EGS 2405- Geostatistics

Spatial correlation/dependency

Lecturer: Mr H. Kipkulei, hkipkulei@jkuat.ac.ke
Technologist: Ms. Sarah Orado

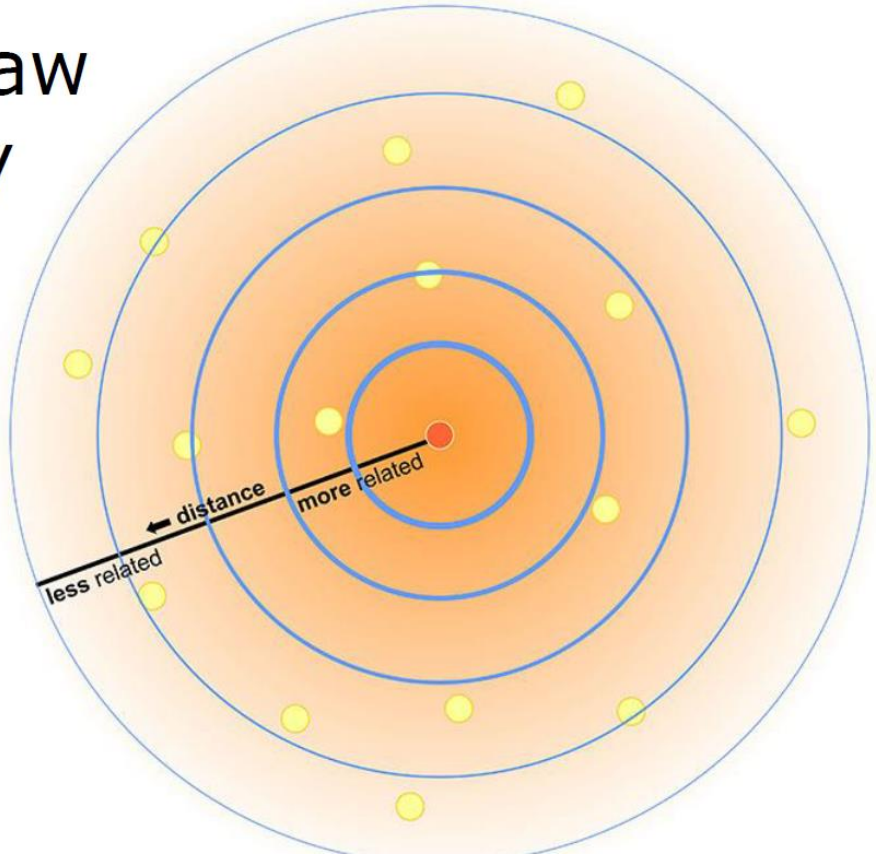
Introduction

- Value of a variable of a point in space is related to ist value at nearby points
- Recall the 1st Tobler‘s law of Geography
- Everything is **related** to everything Near things are more related
- **Distant** things are less **related**

Waldo Tobler in 2007 (Wikipedia)

Introduction

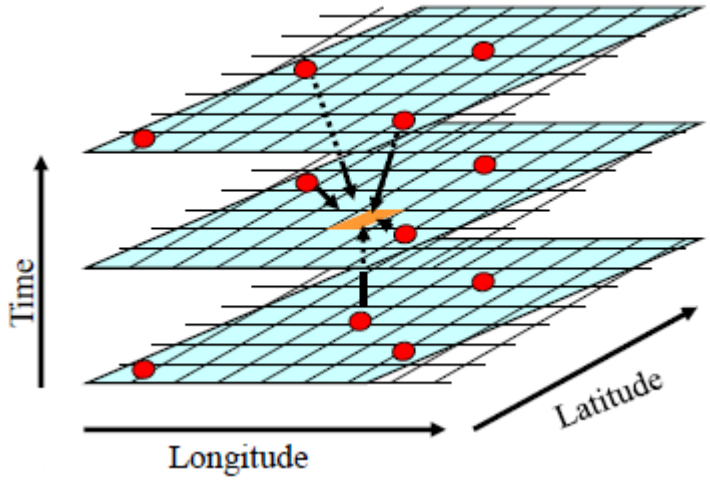
Tobler's First Law of Geography

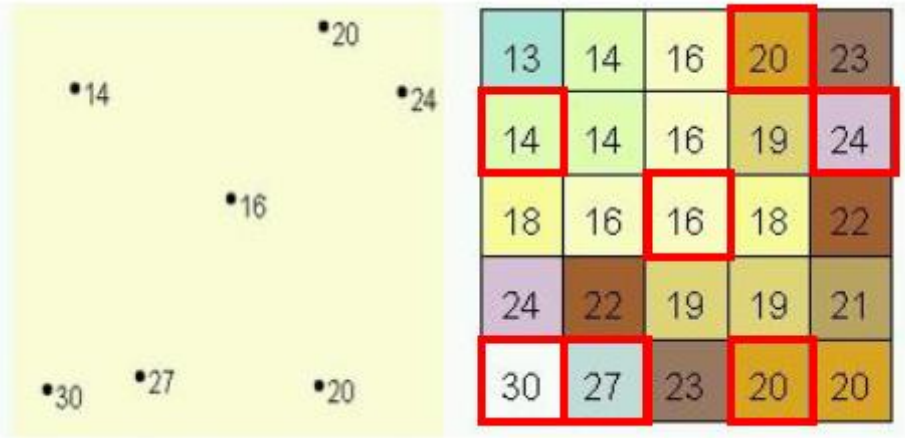


https://www.e-education.psu.edu/maps/l2_p2.html

Introduction

- Tobler's law forms the basic premise behind **interpolation**, and **near** points generally receive **higher weights** than far away points.





Introduction

- All things are related, but closer things more so!
- Covariance(y_i, y_j) is not equal to zero
  - **A variable at one location (y_i) is not independent from the same variables values at neighboring locations (y_j)**
- Spatial influence is (often) subject to a distance decay
  - In spatial statistics, we are interested in how strong the spatial influence is and over what range we can observe it

Introduction

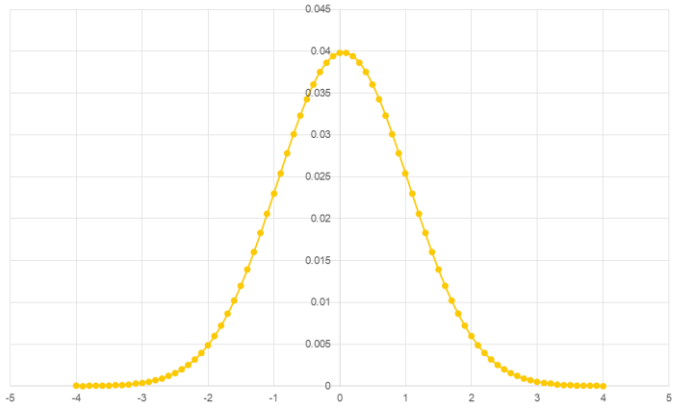
- Standard statistics cannot quantify spatial dependency

- Therefore. Components of Geostatistics adopted include;

