Introduction

Data: set of n attribute measurements $\{z(\mathbf{s}_i), i = 1, \ldots, n\}$, available at n sample locations $\{\mathbf{s}_i, i = 1, \ldots, n\}$

Objectives:

Slide 1

- quantify spatial *auto-correlation*, or attribute dissimilarity typically expressed as: $\frac{1}{2}[z(\mathbf{s}_i) - z(\mathbf{s}_j)]^2$ as a function of separation distance between sample pairs $\mathbf{s}_i$ and $\mathbf{s}_j$
- introduce the sample semivariogram, its characteristics, and provide some examples
  NOTE: *Spatial auto-correlation is a second-order characteristic of spatial variation, and hence the sample semivariogram should be computed from data whose spatial variation is not explained by first-order effects*
- justify the need of going beyond the sample semivariogram to a semivariogram model
- introduce parametric functions of distance that can be used as formal theoretical semivariogram models
- discuss issues of fitting semivariogram models to sample semivariogram values

Semivariogram Cloud

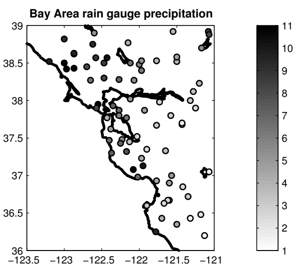
Definition: A scatter-plot of *attribute* squared semidifferences between all possible pairs of samples measured at different locations, versus their separation distance

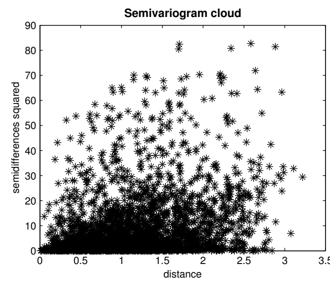
Computational procedure:

Slide 2

1. construct Euclidean distance matrix $\mathbf{D} = [d_{ij}, i = 1, \ldots, n, j = 1, \ldots, n]$ between all n^2 pairs of data locations, where d_{ij} is defined as: $d_{ij} = ||\mathbf{h}_{ij}|| = ||\mathbf{s}_i - \mathbf{s}_j||$
2. construct squared semidifference matrix $\mathbf{E} = [e_{ij}, i = 1, \ldots, n, j = 1, \ldots, n]$ between all n^2 pairs of attribute values, where e_{ij} is defined as: $e_{ij} = \frac{1}{2}[z(\mathbf{s}_i) - z(\mathbf{s}_j)]^2$
3. plot each distance value d_{ij} against the corresponding squared semidifference e_{ij}; in other words, plot $\mathbf{e} = vec(\mathbf{E})$ versus $\mathbf{d} = vec(\mathbf{D})$. The plot of all pairs $\{d_{ij}, e_{ij}\}$ is termed a *semivariogram cloud*

Semivariogram Cloud Example





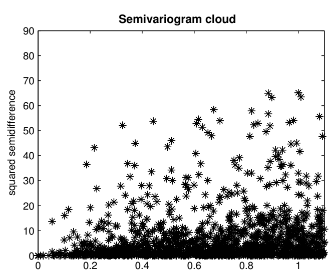
Slide 3

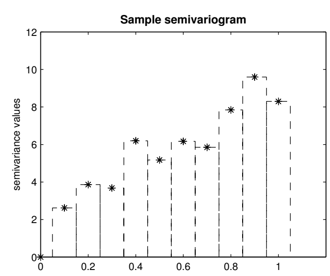
A measure of dissimilarity between attribute values measured at different locations, i.e., a spatial measure of attribute dissimilarity

Expected graph pattern: As the distance d_{ij} between sample pairs increases, the corresponding squared semidifference e_{ij} should also increase

Difficult to interpret, so we consider groups of sample pairs separated by similar distances i.e., average squared semidifferences within distance classes (x-axis bins in the right graph above)

Semivariogram Cloud Versus Plot





Slide 4

Going from the first to the second:

- define a set of L distance classes; the l-th class has limits: $(d_l - t_l, d_l + t_l]$, where d_l is the class midpoint and t_l is half the class width (or distance tolerance)
- for a given distance class $(d_l - t_l, d_l + t_l]$, the semivariogram value $\hat{\gamma}(d_l)$ is the *average* of $n(d_l) << n^2$ squared attribute semidifferences computed from sample pairs whose inter-distances d_{ij} satisfy: $d_l - t_l < d_{ij} \leq d_l + t_l$
- in other words, the semivariogram plot can be regarded as a summary of the semivariogram cloud, according to some distance-based grouping of samples

