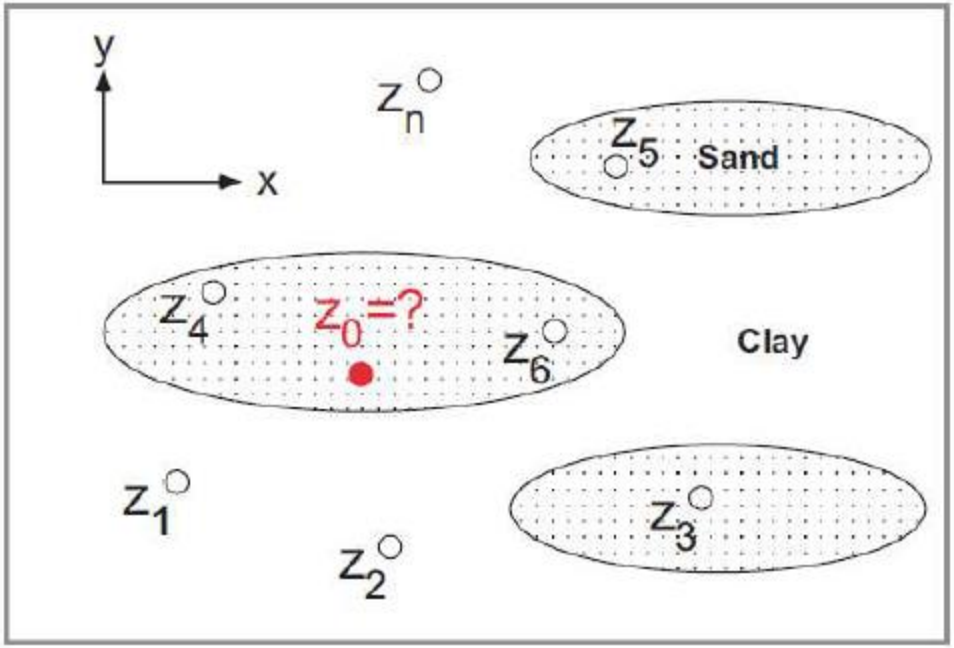
Geostatistical mapping

- It accounts for **spatial dependence (SD)** between observation. (SD- you will encounter this term regularly).
- SD is detected using a **Semivariogram** (another topic coming soon)
- SV is a measure of average **dissimilarity** between **observations** as a function of the separation vector h.

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{\alpha=1}^{N(\boldsymbol{h})} \left[z(\boldsymbol{u}_\alpha) - z(\boldsymbol{u}_\alpha + \boldsymbol{h}) \right]^2,$$

- The experimental semivariogram $\hat{\gamma}(\boldsymbol{h})$ is computed as half the **average squared difference** between the components of every data pair