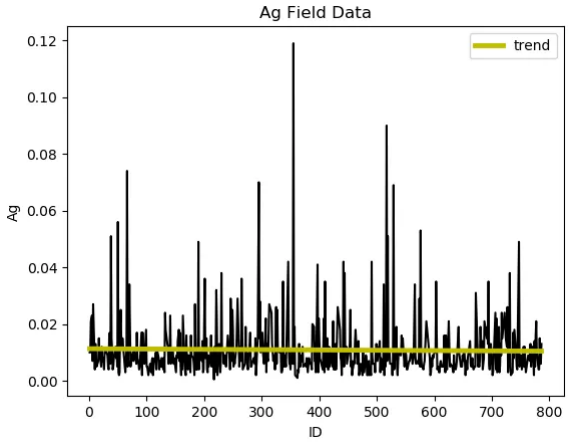
  - (**Semi)variogram analysis** - Characterization of spatial correlation.
  - **Kriging** - Optimal interpolation; generates best linear unbiased estimate at each location; employs semivariogram model.
  - **Stochastic simulation -** Generation of multiple equiprobable images of the variable; also employs semivariogram model

Introduction

- Geostatistical methods are optimal when data are
  - **Normally distributed**

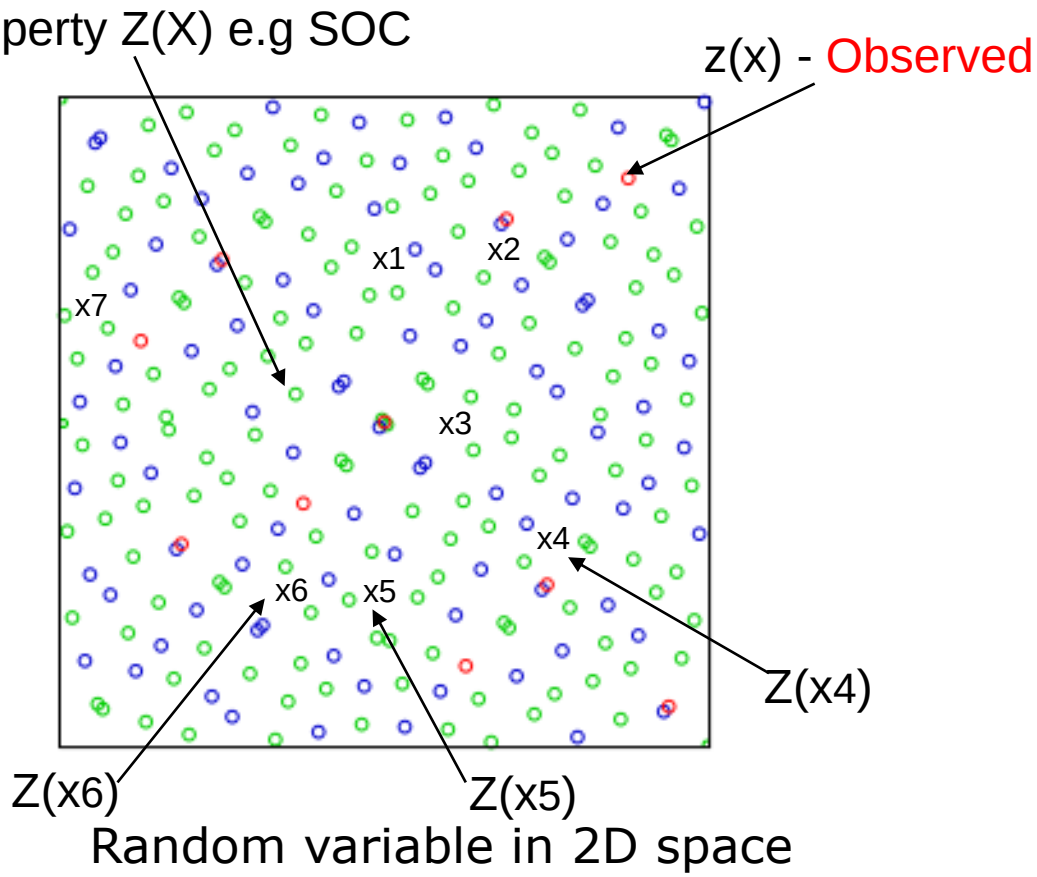


  - **stationary** (mean and variance do not vary significantly in space)



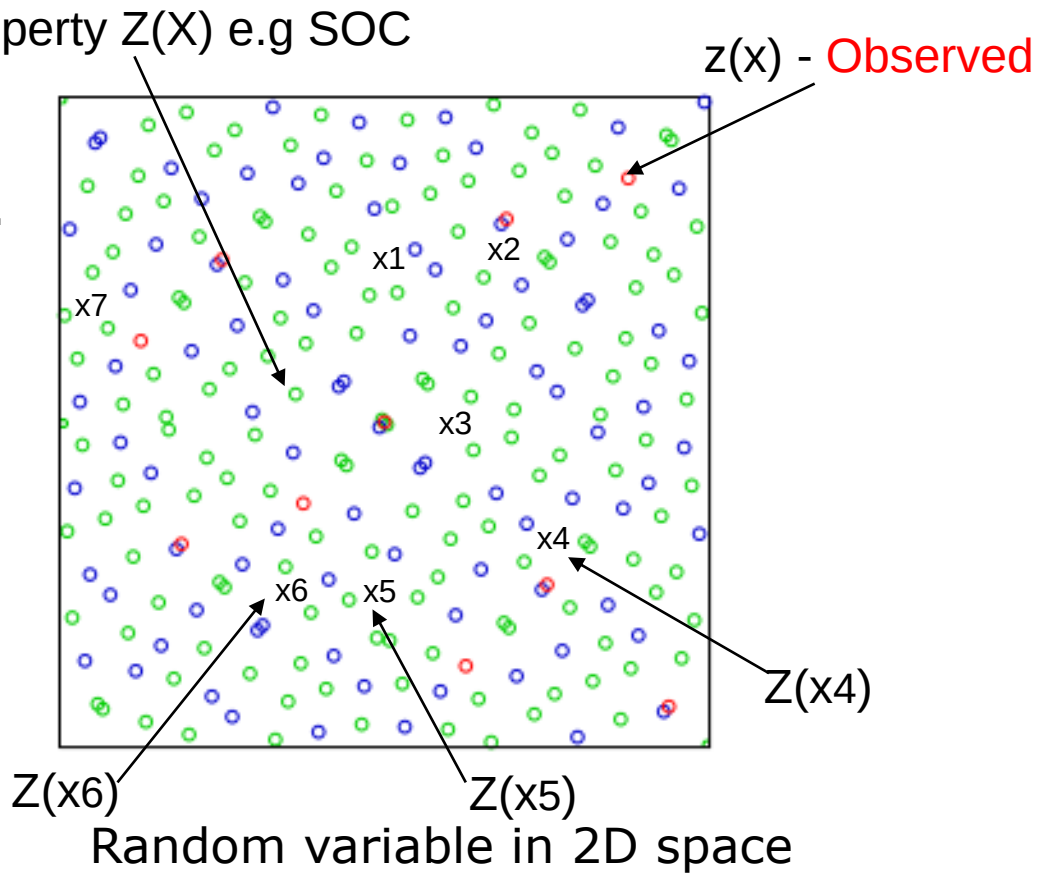
Theory of statistics

- Most spatial properties vary in such a **complex** way
- The variation cannot be defined **deterministically**.
- Therefore reliance on **stochastic** or **probabilistic** approaches in handling spatial uncertainities