Computing Sample Semivariograms

1. compute distance matrix $\mathbf{D} = [d_{ij}, i = 1, \ldots, n, j = 1, \ldots, n]$ and squared semidifference matrix $\mathbf{E} = [e_{ij}, i = 1, \ldots, n, j = 1, \ldots, n]$ between n^2 data pairs

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} & d_{15} \\ d_{12} & 0 & d_{23} & d_{24} & d_{25} \\ d_{13} & d_{23} & 0 & d_{34} & d_{35} \\ d_{14} & d_{24} & d_{34} & 0 & d_{45} \\ d_{15} & d_{25} & d_{35} & d_{45} & 0 \end{bmatrix} \qquad \mathbf{E} = \begin{bmatrix} 0 & e_{12} & e_{13} & e_{14} & e_{15} \\ e_{12} & 0 & e_{23} & e_{24} & e_{25} \\ e_{13} & e_{23} & 0 & e_{34} & e_{35} \\ e_{14} & e_{24} & e_{34} & 0 & e_{45} \\ e_{15} & e_{25} & e_{35} & e_{45} & 0 \end{bmatrix}$$

Slide 5

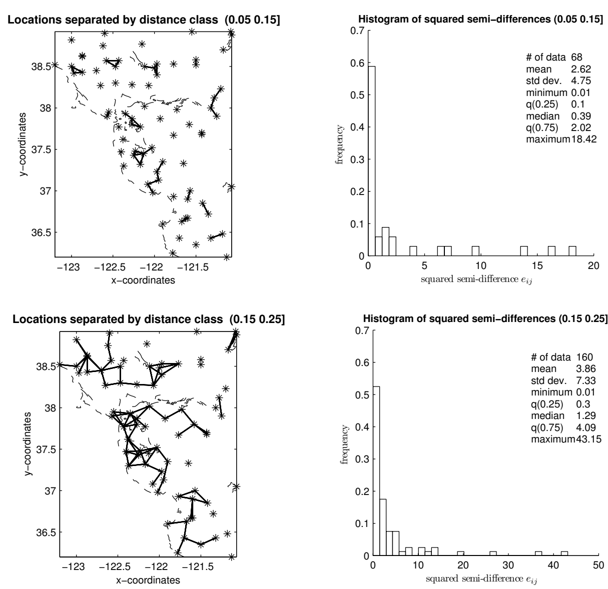
2. for a given distance class $(d_l - t_l, d_l + t_l]$, find entries of $\mathbf{E}$ that correspond to entries of $\mathbf{D}$ falling in that distance class, e.g.:

$$\begin{bmatrix} 0 & \boxed{d_{12}} & d_{13} & d_{14} & d_{15} \\ \boxed{d_{12}} & 0 & d_{23} & \boxed{d_{24}} & d_{25} \\ d_{13} & d_{23} & 0 & d_{34} & \boxed{d_{35}} \\ d_{14} & \boxed{d_{24}} & d_{34} & 0 & d_{45} \\ d_{15} & d_{25} & \boxed{d_{35}} & d_{45} & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} 0 & \boxed{e_{12}} & e_{13} & e_{14} & e_{15} \\ \boxed{e_{12}} & 0 & e_{23} & \boxed{e_{24}} & e_{25} \\ e_{13} & e_{23} & 0 & e_{34} & \boxed{d_{35}} \\ e_{14} & \boxed{e_{24}} & e_{34} & 0 & e_{45} \\ e_{15} & e_{25} & \boxed{e_{35}} & e_{45} & 0 \end{bmatrix}$$

3. sample semivariogram $\hat{\gamma}(d_l)$ for that class is the *average* of the $n(d_l)$ squared semidifferences, e-values, whose corresponding distances, d-values, fall in class $(d_l - t_l, d_l + t_l]$; i.e., the mean of all e-values in boxes in the matrix on the right above

Examples of Semivariogram Computation

Slide 6



$\hat{\gamma}((0.05\ 0.15]) = 2.62$, $\hat{\gamma}((0.15\ 0.25]) = 3.86$ = *averages of values displayed in histograms*

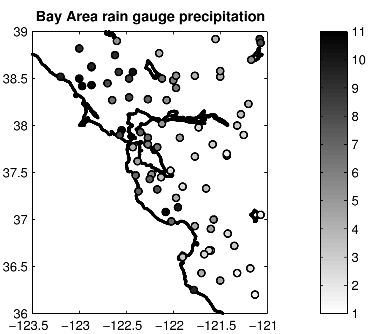
Map views linking sample pairs that contribute to such histograms are extremely informative

