COA 616 Geostatistics in Environmental Sciences

Lecture 4 – Assumptions of geostatistics

Wei Wu

September 20, 2016

Random variables

Random variable: A well-defined numerical description of the outcomes in the sample space of a random experiment.

A sample space associated with a random experiment can be classified as discrete or continuous.

Discrete sample space: Contains a finite number of elements.

Continuous sample space: Contains an infinite and uncountable number of outcomes.

Discrete random variable: Random variable defined over discrete sample spaces.

Continuous random variable: Random variable defined over continuous sample spaces.

Normal distribution

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad where -\infty < x < \infty$$

Four important properties:

1. The mode, median, and mean are all equal.
2. The curve is symmetric around the vertical axis drawn through the mean.
3. The curve is asymptotic to the x-axis in both the positive and negative directions.
4. The total area under the curve is 1.

Random or Deterministic?

Example: total nitrogen at $\mathbf{x}_i$

A full deterministic solution to our problems seems out of reach at present.

Stochastic view – We regard the observations as one drawn at random from the set of values according to some law, from some probability distribution.

I.e. at a point **x** a property Z(**x**) is treated as a random variable with a mean μ, a variance σ^2, and higher order moments, and an accumulative distribution function (cdf).

Random processes

Distribution property $Z(\mathbf{x}_i)$ in space

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots \mathbf{x}_n$, as sampling locations

We have a random variable (RV) at each of these sampling locations

$z(\mathbf{x}_1)$ – realization of RV $Z(\mathbf{x}_1)$

$z(\mathbf{x}_2)$ – realization of RV $Z(\mathbf{x}_2)$

$z(\mathbf{x}_3)$ – realization of RV $Z(\mathbf{x}_3)$

…

$z(\mathbf{x}_n)$ – realization of RV $Z(\mathbf{x}_n)$

$[Z(\mathbf{x}_1), Z(\mathbf{x}_2), Z(\mathbf{x}_3), \ldots, Z(\mathbf{x}_n)]$ = Random function

$[z(\mathbf{x}_1), z(\mathbf{x}_2), z(\mathbf{x}_3), \ldots, z(\mathbf{x}_n)]$ = Regionalized variables

Regionalized variable (Data) → Estimation / Modeling → Infer property of the random function (spatial dependence) → Kriging → Predict values of RF at unsampled locations

Assumptions

$$C(\vec{x}_1, \vec{x}_2) = E[\{Z(\vec{x}_1) - \mu(\vec{x}_1)\}\{Z(\vec{x}_2) - \mu(\vec{x}_2)\}]$$

Stationarity: The distribution of the random process has certain attributes that are the same everywhere.

Second-order stationarity (weak stationarity) :

1) Assume the mean $\mu = E(Z(\mathbf{x}))$ is constant for all $\mathbf{x}$.

$$C(\vec{x}_1, \vec{x}_2) = E[\{Z(\vec{x}_1) - \mu\}\{Z(\vec{x}_2) - \mu\}]$$

2) When x_1 and x_2 coincide

$$\sigma^2 = E[\{Z(\vec{x}) - \mu\}^2] \text{ which is assumed to be the same everywhere}$$

3) When x_1 and x_2 do not coincide

$$C(\vec{x}_i, \vec{x}_j) = E[\{Z(\vec{x}_i) - \mu\}\{Z(\vec{x}_j) - \mu\}] \text{ which is assumed to be constant for any } \vec{h} = \vec{x}_i - \vec{x}_j$$

Assumptions

Strictly or fully stationary: Higher moments depend on the separation **h** only.

Why does full stationarity not matter in practice?

$$C(\vec{x}_i, \vec{x}_j) = E[\{Z(\vec{x}_i) - \mu\}\{Z(\vec{x}_j) - \mu\}] \text{ can be written as}$$
$$\begin{aligned}\mathrm{C}(\mathrm{Z}(\vec{x}), \mathrm{Z}(\vec{x}+\vec{h})) &= E[\{Z(\vec{x}) - \mu\}\{Z(\vec{x}+\vec{h}) - \mu\}] \\ &= \mathrm{E}[\{\mathrm{Z}(\vec{x})\}\{Z(\vec{x}+\vec{h})\} - \mu^2] \\ &= \mathrm{C}(\vec{h})\end{aligned}$$

Covariance function does not exist if weak or second-order stationarity does not meet.

Intrinsic stationarity

Matheron (1965)

Instead of trying to model $Z(\vec{x})$, we will model the difference $Z(\vec{x}) - Z(\vec{x}+\vec{h})$

$$E[Z(\vec{x}) - Z(\vec{x}+\vec{h})] = 0 \text{ for sufficiently small } \vec{\mathrm{h}} \text{ even if } \mathrm{E}(\mathrm{Z}(\vec{\mathrm{x}})) \text{ is not constant}$$
$$\begin{aligned}\mathrm{var}[Z(\vec{x}) - Z(\vec{x}+\vec{h})] &= E[\{Z(\vec{x}) - Z(\vec{x}+\vec{h})\}^2] \\ &= 2\gamma(\vec{\mathrm{h}}) \text{ dependes on } \vec{\mathrm{h}} \text{ only}\end{aligned}$$

$\gamma(\vec{\mathrm{h}})$ is semivariance.