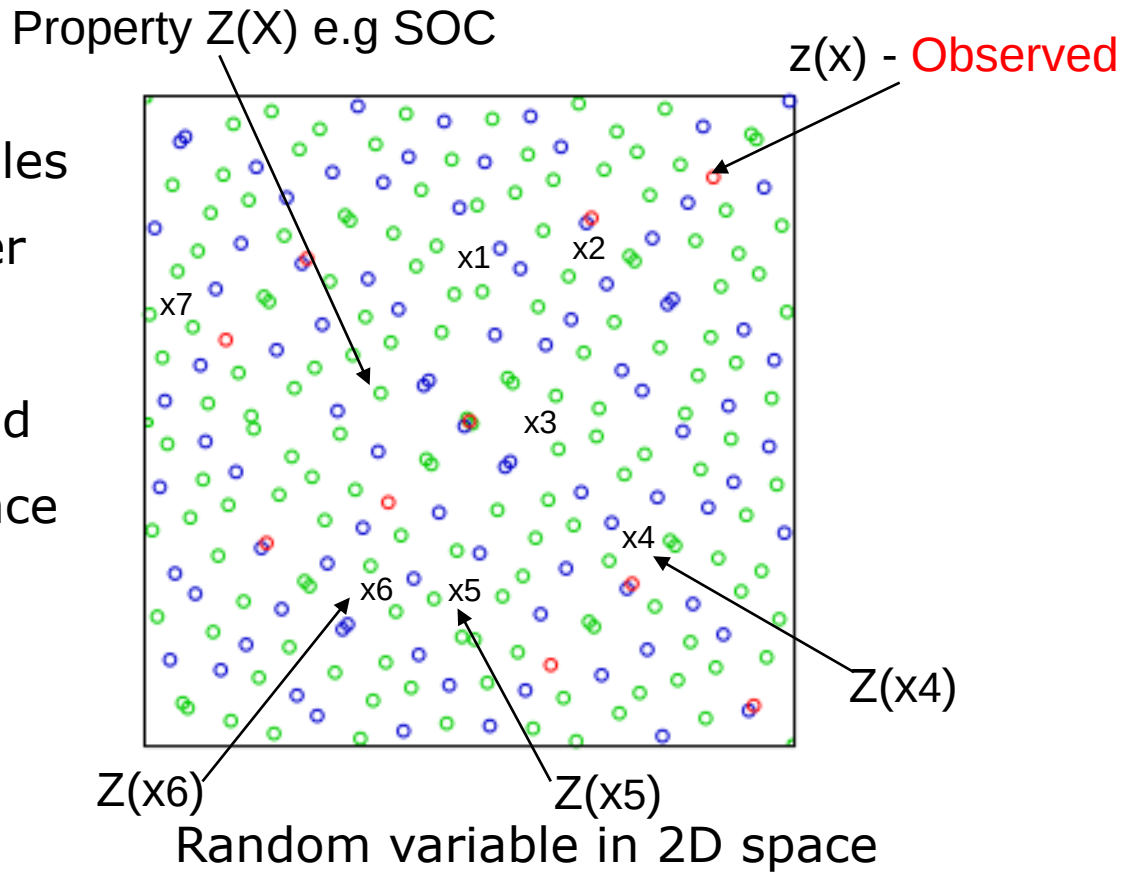
Theory of statistics

- At x, Z(x) is a random variable with a mean, μ and variance,σ^2. The set of random variables, Z(x1), Z(x2)……Z(xn) is a **random** process
- The actual value of Z observed is **<u>just one</u>** of potentially any number of realizations of the random process



Theory of statistics

- Revisiting the Tobler's law, values of regionalized variables at places near to one another tend to be related.
- So to estimating mean, μ and variance,σ2, spatial covariance can be used to describe the relation between points.



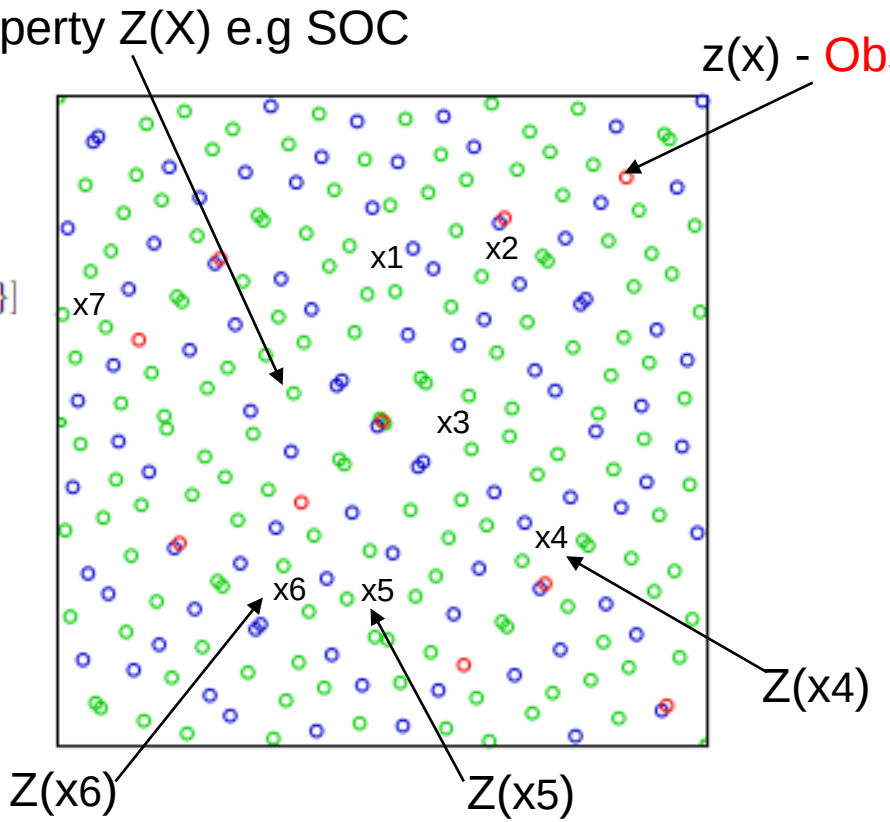
Theory of statistics

- Covariance of random variables is given as

$$C[Z(\mathbf{x}), Z(\mathbf{x}+h)] = E[\{Z(\mathbf{x}) - \mu(\mathbf{x})\}\{Z(\mathbf{x}+h) - \mu(\mathbf{x}+h)\}]$$

where

- (x) and (x + h) are the means of Z at x and x + h
- h is the spatial lag
- E denotes the expected value.

