Spatial statistics

Summer 2011

Carlos III

Fuentes

Lecture notes

Introduction

1. What is Spatial Statistics
2. Three important prototypes
3. Spatial Structure
4. Stationarity and Isotropy
5. Objectives of Spatial Statistics
6. Temporal vs. Spatial Statistics

1. What is Spatial Statistics

Spatial Statistics. Why?

The approach taken in this class is to broaden the classical models and methods to those that recognize the presence and importance of spatial information.

Some simple spatial models will be given to show effect of correlation on estimation, prediction, and design.

Spatial statistics is a vast subject, in large part because spatial data are of so many different types:

- univariate or multivariate
- categorical or continuous
- real-value (numerical) or not real-valued.
- observational or experimental

The data locations may:

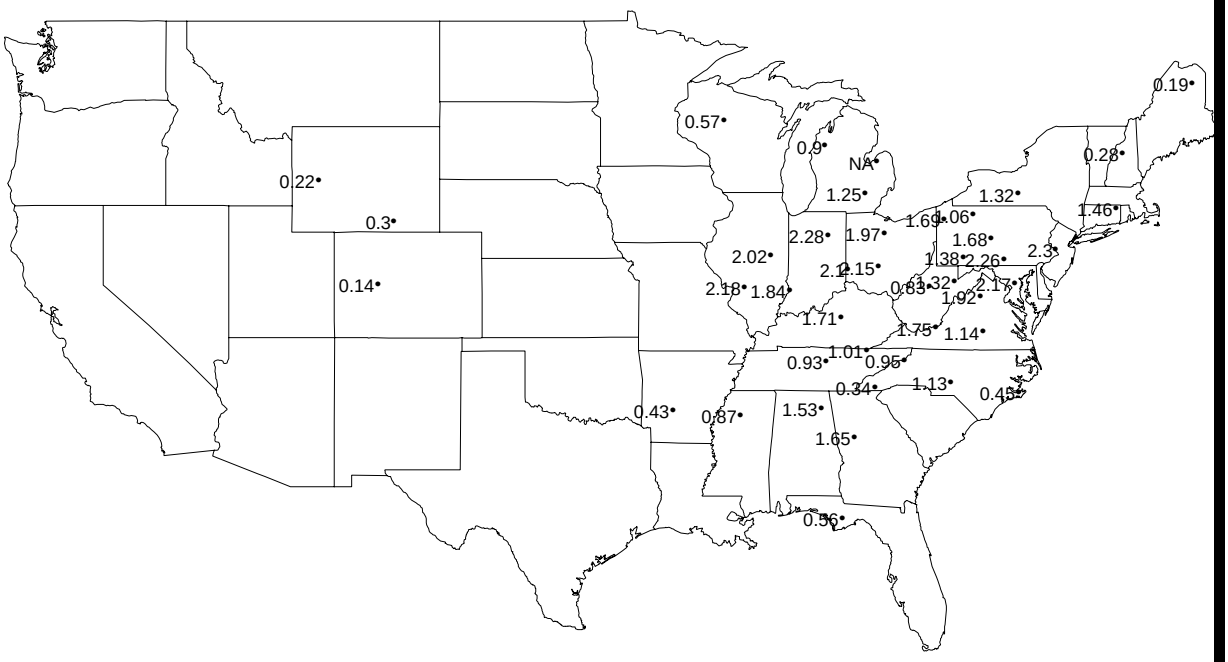
- be points, regions, line segments, or curves

- be regularly or irregularly spaced
- regularly or irregularly shaped
- belong to a Euclidean or non-Euclidean space

Three important prototypes:

1. Geostatistical data
2. Lattice data
3. Spatial point patters

The distinctions between these three types are not always clearcut.



Week 445 collected July 11, 1995

Figure 1: Example of spatial data: Weekly concentrations of HNO3.

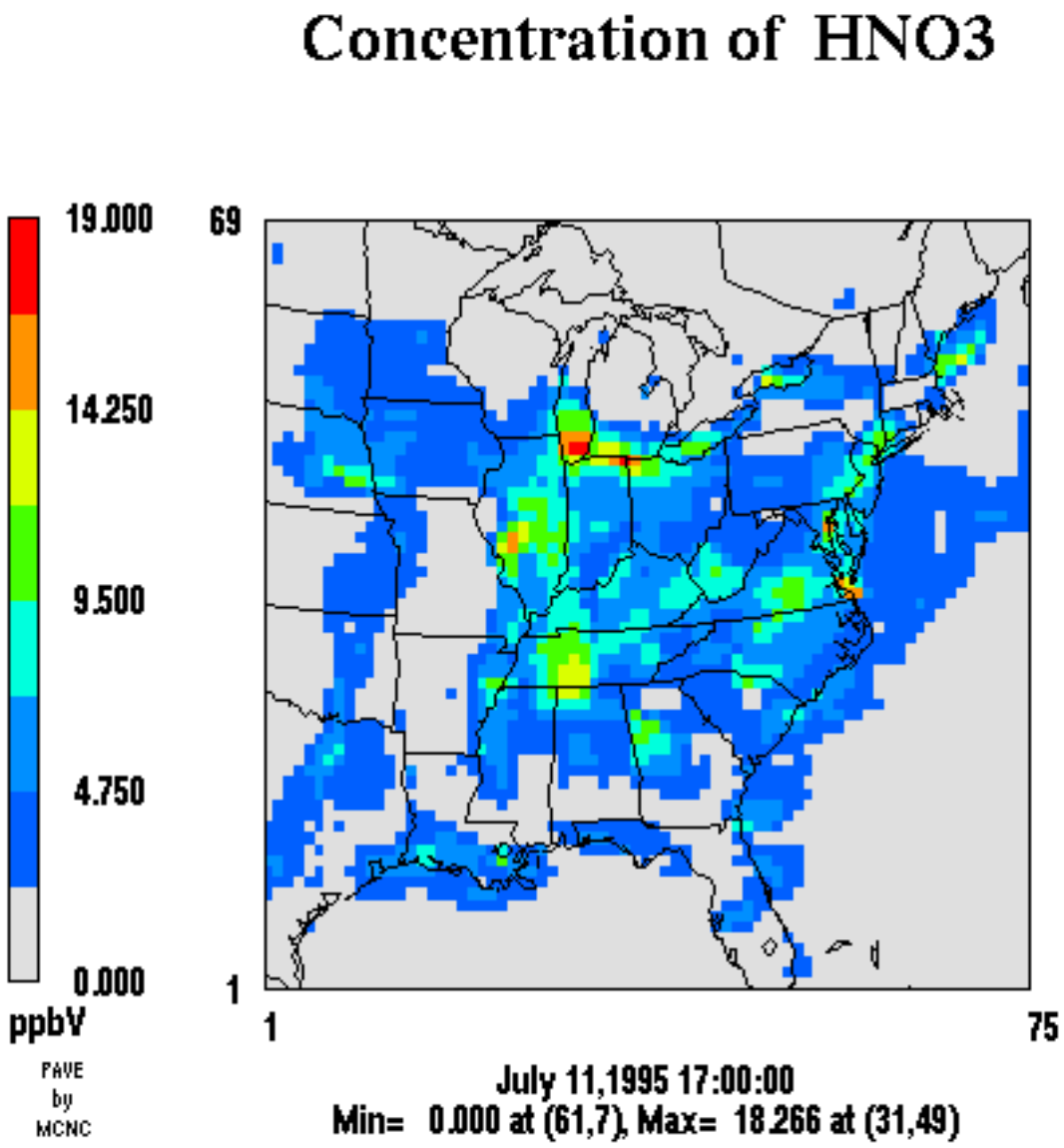


Figure 2: Example of spatial data: concentrations of nitric acid. The resolution is 36 km.

2. Three important prototypes

GEOSTATISTICAL DATA:

Point observations of a continuously varying quantity over a region.

Examples:

- Weekly concentrations of ozone in U.S.
- Annual acid rain deposition in U.S.
- Richness of iron ore within an ore body

One important problem in geostatistics is to predict the ore grade in a mining block from observed samples (Matheron, 1963).

LATTICE DATA:

Counts or averages of a quantity on subregions that make up a larger region. A lattice of locations evokes an idea of regularly spaced points. These will be referred to as *regular* lattices, allowing for the possibility of *irregular lattices*, whose relative displacements do not

follow a predictable pattern. Statistical models for lattice data need to express the fact that observations nearby (in time or space) tend to be alike.

Examples:

- Presence or absence of a plant species in square quadrats over a study area
- Number of deaths due to SIDS in the counties of NC
- Pixel values from remote sensing (satellites)

SPATIAL POINT PATTERNS:

Points on a map. Point patterns arise when the important variable to be analyzed is the location of *events*. The question of interest is whether the pattern is exhibiting complete spatial randomness, clustering, or regularity.

Examples:

1. Location of bird's nests in a suitable habitat – evidence of territoriality?
2. Location of longleaf pines in a natural

forest in NC – evidence of clustering?

3. Diameter of the longleaf pines – Do larger trees cluster?

The size variable (in example 3) is usually called a *mark variable*, and the process is called a *marked spatial point process*.

3. Spatial Structure

- Large-scale structure (Global, over the entire region)
  - Mean function of a geostatistical process
  - Intensity of spatial point process
  - Mean vector of lattice data.
- Small-scale structure (Local, highly localized region)
  - Variogram, covariance function of geostatistical process (and lattice)
  - Nearest-neighbor functions for spatial point process.
  - Neighbor weights for lattice process

4. Stationarity and Isotropy

- Stationarity:
  - constant large-scale structure
  - small-scale structure which depends on the spatial locations only through their relative positions.
- Isotropy. the small-scale structure depends on the spatial locations only through the Euclidean distance between them.