Geostatistical mapping example of permeability

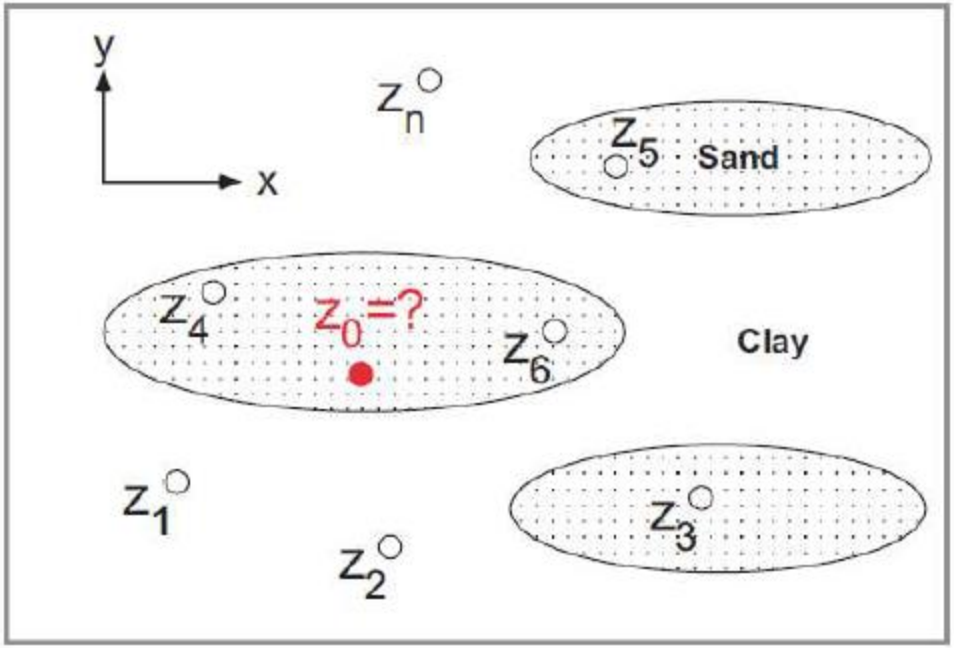


- Permeability within the elongated sand body should be more similar in the lateral direction.
- Thus, points **4** and **6** should be given **higher weights** than points 1 and 2.

Geostatistical mapping example of permeability

- Thus Geostatistical mapping comprise of three main steps:
- Examining the **similarity** between a set of sample (known) data points via an experimental variogram analysis;
- **Fitting** a **permissible** mathematical function to the experimental variogram
- Conducting **kriging** interpolation based on this function

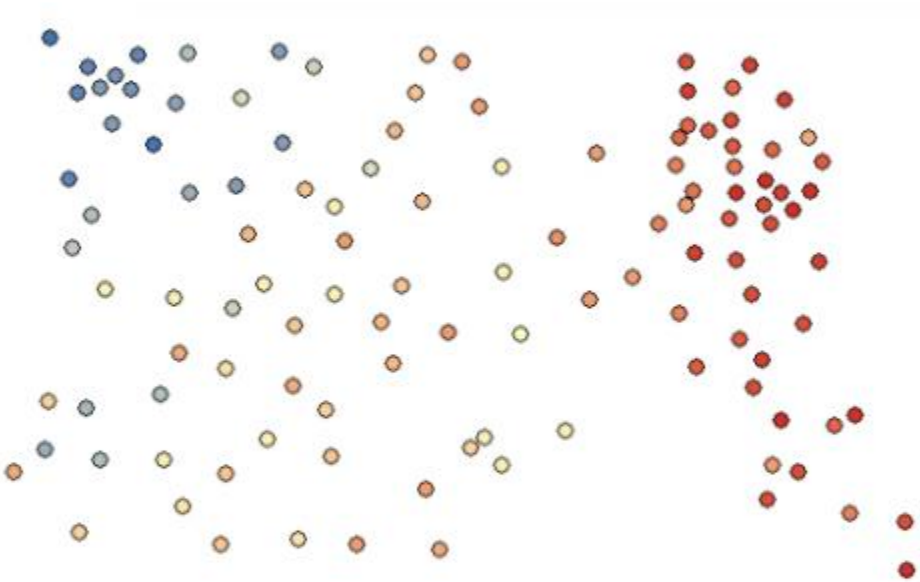
Geostatistical mapping example of permeability