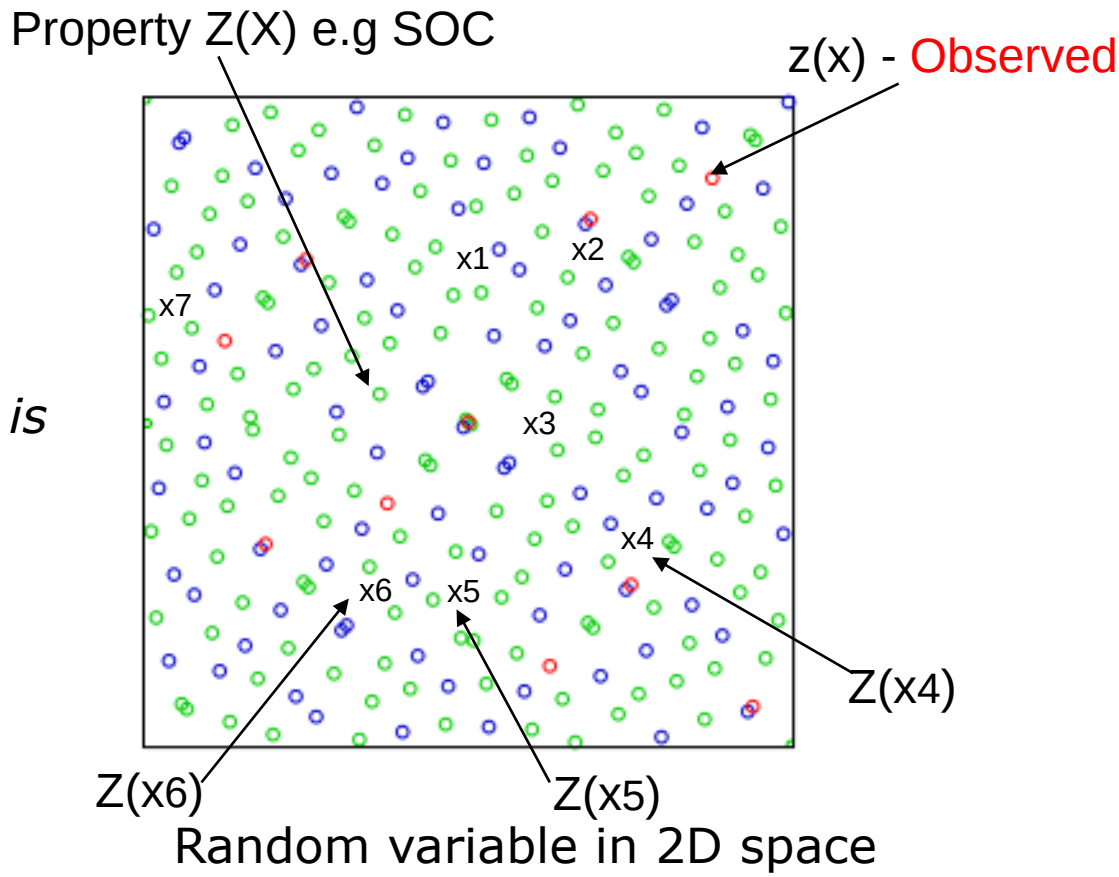


Random variable in 2D space

- Means are unknown as there is only ever one realization of Z at each point, hence we invoke **stationarity**

Stationarity

- Stationarity means that the **distribution** is **invariant** under translation, i.e. *mean is assumed constant between samples*, independent of location



Stationarity

- The covariance between any two points x and x + h is independent of x (do not coincide). It depends only on the vector h. Thus:

$$C[Z(\mathbf{x}), Z(\mathbf{x}+h)] = E[Z(\mathbf{x})Z(\mathbf{x}+h) - \mu^2]$$
$$= C(h).$$

- **Stationarity** of the **covariance** implies stationarity of **variance**, and the variogram.
- The covariance is a function of the lag, C(h), and the lag only.

Stationarity

- In spatial context, we require:
  - The expected value (or mean) of the random variable function Z(x) to be constant for all points x. That is E[Z(x)]=μ(x) =μ which is independent of x and h where h is the spatial lag. With this assumption the covariance function in becomes;

$$C[Z(\mathbf{x}), Z(\mathbf{x}+h)] = E[\{Z(\mathbf{x}) - \mu)\}\{Z(\mathbf{x}+h) - \mu\}]$$

  - If the two points x; x+h coincide i.e., x = (x+h) = x, then the above equation defines the variance.

$$\sigma^2 = E[\{Z(\mathbf{x}) - \mu)\}\{Z(\mathbf{x}) - \mu)\}]$$

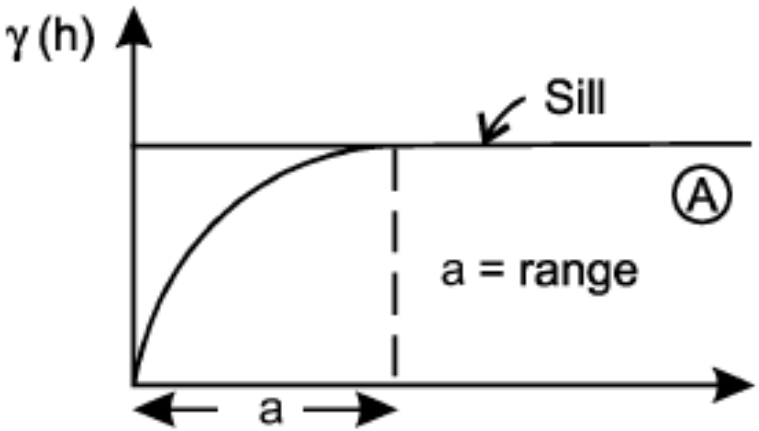
$$\sigma^2 = E[\{Z(\mathbf{x}) - \mu)\}^2]$$

Intrinsic hypothesis

- A random function is said to be **intrinsic** if
- The mathematical expectation exists and does not depend on the support point x, i.e., $EZ(x) = \mu$ and
- For any vector h, the increment $[Z(x + h) - Z(x)]$ has a finite variance which is independent of the point x. In other words,
- $E[Z(x + h) - Z(x)] = 0$ and $Var[Z(x + h) - Z(x)] = 2\gamma(h)$, a finite value which does not depend on x.
- The function $2\gamma(h)$ is called the s**emi-variogram**. This can simply be called **variogram**.