5. Objectives of Spatial Statistics

1. Estimation. Estimating treatment effects in spatial experiments. Estimate the autocorrelation structure.
2. Prediction. Prediction of unobserved variables (kriging)
3. Design Issues. Where to take observations or how to arrange treatments in a spatial experiment.

6. Temporal vs. Spatial Statistics

- Time flows in one direction only, from past to present to future. Not so in space.
- Contrast between time series and geostatistics/lattice data analysis:
  1. In time series, observations usually regularly spaced.
  2. In time series models, observations usually are assumed to be dependent by identically distributed. In Geostatistics observations are assumed to be non-identically distributed (trend).
  3. In time series, interaction is unidirectional. In space, interaction occurs in all directions.
  4. In time series, prediction usually is extrapolating to a future time. In space, interpolation is usually more important.
- Geostatistics and lattice data analysis are

most similar to longitudinal data analysis (data collected repeatedly on experimental units over time). Key difference: independent replications generally exist in the longitudinal case but not in the spatial case.

- spatial point pattern (SPP) analysis is most similar to failure time data analysis.
  1. A SPP is a window of a process which actually occurs over a larger region.
  2. SPP models feature neighbor interactions prominently.

SPATIOTEMPORAL STATISTICS

Spatiotemporal data are observations with identifiable and observed spatial and temporal labels.

Examples:

- earthquakes (locations random in time and space)
- change in locations of trees over time (locations random in space but nonrandom in time)
- environmental monitoring of water quality (locations nonrandom in time and space)

We can model space-time data as:

- a collection of temporally correlated spatial random fields, lattice processes, or spatial point processes, or
- a collection of spatially correlated time series.

Outline

1. Exploratory Data Analysis
2. Geostatistical Models.
3. Estimating the Mean Function.
4. Nonparametric Estimation of Variogram.
5. Variogram Model Fitting.
6. Spatial Regression.
7. Kriging and Bayesian kriging
8. Multivariate Geostatistics.
9. Spatial-Temporal Processes.
10. Nonstationary Processes.

1. Exploratory Data Analysis

Traditional Numerical Summaries:

Mean, median, mode, standard deviation, range, interquartile range (IQR). They reduce the data to a few numbers, not useful for geostatistical data, because the numerical summaries ignore the location.

Stem-and-leaf plot:

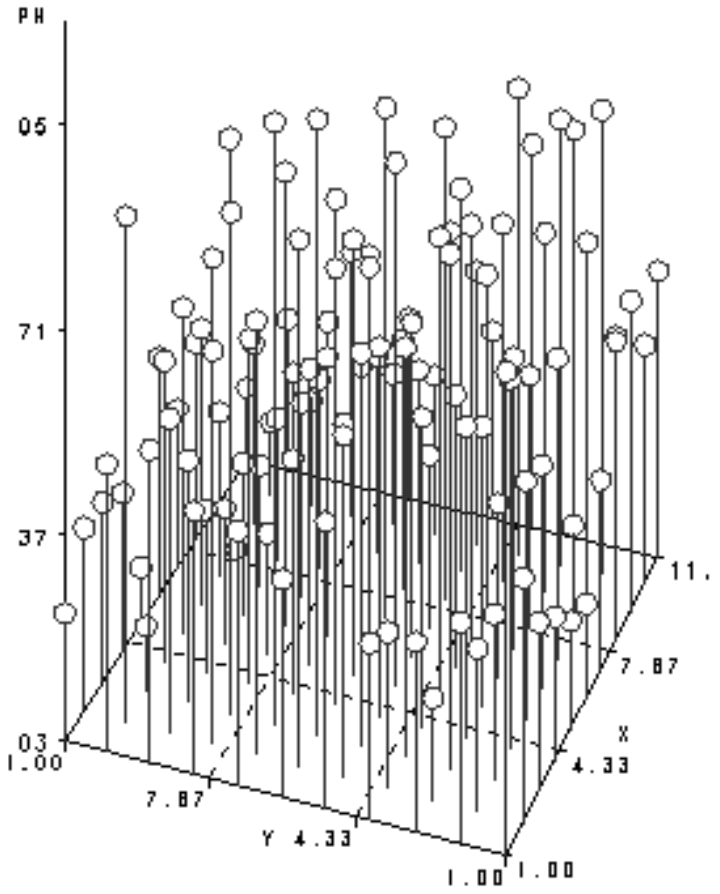
Gives more information about the distribution of the data (if the data are iid), but location is ignored.

Graphs to explore the large-scale structure:

Let Z be the spatial process of interest (eg. ozone concentrations), and Z_i the value of Z at a particular location.

- 3-D Scatter plot: a plot of Z_i versus location. This presents the variable's values over space.
- Plot of Z_i versus each marginal coordinate (latitude, longitude).
- Plot of mean or median of Z_i versus row index or column index.
- 2-D scatterplot of data locations with symbols indicating whether Z_i is above or below the median.
- Contour plot of Z_i. This needs some smoothing.

Three Dimensional Scatter Plot



Methods to explore the large-scale structure:

Mean polish or median polish of Z

1. Assume data on a $p \times q$ rectangular grid $\{(x_l, y_k) : k = 1, \ldots, p; l = 1, \ldots, q\}$. Regard grid nodes as cells in a 2-way table.
2. Operate iteratively on the data. Alternately subtracting row means (or medians) and column means (or medians) and accumulate these means or medians in an extra column and row of cells.
3. Repeat this procedure until another iteration produces virtually no change.
4. Final entries in the extra cells are the mean or median polish estimates or row effects $r_1, \ldots, r_p$, column effects $c_1, \ldots, c_q$ and an overall effect a.
5. Final entries in the body of table are residuals, $\hat{e}_{kl}$, such that

$$Z(x_l, y_k) = \hat{a} + \hat{r}_k + \hat{c}_l + \hat{e}_{kl}$$

(Numerical example presented in class)

Methods to explore the small-scale dependence:

1. h-scatterplots (lag scatterplots)

- Assume we observe the process Z at equally spaced locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$.
- Plot $Z(\mathbf{s}_i + h\mathbf{e})$ versus $Z(\mathbf{s}_i)$ for a fixed vector $\mathbf{e}$ (defines direction) of unit length, a fixed scalar h, and for all $i = 1, \ldots, n$ (e.g. $\mathbf{e} = (1, 0)$, h=1)
- Outliers can be detected with this graph. The graph may reveal the existence of anisotropy and/or nonstationarity in the mean and/or variance.

(Numerical example will be presented in class)

2. Variogram cloud

- Plot $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$ versus $\|\mathbf{s} - \mathbf{s}_j\|^{1/2}$ (Euclidean distance) for all pairs of observations.
- It may be advisable to bin the lags and plot a boxplot for each bin.
- The square-root differences $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^{1/2}$ are more resistant to outliers.
- The variogram cloud implicitly assumes isotropy (does not differentiate any directions)

3. Sample semivariogram

- The traditional sample semivariogram $\hat{\gamma}$ suggested by Matheron (1971) is:

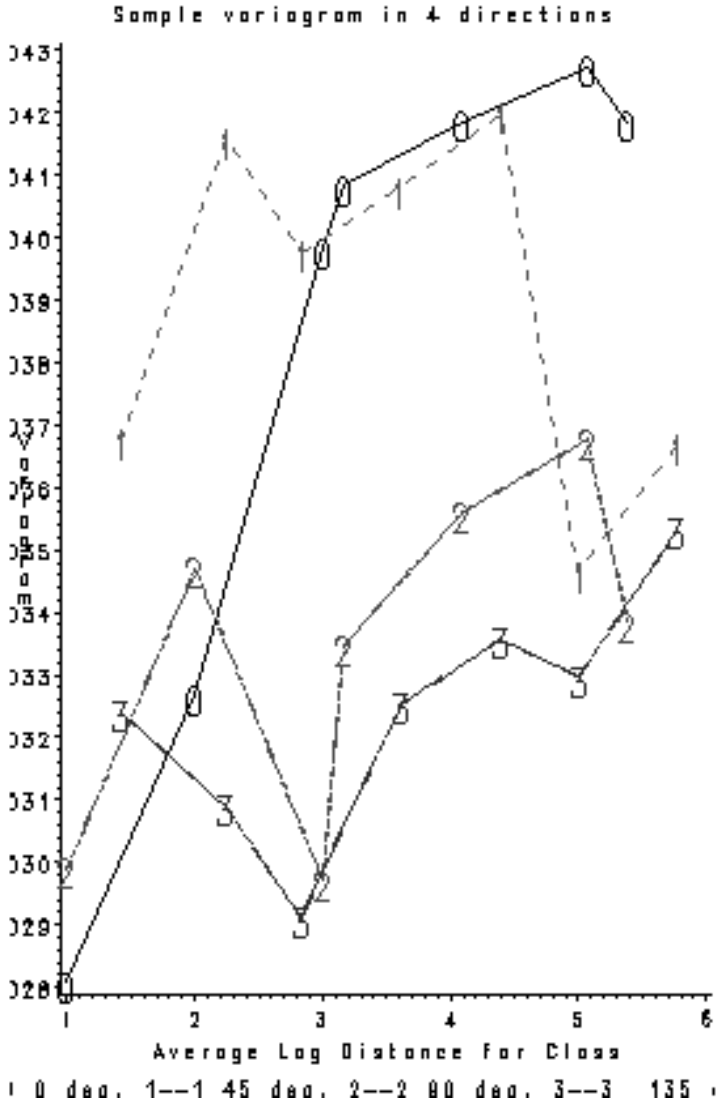
$$\hat{\gamma}(\mathbf{v}) = \frac{1}{2N(\mathbf{v})} \sum_{N(\mathbf{v})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2. \quad (1)$$

  where $N(\mathbf{v})$ are the number of data pairs $\mathbf{s}_i$ and $\mathbf{s}_j$ separated by $\mathbf{v}$.