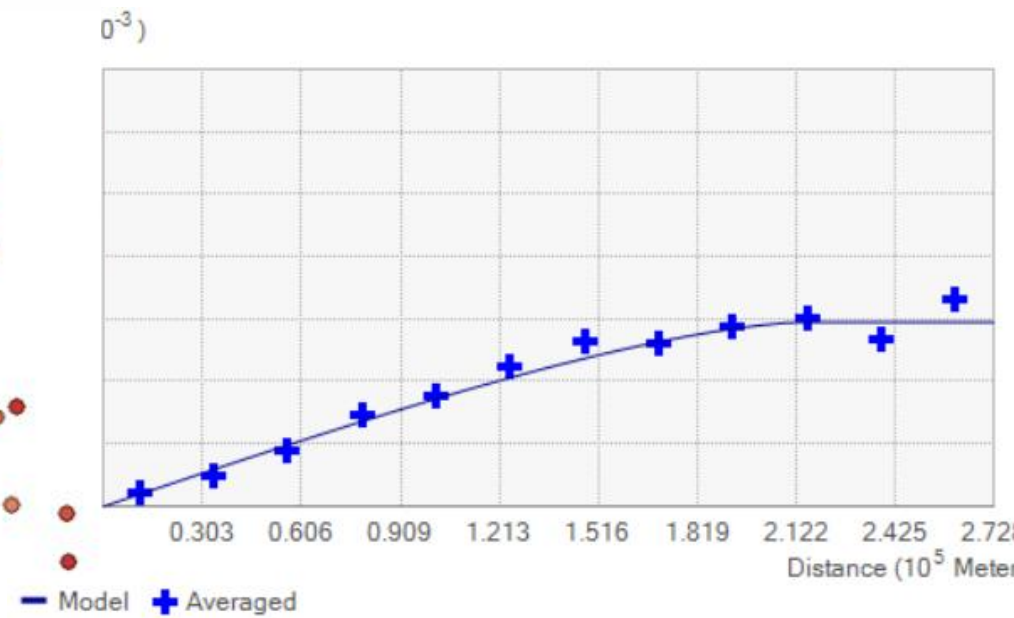


- In the above example, **the spatial correlation** will be revealed by the more similar values of Z4 and Z6 step (1)).
- It will be modeled via step (2) (variogram modeling).

- Using **kriging**, the weights assigned to points 4 and 6 will increase, while those of 1 and 2 will decrease. Total weight must sum to 1.0) (step (3)).
- In kriging, based on the new weights, a **best linear unbiased** estimate of Zo is obtained.

Geostatistical mapping



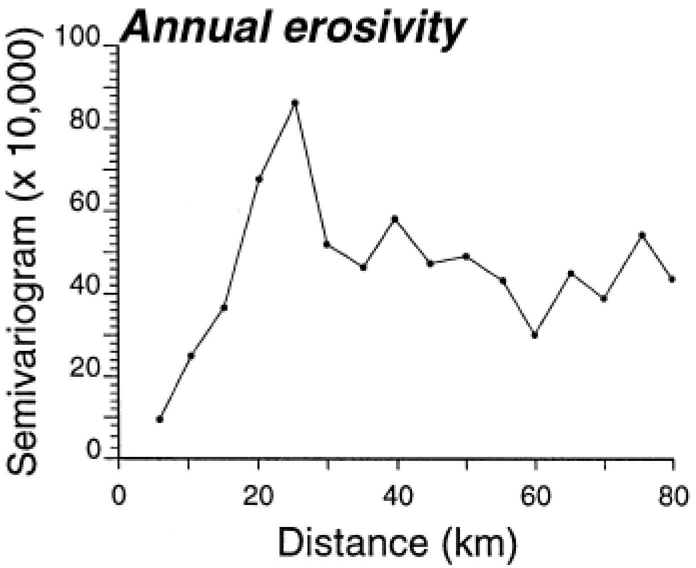


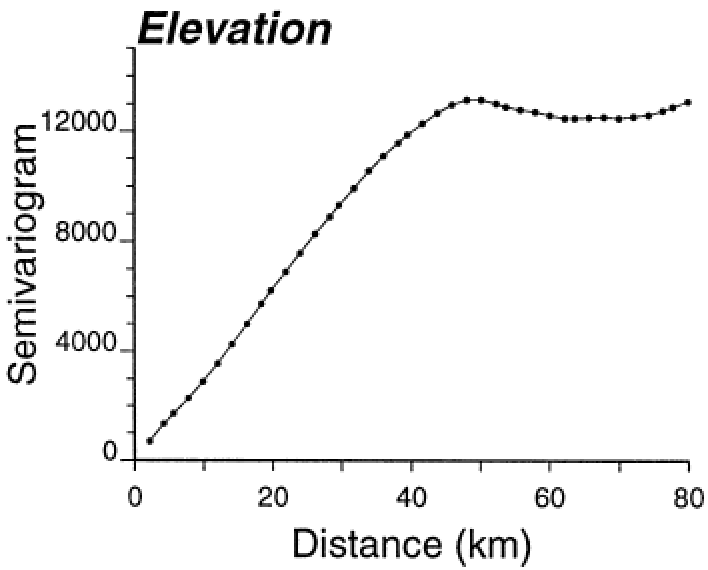
SOC Samples

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} \left[z(u_\alpha) - z(u_\alpha + h) \right]^2,$$