Variogram

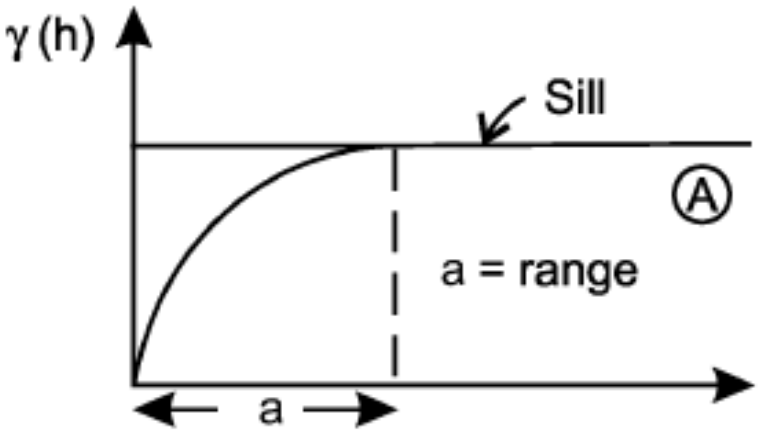
- The spatial correlations of an intrinsic random function are characterised by the **theoretical semi-variogram function**. The semi-variogram can be estimated in two ways namely:
  - Matherons method of moments (MoM)
  - Residual maximum likelihood (REML)



- In the MoM method (Matheron, 1967) the semi-variance can be estimated as: $$\gamma(h) = \frac{1}{2}\mathrm{Var}[Z(\mathbf{x}+h) - Z(\mathbf{x})].$$

The variogram

- The spatial correlations of an intrinsic random function are characterised by the **theoretical semi-variogram function**. The semi-variogram can be estimated in two ways namely:
  - Matherons method of moments (MoM)
  - Residual maximum likelihood (REML)



- In the MoM method (Matheron, 1967) the semi-variance can be estimated as: $$\gamma(h) = \frac{1}{2}\mathrm{Var}[Z(\mathbf{x}+h) - Z(\mathbf{x})].$$

Variogram

- Since it has been assumed that the mean of Z(x + h)- Z(x) is zero, γ(h) is just half the mean square value of the difference

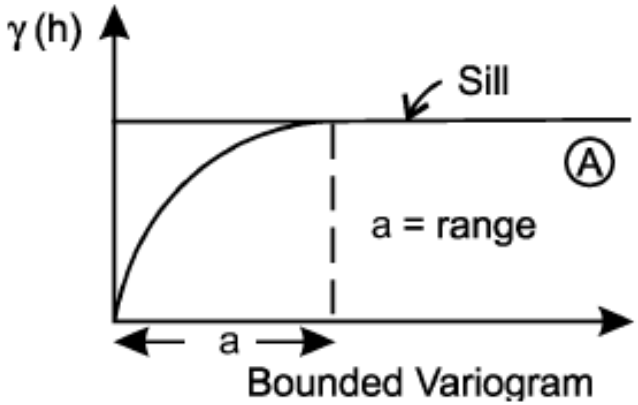
$$\gamma(h) = \frac{1}{2}E[Z(\mathbf{x}+h) - Z(\mathbf{x})]^2$$

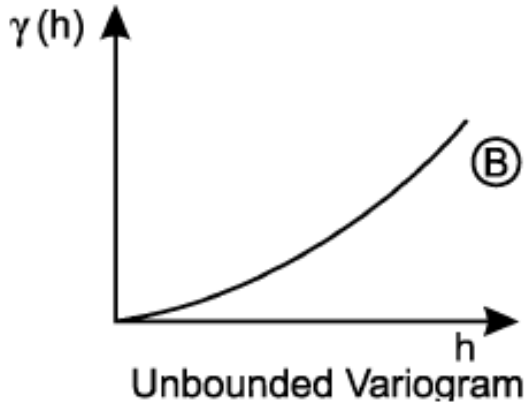
$$\hat{\gamma}(h) = \frac{1}{2N}\sum_{i=1}^{N}[Z(\mathbf{x}_i+h) - Z(\mathbf{x}_i)]^2$$ For a discrete case

  - N = Number of pairs. The model above is known as the **empirical variogram**

Description of a variogram

- A variogram may simply be described as follows:
- It starts at 0 [for h = 0;Z(x + h) = Z(x)]
- It generally increases with h.
- It rises up to a certain level called the sill and then flattens out in some cases. In other cases, it continues to rise





Variogram computation example

- Consider the following set of gold assay values (grade in units of dwts2/ton of ore) of samples, each separated by a distance of 3 ft taken from a segment of a gold bearing
- lode 'O' of gold field I, southern India.

Data in dwts/ton of ore: (one dwt = 1.55517 gms ton of ore).

4	3	3	5	5	5	4	4	5	4
5	17	8	2	2	3	7	7	1	6
10	9	9	10	11	12	11	3	3	4

$$\hat{\gamma}(h) = \frac{1}{2 \times 29}[(4-3)^2 + (3-3)^2 + (3-5)^2 + (5-5)^2 + (5-5)^2 + (5-4)^2 + (4-4)^2 +$$
$$(4-5)^2 + (5-4)^2 + (4-5)^2 + (5-17)^2 + (17-8)^2 + (8-2)^2 + (2-2)^2 + (2-3)^2 +$$
$$(3-7)^2 + (7-7)^2 + (7-1)^2 + (1-6)^2 + (6-10)^2 + (10-9)^2 + (9-9)^2 + (9-10)^2 +$$
$$(10-11)^2 + (11-12)^2 + (12-11)^2 + (11-3)^2 + (3-3)^2 + (3-4)^2] = 7.48$$

Variogram computation example

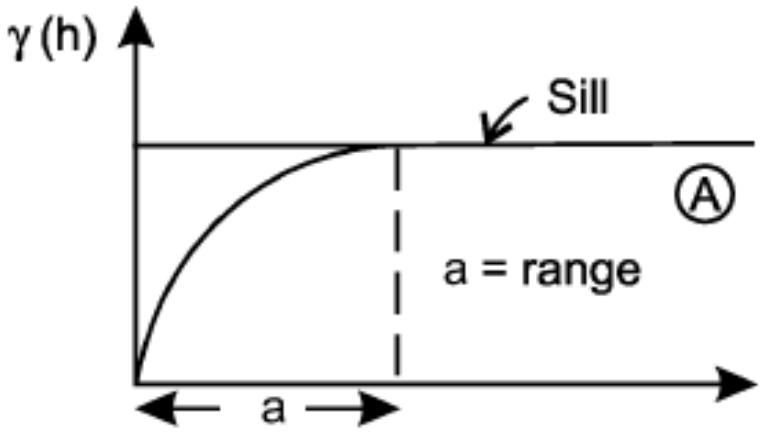
The set-up for $\gamma(6)$ can be computed as: In this case the eligible pairs are:

What is the value for γ6? similarly, compute for (γ9) and (γ12), showing the number of pairs for each lag.

Properties of a Variogram

Range

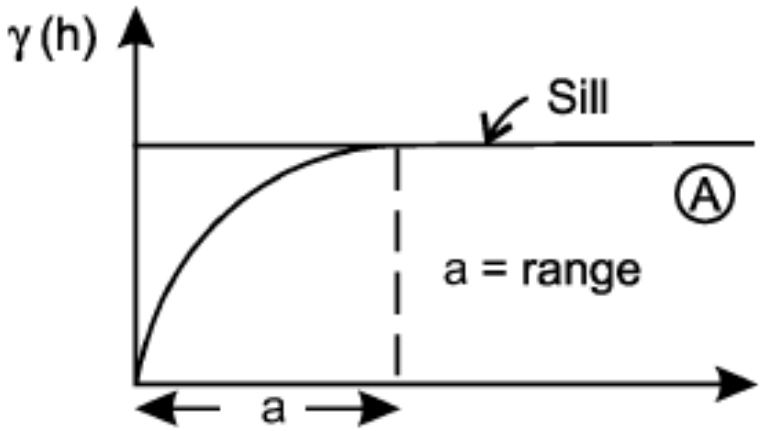
- The rate of increase of γ(h) is an indicator of the rate at which the 'influence' of a sample decreases with increasing distances from the sample site.
- In summary, the range is the distance at which there is no evidence of spatial dependence/correlation.