- Plot $\hat{\gamma}(\mathbf{v})$ versus different values of $\mathbf{v}$.

- Note that this implicitly assumes stationarity of some kind.

- You can display the variogram along selected directions (e.g., N-S, NW-SE, E-W, and NE-WE) on the same 2-D graph.


Sample variogram in 4 directions
Average Log Distance for Class
0 deg. 1--1 45 deg. 2--2 90 deg. 3--3 135

3. Sample autocovariance

- Similar to sample semivariogram
- Plot of $\hat{C}(\mathbf{v})$ versus $\mathbf{v}$, where

$$\hat{C}(\mathbf{v}) = \frac{1}{N(\mathbf{v})} \sum_{N(\mathbf{v})} (Z(\mathbf{s}_i) - \bar{Z})(Z(\mathbf{s}_j) - \bar{Z}).$$

$\bar{Z}$ is the sample mean.

4. Pocket plot

- Data lie on a grid. For each pair of rows j and $k = j + h$ define

$$\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{l} |Z(\mathbf{s}_{jl}) - Z(\mathbf{s}_{kl})|^{1/2}.$$

  and

$$\tilde{Y}_{h} = \frac{1}{N_{h}} \sum_{i,l} |Z(\mathbf{s}_{il}) - Z(\mathbf{s}_{i+h,l})|^{1/2}.$$

  where N_{jk} and N_h are the number of such data differences that exist.

- Then define $P_{jk} = \bar{Y}_{jk} - \tilde{Y}_h$.

- For each fixed j, do a boxplot of $\{P_{jk} : k = 1, 2, \dots\}$ and put the boxplots of different j side by side, using row index (j) as abscissa.

- Study unusual rows "pockets" of nonstationarity.

- Can do the same thing for columns.

Methods to explore the small-scale variability:

- 3-D plot of standard deviation versus spatial location (using moving windows). May reveal nonstationarity in variability.
- Scatterplot of standard deviation versus mean (using moving windows). May also reveal nonstationarity in variability.

Methods to detect outliers:

Outliers: atypical observations. In geostatistical data, outliers may be of two types:

- (a) Distributional outliers. Observations that seem unusual with respect to the data's overall distribution.
- (b) Spatial outliers. observations that may not be unusual with respect to the data's overall distribution, but are unusual with respect to their neighbors.
- Stem-and-leaf display can identify distributional outliers.
- **Outlier-detect statistic**. Compute for each row and column,

$$n^{1/2}|\bar{Z} - med(Z_i)|/\sigma(.5708)^{1/2}$$

where $\sigma = IQR/1.349$, $\bar{Z}$ is the sample mean, and $med(Z_i)$ is the sample median within that row or column. Assess as a

standard normal deviate. Values of 3 or larger are of concern.

- Plot of each datum versus its nearest neighbor (or versus the average of its m nearest neighbors). This can detect spatial outliers.
- Median polish: large residuals correspond to outliers.

GEOSTATISTICS

1. Exploratory Data Analysis
2. *** Geostatistical Models.**
3. Estimating the Mean Function.
4. Nonparametric Estimation of Variogram.
5. Variogram Model Fitting.
6. Spatial Regression.
7. Kriging.
8. Multivariate Geostatistics.
9. Spatial-Temporal Processes.
10. Nonstationary Processes.

Geostatistical Models

1. Model Form:

A very popular statistical model for many kinds of data:

Datum = mean + residual

mean: *nonrandom* quantity (a number).

residual: random variable with mean zero, the residuals have all the *same variance* and are *independent.*

Geostatistical model:

$$Z(\mathbf{s}) = m(\mathbf{s}) + \epsilon(\mathbf{s})$$

Here:

- Large-scale variation: $m(\mathbf{s}) \equiv E[Z(\mathbf{s})]$
- Small-scale variation: $\{\epsilon(\mathbf{s}) : \mathbf{s} \in D\}$ is a zero-mean random field, associated with a *covariance function* C:

$$C(\mathbf{s}, \mathbf{t}) \equiv \text{cov}\{Z(\mathbf{s}), Z(\mathbf{t})\} \equiv \text{cov}\{\epsilon(\mathbf{s}), \epsilon(\mathbf{t})\}$$

The covariance function C satisfies:

1. *Symmetry,* i.e. $C(\mathbf{s}, \mathbf{t}) = C(\mathbf{t}, \mathbf{s})$ for all $\mathbf{s}, \mathbf{t} \in D$.

2. *Nonnegative definiteness,*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i, \mathbf{s}_j) \geq 0$$

   for all n, all sequences $\{a_i : i = 1, \ldots, n\}$ and all sequences of spatial locations $\{\mathbf{s}_i : i = 1, \ldots, n\}$. *(Nonnegative definiteness).*

2. Stationarity:

Three types of stationarity:

1. **Strict stationarity**. The joint probability distribution of the data depends only on the relative positions of the sites at which the data were taken, i.e. the joint distribution of

   $$(Z(\mathbf{x}_1), Z(\mathbf{x}_2), \ldots, Z(\mathbf{x}_m))$$

   is the same as

   $$(Z(\mathbf{x}_1 + \mathbf{v}), Z(\mathbf{x}_2 + \mathbf{v}), \ldots, Z(\mathbf{x}_m + \mathbf{v}))$$

   for any m spatial points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and any $\mathbf{v} \in D$.

2. **Weak stationarity** (second-order stationarity).
   - the mean is constant, $m(\mathbf{s}) \equiv \beta_0$
   - the covariance at two sites depends on only the sites **relative positions**.

     $$C(\mathbf{s}, \mathbf{t}) = C(\mathbf{s} - \mathbf{t})$$

3. **Intrinsic stationarity**. This is the more general stationarity
   - the mean is constant, $m(\mathbf{s}) \equiv \beta_0$
   - the $\text{var}[Z(\mathbf{s}) - Z(\mathbf{t})]$ depends on only the sites **relative positions** $\mathbf{s} - \mathbf{t}$, i.e. there is a function γ such that:

$$\text{var}[Z(\mathbf{s}) - Z(\mathbf{t})] = 2\gamma(\mathbf{s} - \mathbf{t})$$

3. Mean functions:

It is reasonable to assume that sites close to one another should have *similar means* but sites far apart need not. This kind of *local stationarity*, rather than *global stationarity*, leads to the postulation of a continuous, relatively smooth (but non-constant 0 function for the mean).

- A class of mean functions are the polynomials, i.e

  $$m(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

  or

  $$m(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_{11} x^2 + \beta_{12} xy + \beta_{22} y^2$$

  We can write the mean as $m(\mathbf{s}; \boldsymbol{\beta})$ to emphasize the dependence of unknown parameters $\boldsymbol{\beta}$.

- Another continuous (but less smooth) function is the surface that results from performing a median polish

- We can also model the mean function with

a non-parametric approach (splines).

4. Covariance functions

Issues in choosing a model for the covariance function:

- Symmetry.
- Nonnegative definiteness.

We mainly (though not exclusively) will consider covariance functions that have one or more of the following additional properties:

1. $C(\mathbf{h})$ decreases as $\|\mathbf{h}\|$ increases.
2. $C(\mathbf{h}) \to 0$ as $\|\mathbf{h}\|$ increases.
3. $C(\mathbf{h}) \geq 0$ for all $\mathbf{h}$

Still further, issues such as model parsimony, computational efficiency, and replication suggest that, if possible, we strongly consider models that have the following properties:

1. **Isotropy** The covariance between any two values depends only on the Euclidean distance between their corresponding locations:

$$C(\mathbf{h}) = C(\|\mathbf{h}\|)$$

for all $\mathbf{h}$

Types of *anisotropy*:

- *Geometric anisotropy.* When exist a matrix $\mathbf{A}$ such that

$$C(\mathbf{h}) = C([\mathbf{h}'\mathbf{A}\mathbf{h}]^{1/2})$$

  $\mathbf{A}$ is a rotation and stretching of the axis.

- *Zonal anisotropy* (when it is not geometric)

2. **Separability** A (2-D) covariance function C is said to be separable if

$$C(\mathbf{h}) = C(h_1, h_2) = C_1(h_1)C_2(h_2)$$

   for two valid (in $\mathbb{R}$) covariance functions C_1 and C_2.

Examples and a **handout** of isotropic covariance function will be presented in class:

Triangular model (Tent model), Spherical model, Exponential model, Gaussian model, Rational quadratic model, Matern class of models, Cosine model, Wave or hole-effect model.

Some notes about these covariance functions:

- The tent, exponential, spherical, Gaussian, rational quadratic, and Matern models all decrease monotonically to 0 as distance increases.
- The wave, or hole-effect model, also tends to 0 but not monotonically.
- The cosine model does not tend to 0 as distance increases.
- In the Matern class, K_{θ_3} is a function called the modified Bessel function of the third kind of order θ_3
- The exponential model is a special case of

the Matern model with $\theta_3 = 1/2$, the Gaussian model is the limiting case of the Matern model as $\theta_3 \to \infty$.