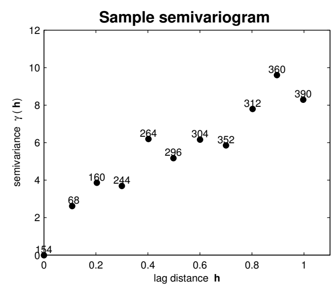
Sample Semivariogram Plots

Consider a set of L distance classes with midpoints $\{d_l, l = 1, \ldots, L\}$ and tolerances $\{t_l, l = 1, \ldots, L\}$. The plot of semivariance values $\{\hat{\gamma}(d_l), l = 1, \ldots, L\}$ versus the average sample inter-distance for each class is called a sample semivariogram

$$\hat{\gamma}(d_l) = \frac{1}{n(d_l)} \sum_{c=1}^{n(d_l)} e_c = \frac{1}{2n(d_l)} \sum_{d_{ij} \in (d_l - t_l, d_l + t_l]}^{n(d_l)} [z(\mathbf{s}_i) - z(\mathbf{s}_j)]^2$$

Slide 7

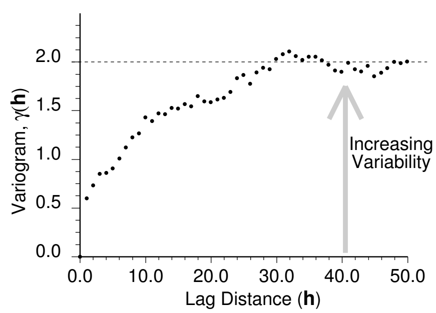




numbers above bullets denote # of sample pairs contributing to $\hat{\gamma}(d_l)$ at each lag distance

could also graph variances of e-values within the distance classes; $\hat{\gamma}(\mathbf{0}) = 0$, always

Semivariogram Characteristics



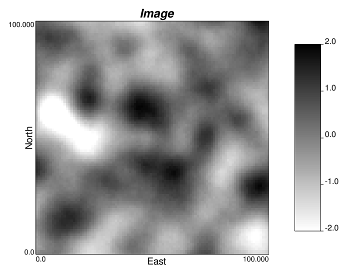
Slide 8

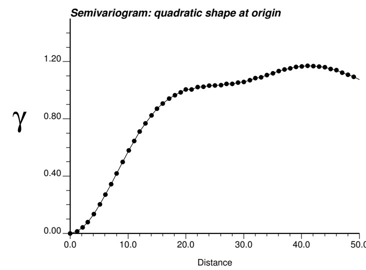
- *sill*: limit semivariogram value (plateau) is approximately equal to sample variance (for representative sample)
- *range*: distance at which semivariogram reaches (or starts oscillating around) sill = distance of influence of any datum on another
- *nugget effect*: discontinuity at origin ($\hat{\gamma}(\epsilon) > \epsilon$); sum of measurement error and micro-structures (variability at scales smaller than sampling interval)
  watch out for sparse data, outliers and positional or attribute errors
- transformation of Euclidean distance into statistical "distance" bearing imprint of specific phenomenon

Sample Semivariogram Shape & Interpretation (1)

Slide 9

Quadratic shape near origin:





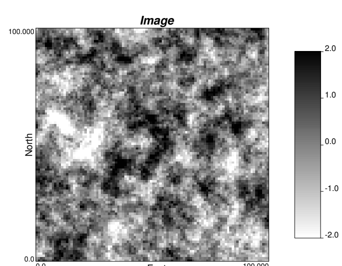
Interpretation:

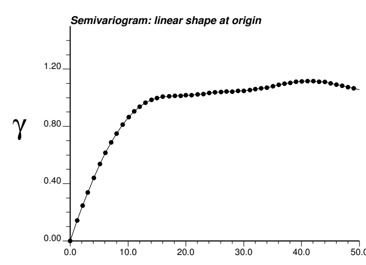
- highly continuous (extremely smooth) spatial attribute variability
- spatial attribute is differentiable
- typical variables: elevation, temperature, ...

Sample Semivariogram Shape & Interpretation (2)

Slide 10

Linear shape near origin:





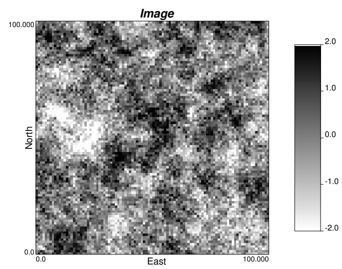
Interpretation:

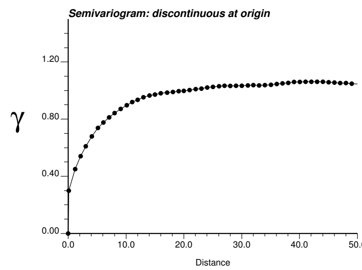
- continuous variability (not extremely smooth) of spatial attribute
- attribute is not differentiable
- typical variables: ore grades, ...