Properties of a Variogram

Sill

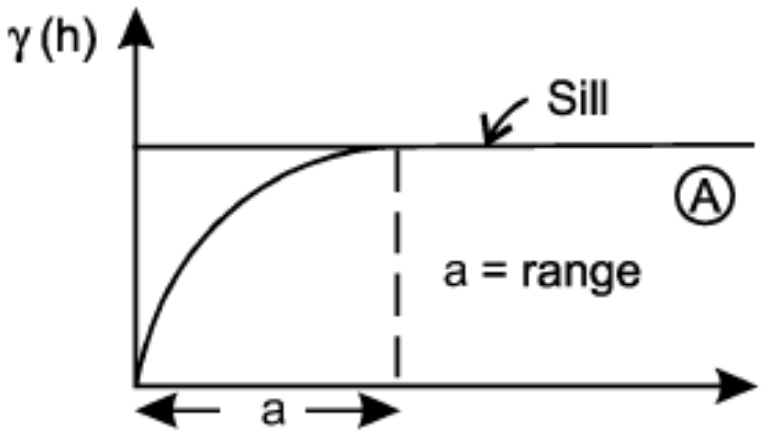
- Sill is the maximum semi-variance.
- It represents variability in the absence of spatial dependence



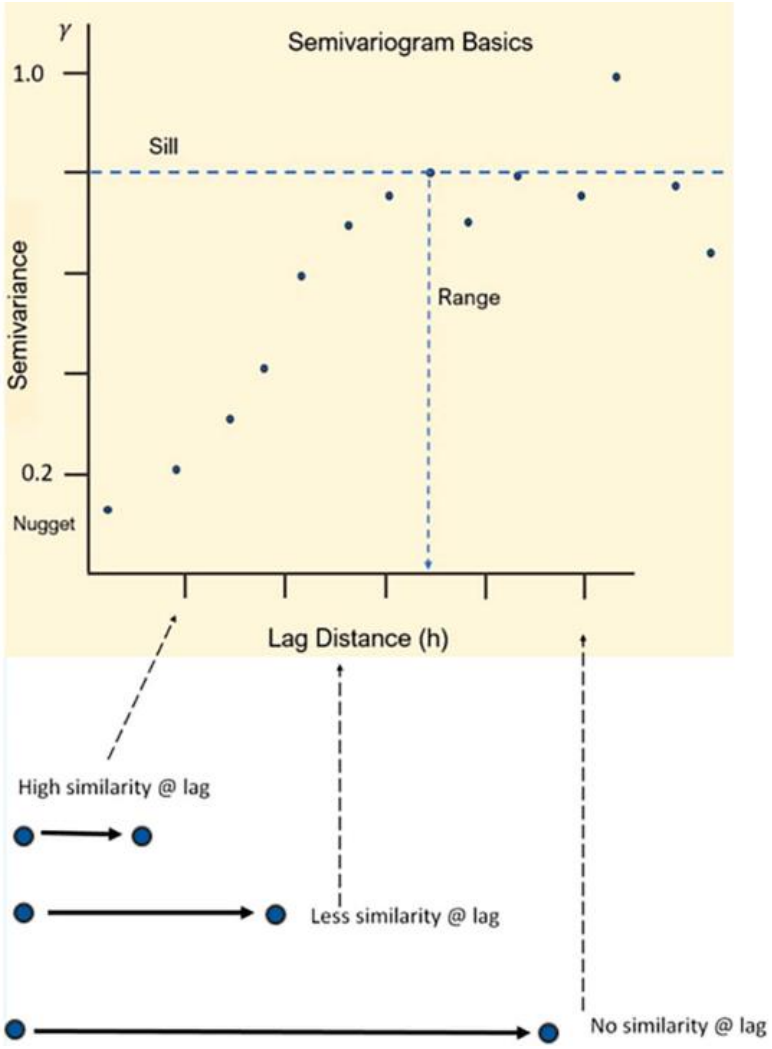
Properties of a Variogram

Nugget

- In theory the semi-variogram value at the origin (0 lag) should be zero. If it is significantly different from zero for lags very close to zero, then this semi-variogram value is referred to as the **nugget**.
- Represents variability that cannot be explained by spatial structure



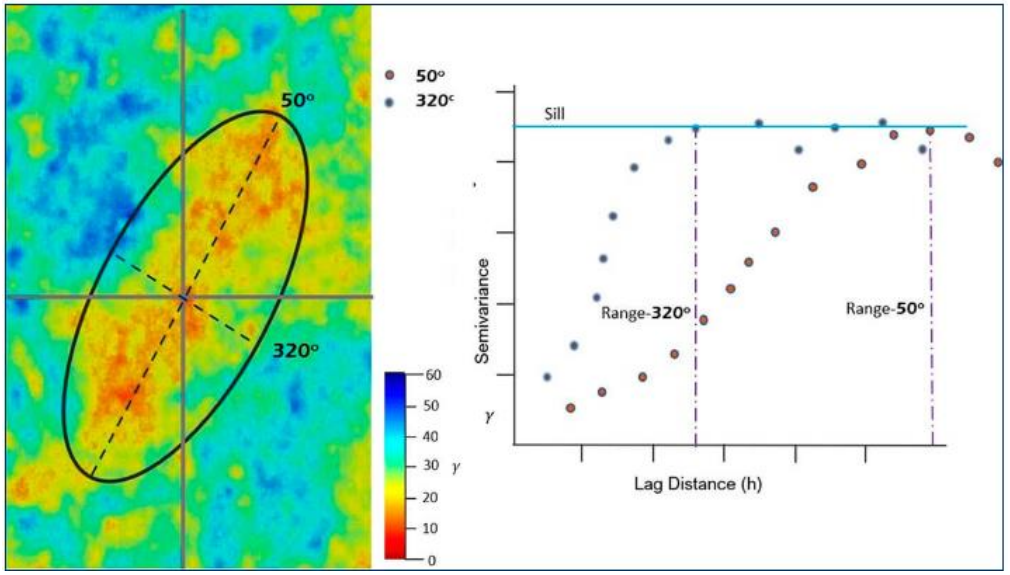
Properties of a Variogram



Properties of a Variogram

Anisotropy

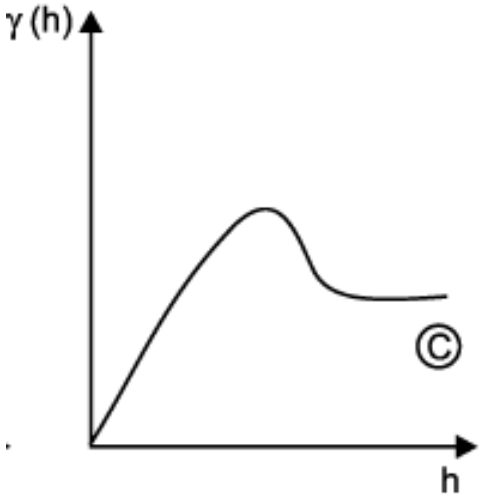
- Spatial variation varies based on directions.
- This behavior is termed as the presence of **anisotropy**.
- In absence of anisotropy, the variogram depends only on the magnitude of the distance between points h and is said to be **isotropic**.