Attributes of these covariance functions models

- **Scale parameter or variance**. $C(0) = \theta_1$. Thus θ_1 is the variance of the random field.
- **Correlation scale parameter**. θ_2 controls how the covariance changes.
- **Shape or smoothing parameter**. For the Matern model θ_3 controls the *shape* of the function.
- **Range, or effective range**. The distance beyond which the covariance function is equal to 0. Only the tent and spherical models have a range(θ_2). The effective range, is defined as the distance beyond which the covariance function does not exceed $0.04 \times$ variance. The exponential, Gaussian, rational quadratic, and Matern models all have effective ranges; the cosine does not.
- **Continuity at 0**. All these covariance functions are continuous at 0. This is not a

required property, but it has implications for the behavior of the random field (as we will see shortly).

From basic covariance models, we can construct more complicated models using the following rules:

- If C_1 and C_2 are valid covariances, then so is $C(\cdot) \equiv C_1(\cdot) + C_2(\cdot)$.
- If C_0 is a valid covariance and $b > 0$, then $C(\cdot) \equiv b \cdot C_0(\cdot)$ is a valid covariance.
- If C_1 and C_2 are valid covariances, then so is $C(\cdot) \equiv C_1(\cdot) \cdot C_2(\cdot)$.
- A valid isotropic covariance function in $\mathbb{R}^{d_1}$ may not be a valid isotropic covariance function in $\mathbb{R}^{d_2}$ where $d_2 > d_1$. The converse is true. With the exception of the tent model, all the models just listed are valid in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Nugget effect

Sometimes a portion of the variation in the data is attributable to measurement error:

$$\epsilon(\mathbf{s}) = \epsilon_{ME}(\mathbf{s}) + \epsilon_{NE}(\mathbf{s})$$

where

- $\epsilon_{ME}(\mathbf{s})$ is the measurement error at site $\mathbf{s}$.
- $\epsilon_{NE}(\mathbf{s})$ is what remains of the residual at $\mathbf{s}$ after subtracting the measurement error.

The covariance for ϵ is then represented as

$$C(\mathbf{h};\theta) = \begin{cases} \theta_0 + C_{NE}(\mathbf{h};\theta) & \text{if } \mathbf{h} = 0 \\ C_{NE}(\mathbf{h};\theta) & \text{if } \|\mathbf{h}\| > 0. \end{cases} \tag{2}$$

where C_{NE} is the covariance for ϵ_{NE}.

θ_0 is called the *nugget effect*.

5. Intrinsic Stationarity and the Semivariogram

Intrinsic stationarity:

- the mean is constant, $m(\mathbf{s}) \equiv \beta_0$
- there is a function γ such that:

$$\frac{1}{2}\text{var}[Z(\mathbf{s}) - Z(\mathbf{t})] = \gamma(\mathbf{s} - \mathbf{t})$$

  γ is called semivariogram.

Some facts:

- The semivariogram can also be expressed as follows:

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

  where $\mathbf{s} - \mathbf{t}$.

- A second-order stationary random process with covariance C is intrinsically stationary, with semivariogram

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

  the converse is not true in general.

Issues choosing a model for the semivariogram:

1. Vanishes at $\mathbf{0}$, i.e., $\gamma(\mathbf{0}) = 0$.
2. Evenness, i.e. $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$.
3. Conditional non-positive definiteness, i.e.

$$\sum_i \sum_j \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$$

   for all $\mathbf{s}_1, \mathbf{s}_2, \ldots$ and all $\lambda_1, \lambda_2, \ldots$ such that $\sum_i \lambda_i = 0$.
4. $\lim_{\|\mathbf{h}\| \to \infty} \{\gamma(\mathbf{h})/|\mathbf{h}\|^2\} = 0$.

Note that separability of the covariance function does not imply that

$$\gamma(\mathbf{h}) = \gamma_1(h_1)\gamma_2(h_2)$$

Semivariogram attributes:

- Sill (variance of the random field)
- Range or effective range
- Nugget effect
- Slope

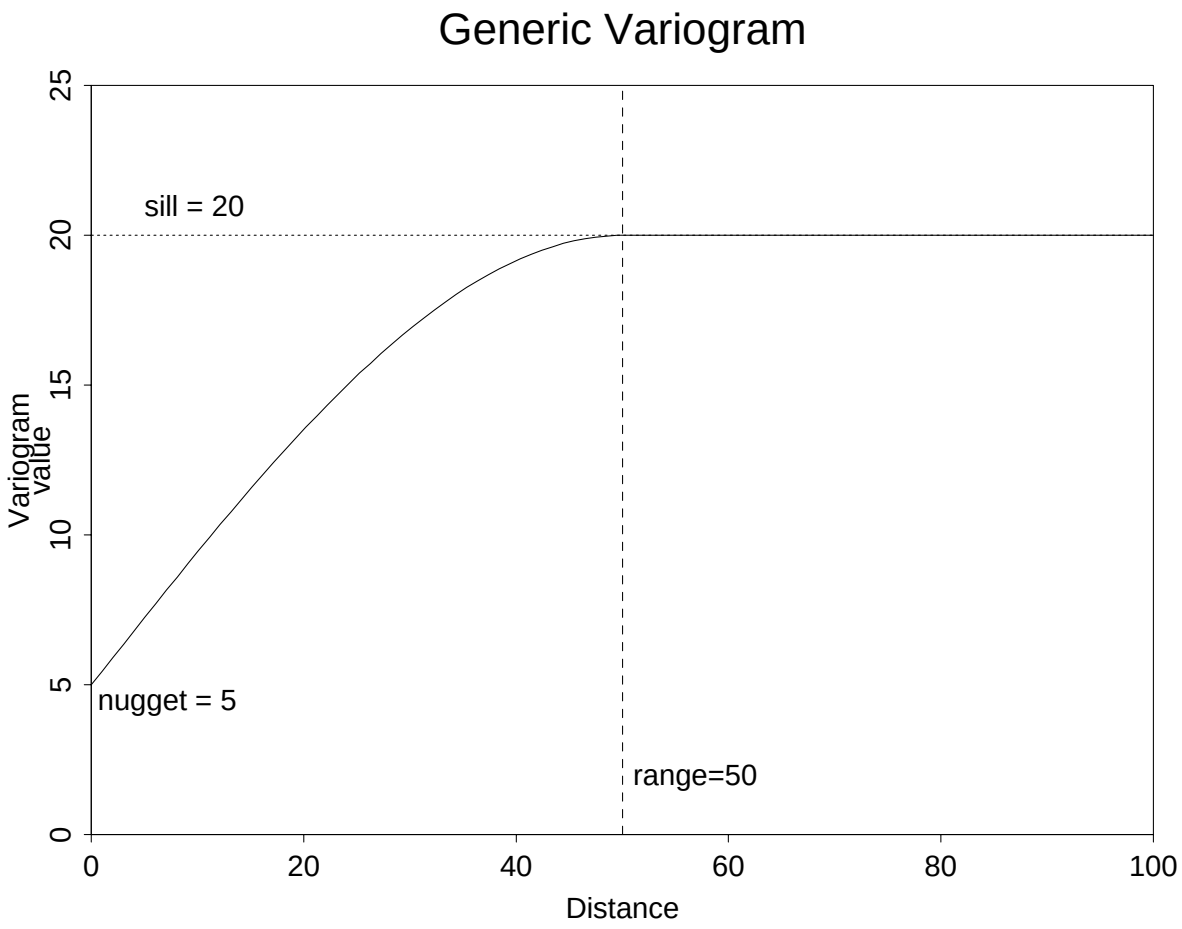


Figure 3: Exponential variogram.

Examples and a **handout** of isotropic semivariogram functions will be presented in class:

Triangular model (Tent model), Spherical model, Exponential model, Gaussian model, Rational quadratic model, Matern class of models, Cosine model, Wave or hole-effect

model, Linear and Power models.

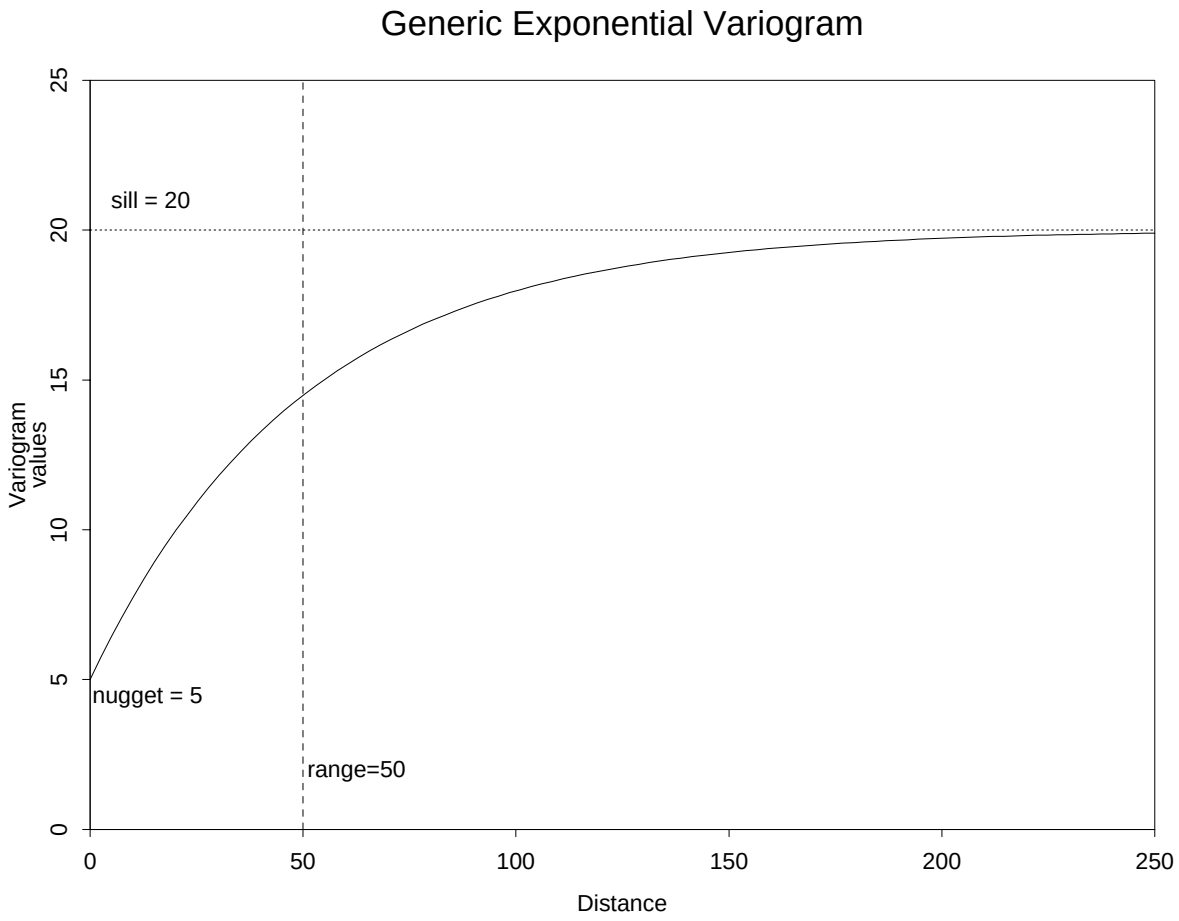


Figure 4: Exponential variogram.