Sample Semivariogram Shape & Interpretation (3)

Discontinuous near origin:

Slide 11





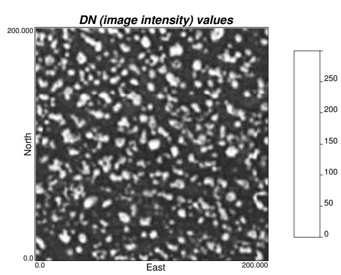
Interpretation:

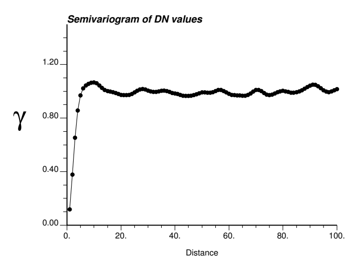
- highly irregular (quasi-random) spatial variability at small scales
- typical variables: precipitation, . . .

Sample Semivariogram Shape & Interpretation (4)

Oscillating (around sill):

Slide 12





Interpretation:

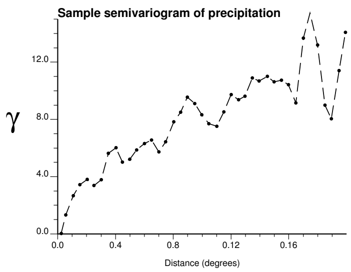
- periodic variability of spatial attribute yields sinusoidal semivariogram
- semivariogram shape possibly due to limited sampling
- need to provide physical evidence for periodicity
- frequently encountered in time series

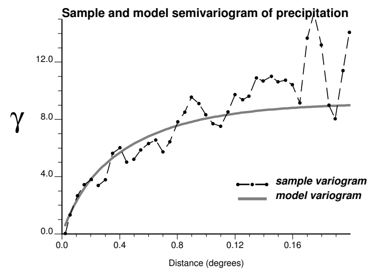
The Need for Semivariogram Models

<u>Problems:</u> (i) sill, range, and relative nugget, cannot be determined directly from the sample semivariogram plot, (ii) a continuum of semivariogram values $\gamma(d)$ for any distance vector d is required in interpolation, but *sample* semivariogram values $\{\hat{\gamma}(d_l), l = 1, \ldots, L\}$ are typically calculated only for few (L) distances $\{d_l, l = 1, \ldots, L\}$.

Slide 13

<u>Semivariogram model definition:</u> parametric function $\gamma(d; \boldsymbol{\theta})$ fitted to sample semivariogram values $\{\hat{\gamma}(d_l), l = 1, \ldots, L\}$; $\boldsymbol{\theta}$ denotes parameter vector with, e.g., range, and sill (for a given semivariogram function)



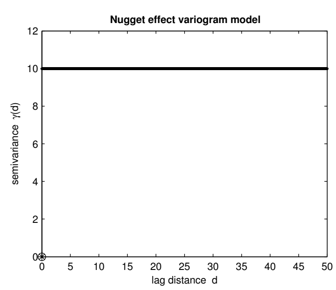


semivariogram modeling is more than a curve fitting exercise;

Warning: *cannot use any curve as semivariogram model !!!*

Valid Semivariogram Models: Pure Nugget Effect

Slide 14



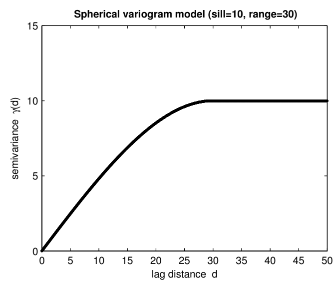
$$\boxed{\gamma(d; \boldsymbol{\theta}) = \begin{cases} 0, & \text{if } d = 0 \\ \sigma, & \text{if } d > 0 \end{cases}}$$

$\boldsymbol{\theta} = [\sigma]$, where σ denotes attribute variance

- indicates complete absence of spatial correlation
- could occur due to measurement error and microstructure, i.e., features occurring at scales smaller than sampling interval