Properties of a Variogram

Hole effect and periodicity.

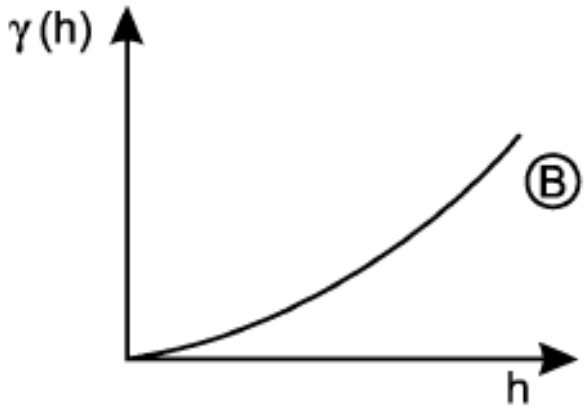
- The variogram may decrease from its maximum to a local minimum and then increase again, indicating greater regularity of repetition



Properties of a Variogram

Unbounded variogram

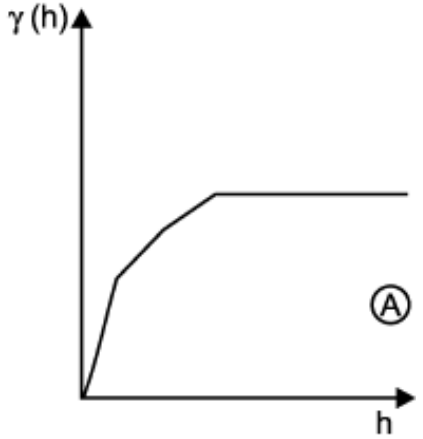
- If the variogram increases indefinitely with increasing lag distance
- The process is intrinsic only.



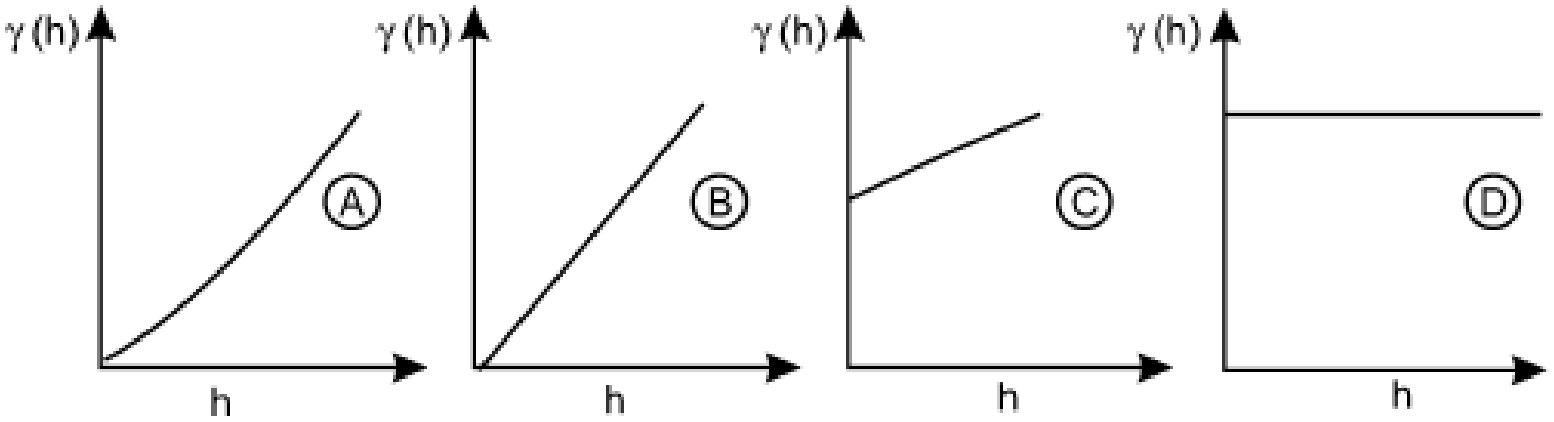
Properties of a Variogram

Nested Structures

- These indicate the presence of variations at different scales such as sample collection, petrographic analysis, remote sensing etc.



Behaviour of variogram near the Origin



Behaviour of variogram at the origin: (A) Highly continuous, (B) Continuous, (C) Discontinuous (D) Purely random

Behaviour of variogram near the Origin

Highly continuous – Regionalized variable highly continuous and differentiable (Parabolic)

Continuous- Regionalized variable highly continuous but not differentiable (Linear shape)

Discontinuous – Discontinuity at the origin, the nugget effect. Variable highly irregular at short distances

Purely random – Lack of total structure

Practical notes on variogram

- At least 50 samples required to construct a variogram
- 250 samples is reliable
- Set of vectors are defined for each bin, thus distance range can be collected into one bin
- Values near the origin can be regarded as important
- More points needed for anisotropic variogram

Modelling the variogram

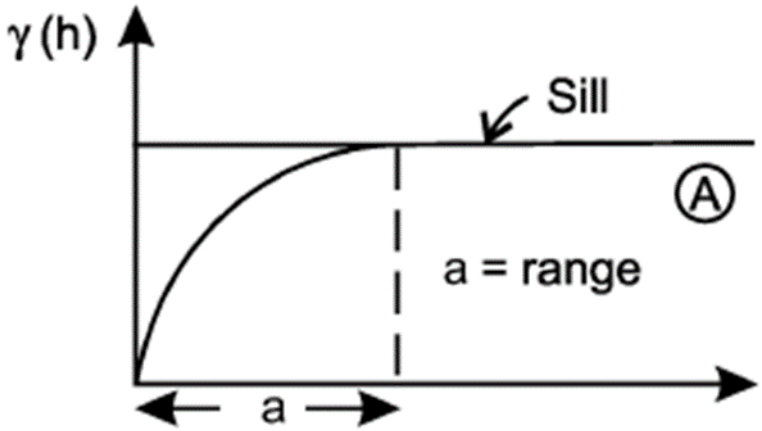
- Characteristics of empirical variogram are **parameterized** using different semi-variogram models
- The following features are considered;
  - A **monotonic** increase with increasing lag distance from the ordinate of appropriate shape;
  - A **constant** maximum or asymptote, or sill;
  - A **positive** intercept on the ordinate, or nugget;
  - Anisotropy.

Modelling the variogram

- The variogram model allows us to:
- Infer the characteristics of the underlying process from the functional form and its parameters;
- Compute the semi-variance between any point-pair, separated by any vector which is used in an optimal interpolator (kriging) to predict at unsampled locations.
- More importantly, the semivariogram models used in the kriging process need to obey certain numerical properties (stated before) in order for the kriging equations to be solvable.