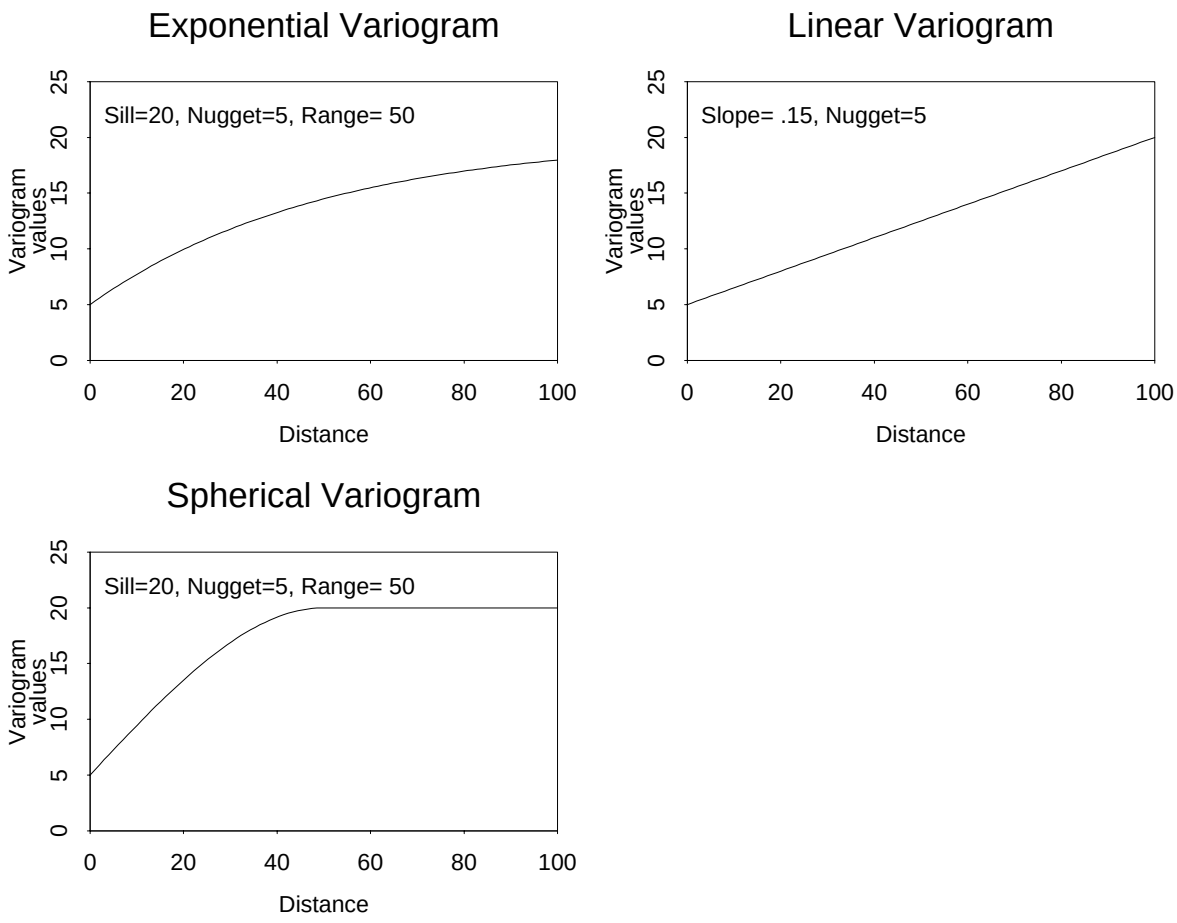


Figure 5: Exponential, linear and spherical variograms.

The Nugget effect for semivariograms. If

$$\epsilon(\mathbf{s}) = \epsilon_{ME}(\mathbf{s}) + \epsilon_{NE}(\mathbf{s})$$

The semivariogram for ϵ is then represented as:

$$\gamma(\mathbf{h};\theta) = \begin{cases} 0 & \text{if } \mathbf{h} = 0 \\ \theta_0 + \gamma_{NE}(\mathbf{h};\theta) & \text{if } \|\mathbf{h}\| > 0 \end{cases} \quad (3)$$

where γ_{NE} is the covariance for ϵ_{NE}.

6. Modeling anisotropy

(a) Range anisotropy. The range is direction-dependent.

- Kind of anisotropy seen most often
- Geometric anisotropy is the easiest to model.

(b) Sill anisotropy. It can be shown that if the sill exists but is direction-dependent. Then either:

- a second-order stationary model is appropriate but the spatial correlation does not vanich in every direction as inter-site distance increases;
- the second-order stationarity is violated; or
- there are measurement errors which are correlated or do not all have mean zero.

(c) Nugget anisotropy.

- Can be caused by correlated measurement errors.

- Typically occurs in one direction only, which is not difficult to model.

(d) Slope anisotropy.

- Can be dealt with in a similar fashion as geometric range anisotropy.

GEOSTATISTICS

1. Exploratory Data Analysis
2. Geostatistical Models.
3. *** Estimating the Mean Function.**
4. Nonparametric Estimation of Variogram.
5. Variogram Model Fitting.
6. Spatial Regression.
7. Kriging.
8. Multivariate Geostatistics.
9. Spatial-Temporal Processes.
10. Nonstationary Processes.

Estimating the Mean Function

The mean function $m(\mathbf{s};\boldsymbol{\beta})$ is a linear (or nonlinear) function of the elements of $\boldsymbol{\beta}$, then linear (or nonlinear) least squares can be used to fit the model to the data. This is called *trend surface analysis.*

Example of linear family of models, for $\mathbf{s} = (x, y)$:

$$m(\mathbf{s};\boldsymbol{\beta}) = \sum_{i=1}^{p} \beta_i f_i(\mathbf{s}),$$

when $p = 3$, $f_1(\mathbf{s}) = 1$, $f_2(\mathbf{s}) = x$, and $f_3(\mathbf{s}) = y$, we get a first-order polynomial

$$m(\mathbf{s};\boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 y,$$

e.g. of nonlinear in $\boldsymbol{\beta}$:

$$m(\mathbf{s};\boldsymbol{\beta}) = log(\beta_1 + x) + \mathrm{e}^{\beta_1 y}.$$

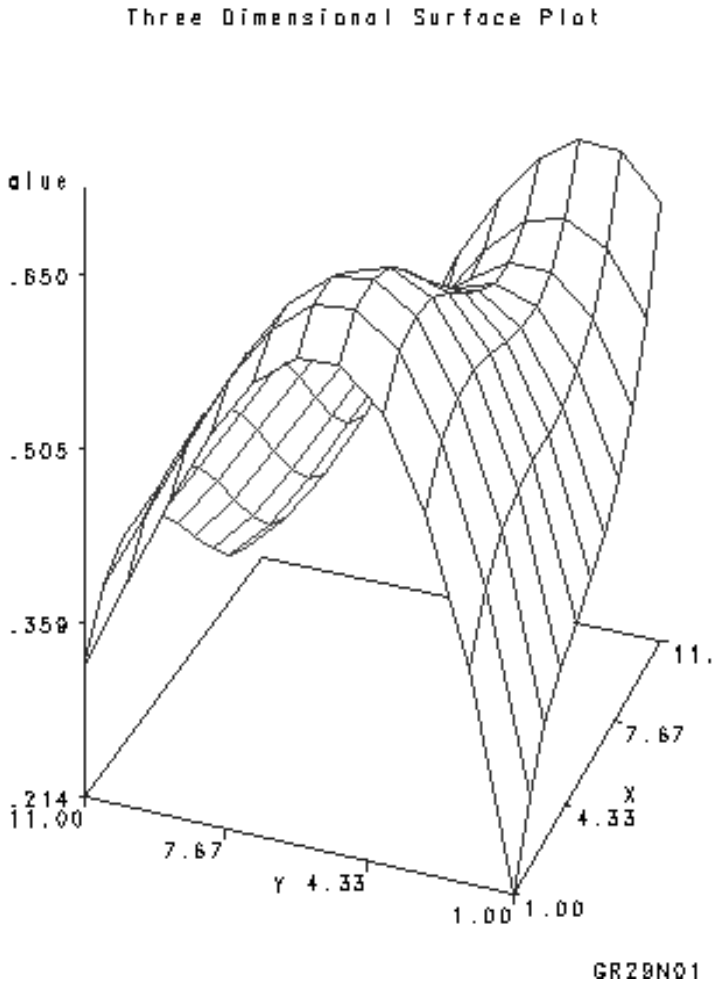


Figure 6: Application to soil PH data (a third-order polynomial surface), $m(\mathbf{s}, \boldsymbol{\beta}) = \beta_o + \beta_1 x + \beta - 2y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2 + \beta_6 x^3 + \beta_7 x^2 y + \beta_8 x y^2 + \beta_9 y^3$.