A function of $\vec{\mathrm{h}}$ is semivariogram or variogram.

These two equations constitute intrinsic stationarity.

For second-order stationarity, $\gamma(\vec{h}) = C(\vec{0}) - C(\vec{h}) = \sigma^2(1 - \rho(\vec{h}))$

Property of covariance and semivariance

Symmetry: $C(\vec{h}) = C(-\vec{h})$

$\gamma(\vec{h}) = \gamma(-\vec{h})$

Positive semidefiniteness: The covariance matrix for any number of points is positive semidefinite. The variogram must be negative semidefinite.

$C(\vec{h})$ and $\gamma(\vec{h})$ are continuous at $\vec{h} = \vec{0}$, then they must be continuous everywhere.

Continuity: The variogram must pass through the origin if the process is continuous. However, calculated $\gamma(\vec{0})$ sometimes appear positive in sample variogram.

The positive value is nugget variance.

$$\gamma(\vec{h}) = \sigma_N{}^2 (1 - \delta(\vec{h})) + \gamma'(\vec{h})$$

where

$\sigma_N{}^2$ is nugget effect

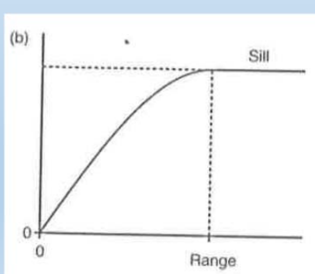
$\delta(\vec{h})$ is the Kronecker delta function taking the valurs 1 when $\vec{h} = \vec{0}$ and 0 otherwise.

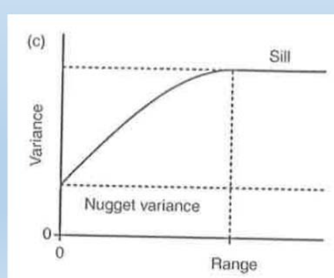
Semivariograms

Monotonic increasing: The variances increases with increasing lag distance.

Sill and range

- Sill: An upper bounds of a semivariogram. The maximum variance, a prior variance, σ^2, of the process.
- Range: The lag distance a semivariograms reach sill. The effective ranges are the lag distances at which semivariograms reach 0.95 of their sills.

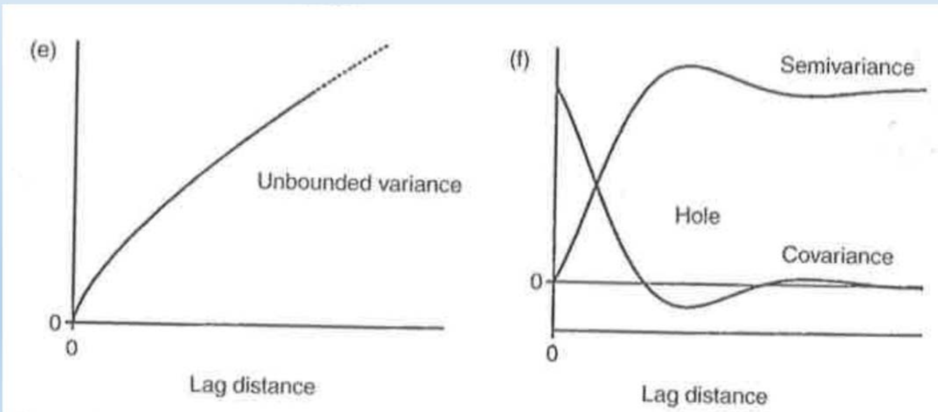




Semivariograms

Unbounded variogram: The process may be intrinsic but not second-stationary

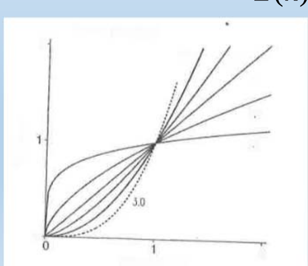
Hole effect: Due to regular repetition in the process

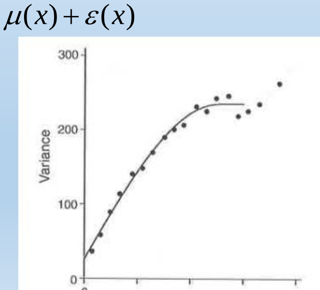


Semivariogram

- Anisotropy: Spatial variation is not the same in all directions. Geometric, zonal
- Trend
  Local trend: The curves show concave at the origin
  Long range trend: Semivariograms increase after having appeared to reach sill

$$Z(\vec{x}) = \mu(\vec{x}) + \varepsilon(\vec{x})$$





Support

Measurements must be made on finite volumes. The volume, with its particular size, shape and orientation, is the support of the sample.

Practical consequence

1) It sets the minimum to the resolution of spatial variation that can be detected and measured by that sample.
2) Variogram in practice is a function of support