Slide 15

Valid Semivariogram Models: Spherical



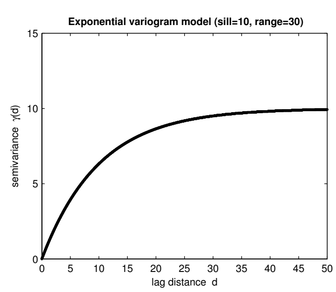
$$\boxed{\gamma(d;\boldsymbol{\theta}) = \begin{cases} \sigma\left[\frac{3}{2}\left(\frac{d}{r}\right) - \frac{1}{2}\left(\frac{d}{r}\right)^3\right], & \text{if } d < r \\ \sigma, & \text{if } d \geq r \end{cases}}$$

$\boldsymbol{\theta} = [\sigma \;\; r]$, where r is the model range

- linear behavior at origin
- clearly defined range parameter r

Slide 16

Valid Semivariogram Models: Exponential

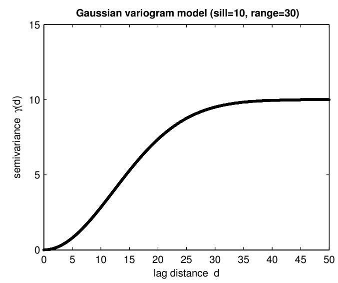


$$\boxed{\gamma(d;\boldsymbol{\theta}) = \sigma\left[1 - \exp\left(-\frac{3d}{r}\right)\right]}$$

$$\boldsymbol{\theta} = [\sigma \;\; r]$$

- linear behavior at origin; rises faster than spherical; reaches sill asymptotically
- <u>effective</u> range parameter r; distance at which 95% of sill reached

Valid Semivariogram Models: Gaussian



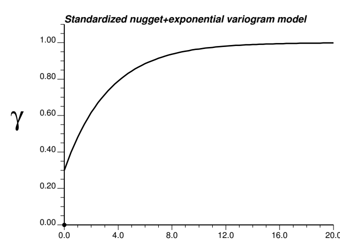
Slide 17

$$\boxed{\gamma(d;\boldsymbol{\theta}) = \sigma\left[1 - \exp\left(-\frac{3d^2}{r^2}\right)\right]}$$

$$\boldsymbol{\theta} = [\sigma \;\; r]$$

- quadratic behavior at origin; implies smooth spatial variability of attribute values; reaches sill asymptotically
- <u>effective</u> range parameter r; distance at which 95% of sill reached

Valid Semivariogram Models: Nugget + Exponential



Slide 18

$$\boxed{\gamma(d;\boldsymbol{\theta}) = \begin{cases} 0, & \text{if } d = 0 \\ a + ([\sigma - a][1 - \exp(\frac{3d}{r})]), & \text{if } d \geq \epsilon \end{cases}}$$

$$\boldsymbol{\theta} = [\sigma \;\; a \;\; r]$$

- discontinuous at origin; reaches sill asymptotically
- practical range parameter r; distance at which 95% of sill reached
- a/σ = relative nugget contribution = proportion (to total sill) of purely random spatial variability
- more complex models can be built by adding or multiplying valid models

Fitting Semivariogram Models to Sample Data

Or fitting valid semivariogram functions (curves) to sample semivariogram values

Manual fitting:

- select number of semivariograms, their type (functional form), sill, and range
- model behavior at origin (nugget effect, shape of semivariogram at distances smaller than first lag) using prior knowledge about phenomenon

Automatic fitting:

Slide 19

- least squares fit (ordinary, generalized, weighted): choose semivariogram model parameters (typically iteratively) so as to minimize discrepancy between model and sample semivariogram values over all lags; other methods also available
- treat with caution, especially with sparse data and outliers

Cross-validation:

- given a proposed parameter set, i.e., a semivariogram model, perform cross-validation using geostatistical interpolation, and record resulting error statistics
- repeat with different model parameters, and select as "optimal" model the one whose parameters yield best cross-validation error statistics

Summary

- Spatial auto-correlation can be quantified by looking at attribute dissimilarity as a function of separation distance
- The semivariogram cloud is "too cloudy" for detecting meaningful patterns
- The semivariogram plot is constructed by averaging squared semidifferences within distance bins to "smooth" out the variability in the semivariogram cloud
  NOTE: *Watch out for trends (first-order effects) in the data; a sample semivariogram quantifies second-order effects and might be contaminated by variations due to trends/drifts*

Slide 20

- A quantitative way to encapsulate a sample semivariogram is through a parametric semivariogram model
- Fitting procedures exist for estimating the parameters of semivariogram models, i.e., for fitting model semivariograms to sample semivariograms
- The final semivariogram model can be used for simulation (pattern generation) and geostatistical interpolation

NOTE: *A semivariogram model is a spatial process model, whose parameters are inferred from the sample data through the sample semivariogram*