This approach is quite easy to implement, with PROC REG in SAS, and in R. It does have some drawbacks:

- the fitting procedure is very sensitive to outliers;
- the regression variables (e.g. spatial coordinates) tend to be highly correlated, which causes the fitting procedure to be numerically unstable;

Two alternative approaches:

- Locally weighted least squares (LOESS)
  - Only assumes that the mean function is smooth
  - Estimates this smooth trend in a moving fashion by fitting a site specific polynomial.
  - Fits using weighted least squares, with weights inversely related to distance from the site.
- Median polish

Median polish takes the model for the mean function to be:

$$m(x_l, y_k; \boldsymbol{\beta}) = a + r_k + c_l.$$

Earlier we saw how median polish fits this model to the data. The fitted values at data locations are:

$$m(x_l, y_k; \hat{\boldsymbol{\beta}}) = \hat{a} + \hat{r}_k + \hat{c}_l.$$

The fitted surface over the remainder of the spatial domain is obtained by linearly interpolating. For $\mathbf{s} = (x, y)$ between the four nodes (x_l, y_k), (x_{l+1}, y_k),(x_l, y_{k+1}),(x_{l+1}, y_{k+1}), the fit is given by the planar interpolant:

$$\hat{a} + \hat{r}_k + \left(\frac{y - y_k}{y_{k+1} - y_k}\right)(\hat{r}_{k+1} - \hat{r}_k) + \hat{c}_l$$

$$+ \left(\frac{x - x_l}{x_{l+1} - x_l}\right)(\hat{c}_{l+1} - \hat{c}_l).$$

(Cressie Pag. 188 and pag. 219).

Median polish avoids some of the pitfass of least squares:

- more flexible
- more resistant to outliers
- residuals from a median polish are less biased than those from OLS

However, the assumed row-column additivity and the blurring associated with irregularly spaced data are limitations of the method.

GEOSTATISTICS

1. Exploratory Data Analysis
2. Geostatistical Models.
3. Estimating the Mean Function.
4. *** Nonparametric Estimation of Variogram.**
5. Variogram Model Fitting.
6. Spatial Regression.
7. Kriging.
8. Multivariate Geostatistics.
9. Spatial-Temporal Processes.
10. Nonstationary Processes.

Estimation of Variogram

1. Semivariogram estimation

The raw ingredients for semivariogram estimation are either:

- the observations $\{Z_1, \ldots, Z_n\}$, if the mean function is taken to be constant;
- the residuals

$$\hat{\epsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - m(\mathbf{s}_i; \hat{\boldsymbol{\beta}}), \quad (i = 1, \ldots, n)$$

  from a fitted mean function at the data locations, otherwise.

Assume for simplicity that the data locations are points on a regular rectangular grid.

The basic idea is to estimate $\gamma(\mathbf{h})$ by one-half the average squared difference of responses or residuals whose data locations are lagged by $\mathbf{h}$.

The traditional sample semivariogram $\hat{\gamma}$ (also

empirical or estimated semivariogram):

$$\hat{\gamma}(\mathbf{h}_u) = \frac{1}{2N(\mathbf{h}_u)} \sum_{N(\mathbf{h}_u)} (\hat{\epsilon}(\mathbf{s}_i) - \hat{\epsilon}(\mathbf{s}_j))^2$$

$$(u = 1, \ldots, k)$$

where $N(\mathbf{h}_u)$ are the number of data pairs $\mathbf{s}_i$ and $\mathbf{s}_j$ separated by $\mathbf{h}_u$.

- Here $\mathbf{h}_1, \ldots, \mathbf{h}_k$ are the distinct values of $\mathbf{h}$ represented in the data set.
- $N(\mathbf{h}_u)$ is the number of times that lag $\mathbf{h}_u$ occurs in the data set. (We don't double-count.)
- This is a method-of-moments type estimator.
- The estimator is **biased** when the observations themselves are used $(Z_1, \ldots, Z_n)$ it there is a non-constant mean (the mean depends on location). It is approximately unbiased when the residuals are used.

- If the mean is constant $m(\mathbf{s}_i; \boldsymbol{\beta}) = \alpha$, then the estimator is **unbiased** when the observations themselves are used.

When data locations are irregularly spaced, we partition the *lag space* $H = \{\mathbf{s} - \mathbf{t} : \mathbf{s}, \mathbf{t} \in D\}$ into lag classes or *windows* $H_1, \ldots, H_k$, and assign each lag in the data set to one of these classes.

$$\hat{\gamma}(H_u) = \frac{1}{2N(H_u)} \sum_{N(H_u)} (\hat{\epsilon}(\mathbf{s}_i) - \hat{\epsilon}(\mathbf{s}_j))^2$$

$$(u = 1, \ldots, k)$$

- Here $\mathbf{h}_u$ is a representative lag for the whole lag class H_u; typically $\mathbf{h}_u$ is taken to be the centroid of H_u or the average of all the realized lags in the lag class.
- $N(H_u)$ is the class frequency of H_u
- The estimator (assuming the right mean function is fitted) is approximately unbiased; it is not exactly unbiased even in the case where the observations themselves are used because the grouping of lags into classes causes a blurring effect.

- Two main types of partition (examples presented in class):
  - *Polar* partitioning, i.e., angle and distance classes
  - Rectangular partitioning

Remarks:

- Generally we construct a plot of these estimates corresponding to each of several directions.
- The polar partition more naturally allows for the construction of directional semivariograms.

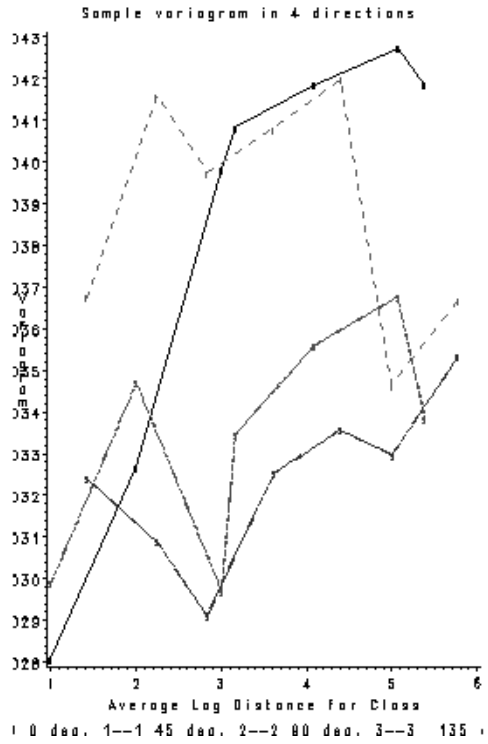


Figure 7: Semivariograms based on polar partition and residuals of a second-order fit for the mean

- Note that we only have estimates of $\gamma(\mathbf{h})$ for a finite number of lags.
- How many lag classes (i.e. how fine a partition) should we use? A rule of thumb will be given later, but there is no harm in trying several different partitions.
- An alternative and more robust (less sensitive to outliers) estimator, proposed by Cressie and Hawkins (1980, *Journal of the International Association for Mathematical Geology*), is

$$\bar{\gamma}(\mathbf{h}_u) = \frac{\left\{\frac{1}{N(H_u)} \sum_{N(H_u)} |\hat{\epsilon}(\mathbf{s}_i) - \hat{\epsilon}(\mathbf{s}_j)|^{1/2}\right\}^4}{.914 + [.988/N(H_u)]}$$

$$(u = 1, \ldots, k)$$

Handout will be given with eight examples of sample semivariograms.

2. Estimation of Covariance.

The usual estimator is

$$\hat{C}(\mathbf{h}) = \frac{1}{N(H_u)} \sum_{N(H_u)} (Z(\mathbf{s}_i) - \bar{Z})(Z(\mathbf{s}_j) - \bar{Z})$$

$\bar{Z}$ is the sample mean.

Remarks:

- This estimator is meaningful only if the process is *second-order stationary*; otherwise it's estimating something that doesn't exist.
- This estimator is the spatial generalization of the sample autocovariance function used by time series analysts.

Comparison with semivariogram estimation:

- $\hat{\gamma}(\mathbf{h}) \neq \hat{C}(\mathbf{h}) - \hat{C}(\mathbf{h})$, but the difference is usually small for large n.
- If the estimates are based on the observations themselves, then $\hat{C}(\mathbf{h})$ is **biased** even when the mean is constant and then $\hat{\gamma}(\mathbf{h})$ is **unbiased**.
- If the estimates are based on residuals from a fitted mean function, then $\hat{\gamma}(\mathbf{h})$ is less biased.
- If there is a trend in the data that is not removed, $\hat{\gamma}(\mathbf{h})$ is not as badly biased.

3. Checking for Isotropy.

Prior to fitting a parametric model to the sample semivariogram, we need to determine how the semivariogram depends on the relative orientation of data locations. We want to investigate whether isotropy would be a reasonable assumption for the data.