NB: This will be covered in detail later

Geostatistical mapping



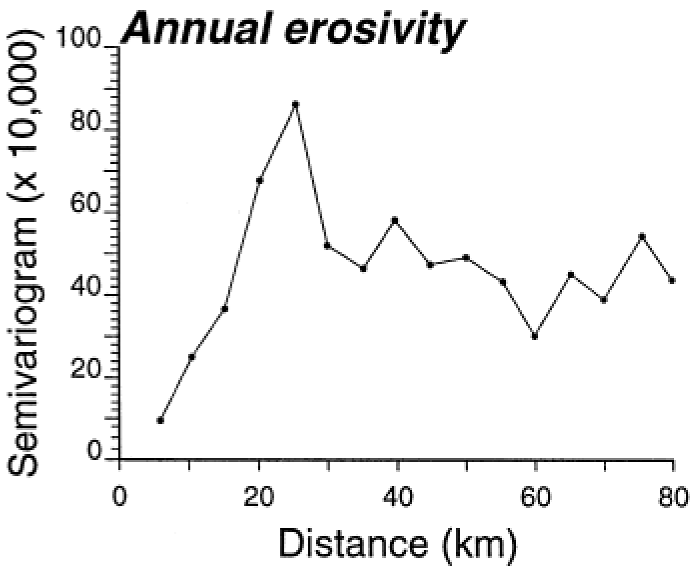


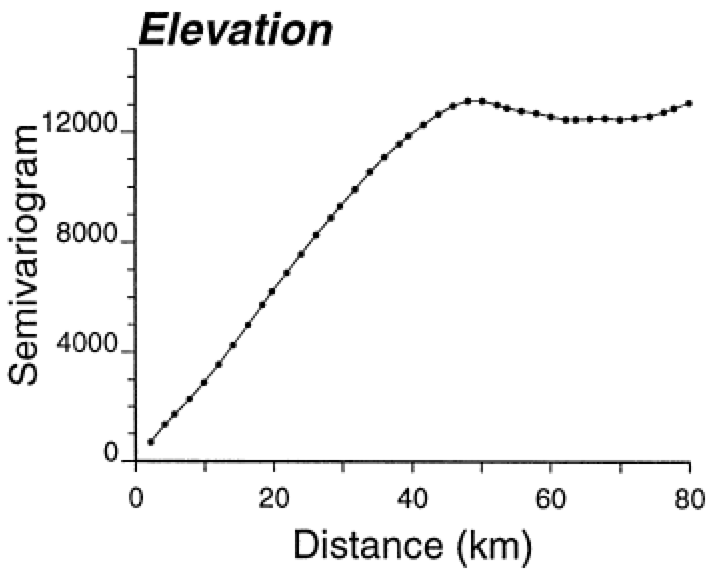
Experimental semivariograms (**Omnidirectional**) of annual erosivity and elevation

Source: Goovaerts

Omnidirectional semivariogram: One direction, few data samples

Geostatistical mapping





- Semivariogram values increase with the separation distance, reflecting our intuitive feeling that two erosivity values close to each other on the ground are more alike
- Thus their squared difference is smaller, than those further apart.

Which way? Geostatistical mapping or SI

- Given the same set of sampled data, interpolation results using Inverse Distance Squared (IDS) and kriging can look drastically **different**.
- There are situations when the sampled data are simply not good for kriging.
- Given such data --either too unreliable or **too sparse** and widely spaced to capture the spatial correlation of the variable, the conventional IDS may give just as good a result.
- The decision of which method to use is in a way **data-driven**

What next then?

- There is no accepted universal algorithm for determining a variogram/covariance model
- Most consequential decisions of any geostatistical study are made early in the exploratory data analysis (**EDA**) – Which is our next topic

Thank you for your attention! Questions?