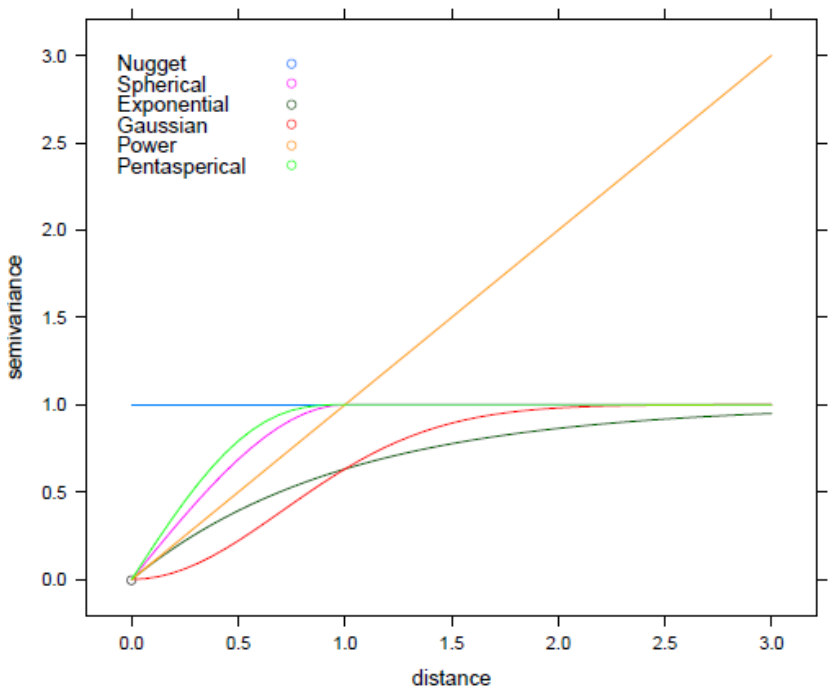
Types of the variogram

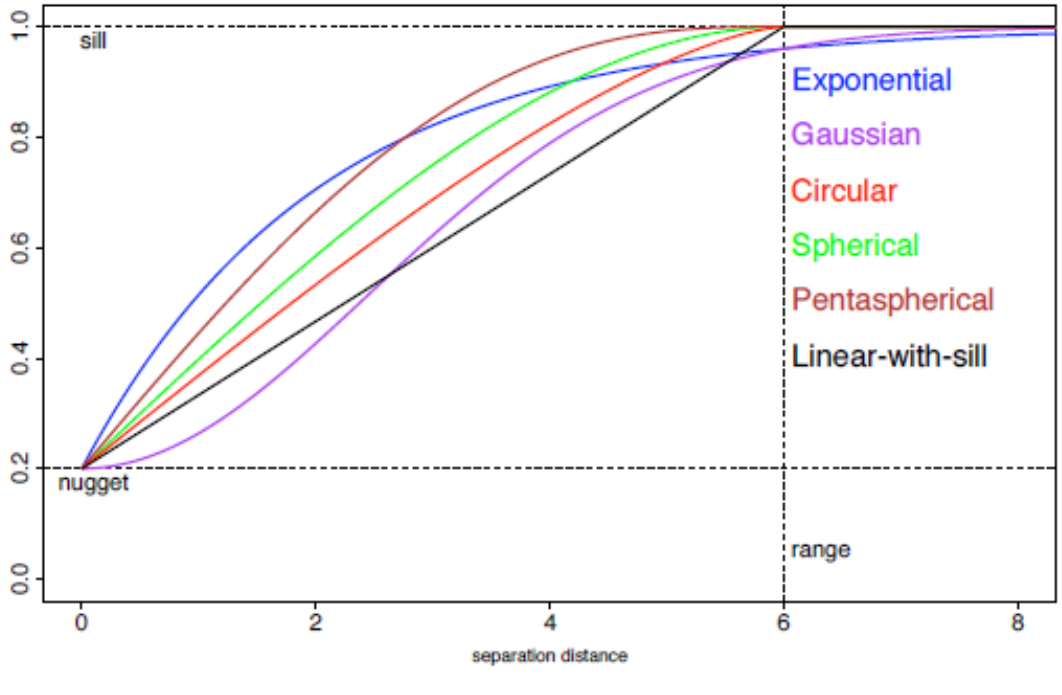
Recall the variogram illustration



- Using h to represent lag distance, a to represent (practical) range, and c to represent sill, various types of variogram models can be defined

Types of the variogram





- **Nugget** - The nugget model represents the discontinuity at the origin due to small-scale variation. Pure randomness with no spatial correlation

Types of the variogram

Spherical

- The spherical model is the most commonly used model.
- The tangent at the origin intersects the sill at a point with an abscissa 2a/3 .
- This model curves more gradually as the sill is reached than the circular one

$$\gamma(h) = \begin{cases} c \cdot \left(1.5\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\right) & \text{if } h \leq a \\ c & \text{otherwise} \end{cases}$$

Types of the variogram

Exponential

- Together with spherical functions together account for a large proportion of the models fitted in the environmental sciences.
- For practical purposes, the range can be taken as 3a.
- The tangent at the origin intersects the sill at a point with an abscissa a.
- The exponential model approaches its sill gently and asymptotically so that it does not have a finite range.

$$\gamma(h) = c \cdot \left(1 - \exp\left(\frac{-3h}{a}\right)\right)$$

Types of the variogram

Gaussian

- The Gaussian model, with its parabolic behavior at the origin, represents an extremely continuous phenomenon.

$$\gamma(h) = c \cdot \left(1 - \exp\left(\frac{-3h^2}{a^2}\right)\right)$$

Power

- Does not reach a finite sill, and does not have a corresponding finite function

$$\gamma(h) = c \cdot h^{\omega} \text{ with } 0 < \omega < 2$$

Choosing a variogram model

- Fitting a model to an empirical semi-variogram is much more of an **art** than a science, More of subjective judgement

Principles usually followed

- **Exponential**: Suitable for first-order autoregressive process: values are random dependency on the nearest neighbor
- **Gaussian**: similar to exponential model, but phenomenon exhibits strong close-range dependency
- **Spherical, circular, pentaspherical**: Patches of similar values; patches have similar size (approx. range) with transition zones (overlap of processes);
- **Historical experience**. Models successfully applied with the type of data at hand or kind of process.

Choosing a variogram model

- **Expectations** from the process. What do we expect from the supposed process? if we have some other evidence of its spatial behaviour. For example, a Gaussian model might be expected for a phenomenon which physically must be very continuous, the surface of a ground-water table.
- **Subjective judgment**. Visual estimate of functional form from the variogram.
- Use **statistical test**. Fit various models, pick the statistically-best fit.

Fitting a variogram model

- Once a model form is selected, then the model parameters must be adjusted for a best fit of the experimental variogram. This can be through the following ways:
  - By **eye**, adjusting parameters for good-looking fit. However, it is hard to judge the relative value of each point.
  - **Automatically**, looking for the **best fit** according to some objective criterion. Various criteria possible in gstat.
  - Favour sections of the variogram with **more pairs** and at **shorter ranges** (because it is a **local interpolator**).
  - **Mixed**: adjust by eye, evaluate statistically; or vice versa

Thank you for your attention! Questions?