Two graphical diagnostics:

- 1. At least three (preferably more) directions are needed to distinguish geometric anisotropy.
- 2. Rose diagram
  - Consists of smoothing the directional sample semivariograms, then in the lag space connecting, with a smooth curve, those lag vectors $\mathbf{h}_u$ for which these smoothed semivariograms are roughly equal.
  - In effect, this plots estimated isocorrelation contours (in the case of a second-order stationary process).

– Circular curves $\Rightarrow$ isotropy;
  elliptical curves$\Rightarrow$ geometric anisotropy.

(Examples presented in class)

GEOSTATISTICS

1. Exploratory Data Analysis
2. Geostatistical Models.
3. Estimating the Mean Function.
4. Nonparametric Estimation of Variogram.
5. *** Variogram Model Fitting.**
6. Spatial Regression.
7. Kriging.
8. Multivariate Geostatistics.
9. Spatial-Temporal Processes.
10. Nonstationary Processes.

Variogram Model Fitting

The next step of the geostatistical method is to select and fit a parametric family of models to the sample semivariogram. Why aren't we satisfied with just the sample semivariogram itself?

- The sample semivariogram may violate the required property of conditional negative definiteness.
- For various purposes (e.g. kriging) we may require an estimate of the semivariogram at a lag not represented in the data.
- The sample semivariogram may be quite bumpy. A smoothed version may be helpful for understanding the nature of the spatial dependence.

Let $\gamma(\mathbf{h}; \boldsymbol{\theta})$ denote the parametric model to be fit to the sample semivariogram and let Θ denote the parameter space for $\boldsymbol{\theta}$.

1. Methods of Fitting

- (a) By eye.
- (b) Ordinary nonlinear least squares (not appropriate: because the lag classes have different variability and they are correlated).
- (c) Weighted nonlinear least squares (Cressie, 1985, Mathematical Geology)

  A weighted nonlinear least squares estimator of $\gamma(\mathbf{h};\boldsymbol{\theta})$ is defined as a value $\hat{\boldsymbol{\theta}} \in \Theta$ that minimizes the weighted residual sum of squares function:

$$\omega(\boldsymbol{\theta}) = \sum_{u \in U} \frac{N(\mathbf{h}_u)}{[\gamma(\mathbf{h}_u;\boldsymbol{\theta})]^2} [\hat{\gamma}(\mathbf{h}_u) - \gamma(\mathbf{h}_u;\boldsymbol{\theta})]^2.$$

  Here:

  - U is a specified subset of lag classes believed to yield reliable estimates of $\gamma(\mathbf{h})$. Generally U is taken to be of the form $U = \{u; N(\mathbf{h}_u) \geq G_1, \|\mathbf{h}_u\| \leq G_2\}$;

one rule-of-thumb is to take $G_1 = 30$ and $G_2 =$ half the largest lag in the data.

– The weights $\frac{N(\mathbf{h}_u)}{[\gamma(\mathbf{h}_u;\boldsymbol{\theta})]^2}$ are small if either $N(\mathbf{h}_u)$ is small or $\gamma(\mathbf{h}_u;\boldsymbol{\theta})$ is large. Thus, nonparametric estimates at large lags tend to receive relatively less weight.

- (d) Generalized nonlinear least squares
  - For two lags $\mathbf{h}_1$ and $\mathbf{h}_2$, $\hat{\gamma}(\mathbf{h}_1)$ and $\hat{\gamma}(\mathbf{h}_2)$ are dependent for two reasons:
    - They may be functions of some of the same observations;
    - Even if they have no observations in common, observations in the one estimate are generally spatially correlated with observations in the other.
  - Consequently we may want to consider a Generalized Nonlinear Least Squares approach, in which we minimize

    $$GRSS(\boldsymbol{\theta}) = [\hat{\gamma} - \gamma(\boldsymbol{\theta})]' [\hat{\text{var}}(\hat{\gamma})]^{-1} [\hat{\gamma} - \gamma(\boldsymbol{\theta})].$$

  - Here $\hat{\gamma}$ is the vector of nonparametric semivariogram estimates and $\gamma(\boldsymbol{\theta})$ is the corresponding vector of modeled semivariogram values
  - Derivation and calculation of $\hat{\text{var}}(\hat{\gamma})$ can be a challenge.

- (e) Maximum likelihood (and restricted maximum likelihood)
  - Applicable to processes with second-order stationary errors only.
  - Estimates $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ simultaneously.
  - Let $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ denote the covariance matrix of $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and let $\mathbf{X}$ denote the model matrix for the model

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

  - The log-likelihood functions is

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{Z}) =$$

$$-\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}).$$

  - a MLE is a value $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ that maximizes $L(\boldsymbol{\beta}, \boldsymbol{\theta})$.
  - Generally a MLE must be found by numerical optimization routines (e.g. Newton-Raphson). Thus we have to be concerned with starting values, and convergence criteria.

– A restricted MLE (REML) is defined as a value $\boldsymbol{\theta} \in \Theta$ that maximizes the log-likelihood function associated with $n - \text{rank}(\mathbf{X})$ linearly independent error contrasts. It is less biased than MLE.

Examples presented in class (handout).

2. Comparison of Fitting methods

The relatively easy-to-compute weighted least squares estimator performs almost as well as the more complicated ML and REML estimators, and has been the estimator of choice for most practitioners. Some statisticians still prefer the less *ad hoc* ML and REML estimators, however.

3. Model Selection Procedures

- (a) Visual inspection of semivariogram plot
- (b) Minimized weighted (or generalized) residual sum of squares function, $\omega(\hat{\boldsymbol{\theta}})$.
- (c) Maximized log-likelihood (or restricted log-likelihood) function, $L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$.
- (d) Penalized likelihood criteria, e.g. Akaike's Information Criterion

  $AIC = L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - \#\text{of estimated parameters.}$

Example presented in class.

GEOSTATISTICS

1. Exploratory Data Analysis
2. Geostatistical Models.
3. Estimating the Mean Function.
4. Nonparametric Estimation of Variogram.
5. Variogram Model Fitting.
6. *** Spatial Regression.**
7. Kriging.
8. Multivariate Geostatistics.
9. Spatial-Temporal Processes.
10. Nonstationary Processes.

Spatial Regression

Generalized least squares (GLS) with known covariance matrix

- Model:

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ E(\boldsymbol{\beta}) = \mathbf{0},\ \text{var}(\boldsymbol{\epsilon}) = \mathbf{V}$$

  where $\mathbf{V}$ is completely specified positive definite matrix.

- Example in geostatistical context, in which $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$:

  Consider an isotropic spherical model with nugget=0.25 range=2, sill =2. Write $\mathbf{V}$.

- GLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}.$$

Estimated generalized least squares (EGLS)

- In practice we don't know the true value of $\boldsymbol{\theta}$ and consequently $\mathbf{V}$ cannot be completely specified. A natural solution to this problem is to replace $\boldsymbol{\theta}$ in the evaluation of $\mathbf{V}$ by an estimator $\hat{\boldsymbol{\theta}}$, thereby obtaining $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}})$.
- Illustration of this in class, using an isotropic spherical model with nugget=0.25 (θ_0), range=2 (θ_1), sill =2 (θ_3), and estimates: $\hat{\theta_0} = 0$, $\hat{\theta_1} = 2.2$ $\hat{\theta_2} = 2.5$. Obtaine the corresponding covariance model using the estimates, $C(\cdot, \hat{\boldsymbol{\theta}})$, and write $\hat{\mathbf{V}}$.
- EGLS estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{EGLS} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Z}.$$

Mean Structure or Covariance Structure?

One issue we've ignored to this points is that the choice, in practice, of a decomposition of the data into mean structure (large-scale variation) and covariance structure (small-scale variation) is not so clearcut. This is often phrased as follows:

"One man's structure is another man's covariance structure."

An illustration of this (for a 1-D process) is provided (handout given in class):

- *Trend by no correlation.* $Z(s) = -0.5 + 0.02s + \epsilon(s)$, where $\{\epsilon(s) : s_1, \ldots, 50\}$ are independent and identically distributed $N(0, 1)$ random variables.
- *Correlation but no trend.* $Z(s) = 0 + \epsilon$, where $\{\epsilon(s) : s_1, \ldots, 50\}$ are normally distributed random variables with mean 0 and covariance structure determined by the

exponential covariance function
$C(r) = \exp(-r/10)$.

If replications of a spatial process are available, statistical procedures exist for distinguishing between these two types of models. In practice, however, geostatistical data are not replicated so we must settle for plausibility, rather than a high degree of certainty, of the proposed decomposition.

GEOSTATISTICS

1. Exploratory Data Analysis
2. Geostatistical Models.
3. Estimating the Mean Function.
4. Nonparametric Estimation of Variogram.
5. Variogram Model Fitting.
6. Spatial Regression.
7. *** Kriging.**
8. Multivariate Geostatistics.
9. Spatial-Temporal Processes.
10. Nonstationary Processes.

Kriging

Spatial Prediction (Ordinary Kriging)

The origin of the word *kriging* is from D.G. Krige, a South African mining engineer who in the 1950's developed empirical methods for predicting ore grades at unsampled locations using the known grades of ore sampled at nearby sites.

Krige's original method is what is now called *ordinary kriging.* There have been several modifications and extensions (e.g., universal kriging, indicator kriging, disjunctive kriging, and others) but they are all based on very similar ideas.

The theory of ordinary kriging is based on the same geostatistical model we have been using all along, with tow important restrictions:

1. The mean $m(\mathbf{s})$ is assumed to be constant.
2. The semivariogram $\gamma(\mathbf{h})$ is assumed to be known.

Let $\mathbf{s}_0$ denote an arbitrary location in D usually this will be an unsampled location but it need not be.

Goal of kriging: to predict the value of $Z(\mathbf{s}_0)$ at $\mathbf{s}_0$.

Properties of ordinary kriging predictor:

1. It is a **linear** combination of the data values, i.e.,

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i).$$

2. It is **unbiased**, i.e., it satisfies

$$E[\hat{Z}(\mathbf{s}_0)] = E[Z(\mathbf{s}_0)].$$

3. Among all function of the data that satisfy the first 2 properties, it is the **best** in the sense that minimizes the variance of prediction error, $\text{var}[\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)]$.

Kriging is also known as the Best Linear Unbiased Predictor (BLUP).

Properties 1 and 2 above imply $\sum_{i=1}^{n} \lambda_i = 1$.

The properties we have imposed on our predictor lead us to minimize

$$\text{var}[\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)]$$

subject to the restriction

$$\sum_{i=1}^{n} \lambda_i = 1.$$

The method of Lagrange multipliers, from calculus, is applicable here. It can be used to show that the optimal coefficients $\lambda_1, \ldots, \lambda_n$ are the first n elements of the vector $\boldsymbol{\lambda}_O$ that satisfies *(the ordinary kriging equations)*:

$$\boldsymbol{\Gamma}_O \boldsymbol{\lambda}_O = \boldsymbol{\gamma}_O$$

where

$$\boldsymbol{\lambda}_O = (\lambda_1, \ldots, \lambda_n, m)'$$

$$\boldsymbol{\gamma}_O = [\gamma(\mathbf{s}_1 - \mathbf{s}_0), \ldots, \gamma(\mathbf{s}_n - \mathbf{s}_0)]'$$

$$\mathbf{\Gamma}_O = \begin{cases} \gamma(\mathbf{s}_i - \mathbf{s}_j) & \text{for } i = 1, \ldots, n; j = 1, \ldots, n \\ 1 & \text{for } i = n+1; j = 1, \ldots, n \\ 0 & \text{for } i = n+1; j = n+1 \end{cases} \tag{4}$$

and m is a Lagrange multiplier and $\mathbf{\Gamma}_O$ is symmetric.

The minimized variance, called the *kriging variance*, is

$$\sigma^2_{OK}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_i - \mathbf{s}_0) + m = \boldsymbol{\lambda}'_O \boldsymbol{\gamma}_O.$$

Example presented in class with $\gamma(\|\mathbf{h}\|) = 1 - \exp(-\|\mathbf{h}\|/2)$.

Define:

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)',$$

$$\boldsymbol{\gamma} = (\gamma(\mathbf{s}_1 - \mathbf{s}_0), \ldots, \gamma(\mathbf{s}_n - \mathbf{s}_0)),$$

$$\mathbf{\Gamma} = \{\gamma(\mathbf{s}_i - \mathbf{s}_j)\}.$$

Prediction intervals

A $100(1-\alpha)\%$ prediction interval for $Z(\mathbf{s}_0)$ is as follows:

$$\hat{Z}(\mathbf{s}_0) \pm z_{\alpha/2}\sigma_{OK}(\mathbf{s}_0),$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

Some remarks:

- Ordinary kriging is derived under an assumption of constant mean. The variation of OK which can handle random fields with trend is called Universal Kriging (will be described later).
- OK is derived under the assumption that the semivariogram is known. In practice, the semivariogram is unknown and must be estimated, and the estimator $\hat{\gamma}()$ replaces $\gamma()$ in the kriging equations and in the expression for the kriging variance.

  Note: the estimated kriging variance tends to underestimate the prediction error variance of the OK predictor because it does not account for the estimation error incurred in estimating θ.
- Ok is a linear combination of *all* the observations. In practice, often only the observations within a *moving window* or *kriging neighborhood* are used.

- *Environmental monitoring programs.* Note that the kriging variance at any given site $\mathbf{s}_0$ does not depend on the data. thus, it can be used to answer sampling design questions, such as where to take one more observation to maximize the reduction in σ^2_{OK} at a certain point, or where to take one more observation to minimize the maximum (or average) value of σ^2_{OK} over the entire spatial domain.

  A Sampling Design Example will be presented in class.

Bayesian spatial prediction

In the conventional geostatistical approaches for interpolation, i.e. kriging, the covariance structure is estimated first, and then the estimated covariance is used for interpolation. The properties of the interpolants based on an estimated covariance structure are not well understood, and it is common practice to ignore the effect of the uncertainty in the covariance structure on subsequent predictions.

A Bayesian approach to interpolation of spatial processes will provide a general methodology for taking into account the uncertainty about parameters on subsequent predictions.

The prediction problem can be stated in the following form: given observations of a vector field,

$$Z = \{Z(\mathbf{x}_0), Z(\mathbf{x}_2), \ldots, Z(\mathbf{x}_n)\},$$

predict the value $Z(\mathbf{x}_0) = z_0$, for some $\mathbf{x}_0 \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. We can take two approaches to this: an approach based on Lagrange multipliers, and a Bayesian approach.

The Lagrange multiplier approach is the most direct derivation of the kriging estimator and is the one most commonly given in textbooks.

The Bayesian approach leads to the same answers as the standard kriging predictor when the model parameters are known, but it also extends to the case where these parameters are unknown.

The Bayesian approach generalizes automatically to the case in which the variogram parameters are unknown, whereas the classical approach essentially makes the assumption that these are known and only deals with the question of uncertainty of model parameters in a very peripheral way.

This is one major reason for viewing the problem in Bayesian terms, and the close parallels between this and the more traditional approaches.

When the goal is to predict $Z(\mathbf{x}_0) = z_0$, the Bayesian solution is the posterior predictive distribution of $Z(\mathbf{x}_0)$ given the observations Z,

$$\pi(Z(\mathbf{x}_0)|Z) \propto \int \pi(Z(\mathbf{x})|Z,\phi)\,\pi(\phi|Z)d\phi. \quad (5)$$

where $\pi(\phi|Z)$ is the posterior of the model parameters. In the kriging prediction, the model parameters $\phi = (\beta, \alpha, \theta)$ are estimated using a likelihood approach or using empirical methods, and then they are treated as known for the subsequent prediction.

GEOSTATISTICS

1. Exploratory Data Analysis
2. Geostatistical Models.
3. Estimating the Mean Function.
4. Nonparametric Estimation of Variogram.
5. Variogram Model Fitting.
6. Spatial Regression.
7. Kriging.
8. *** Multivariate Geostatistics.**
9. Spatial-Temporal Processes.
10. Nonstationary Processes.

Multivariate Geostatistics

Extensions of Ordinary Kriging

1. We have considered *point kriging*, i.e., prediction at a single point. Sometimes it is desirable to predict the average value over an entire region. This can be accomplished by a straightforward extension of OK called ordinary *block kriging*.

2. We have considered *ordinary kriging*, in which the underlying process is taken to have constant mean. A constant-mean assumption may be untenable. Two extensions which allow for non-constant mean are *universal kriging* and *median polish kriging*.

3. The ordinary and universal kriging predictors are estimates of $E\left(Z(\mathbf{s}_0)|\mathbf{Z}\right)$. In some situations, the quantity $P(Z(\mathbf{s}_0) \geq z_0|\mathbf{Z})$ is of more importance than $E\left(Z(\mathbf{s}_0)|\mathbf{Z}\right)$. This is often true in

environmental monitoring when there are prespecified standards, such as "ozone levels in air cannot exceed 2ppm." A method for predicting such a quantity, is called *indicator kriging.*

4. In some situations, measurements at data locations are taken on more than one variable. An extension which simultaneously predicts these variables is called *co-kriging.*

1. Block Kriging

We want to predict the average value of Z over a **region** B, i.e.,

$$Z(B) \equiv \frac{\int_B Z(\mathbf{s}) d\mathbf{s}}{|B|}$$

where $|B|$ is the area of the block.

The theoretical development proceeds along similar lines as for point kriging and yields ordinary block kriging equations,

$$\boldsymbol{\Gamma}_0 \boldsymbol{\lambda}_{OB} = \boldsymbol{\gamma}_{OB}$$

(same as before), where

$$\boldsymbol{\gamma}_{OB} = [\gamma(B, \mathbf{s_1}), \ldots, \gamma(B, \mathbf{s_n}), 1]'$$

$$\gamma(B, \mathbf{s}_i) = |B|^{-1} \int_B \gamma(\mathbf{u} - \mathbf{s_i}) d\mathbf{u}.$$

The ordinary block kriging predictor of $Z(B)$ is given by

$$\hat{Z}(B) = \sum_{i=1}^{n} \lambda_{B,i} Z(\mathbf{s}_i)$$

where $\lambda_{B,1}, \ldots, \lambda_{B,n}$ are the first n elements of $\boldsymbol{\lambda}_{OB}$.

The kriging variance is given by

$$\boldsymbol{\lambda}_{OB}\boldsymbol{\gamma}_{OB} - |B|^{-2} \int_B \int_B \gamma(\mathbf{u} - \mathbf{v}) d\mathbf{u} d\mathbf{v}.$$

Remarks:

- In practice, it generally will be necessary to evaluate the integrals by a a numerical integration procedure.
- We described kriging a block average from point data ("change-of-support") problem. It is also possible to krige a block average from blocks.

2. Universal Kriging (UK)

Assume

$$Z(\mathbf{s}) = \beta_0 + \beta_1 f_1(\mathbf{s}) + \beta_2 f_2(bfs) + \cdots + \beta_p f_p(\mathbf{s}) + \epsilon(\mathbf{s})$$

where the $f_j(\cdot)$' s are functions of spatial location and $\epsilon(\cdot)$ is intrinsically stationary. We find the BLUP, by minimizing

$$\text{var}[\sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i) - Z(\mathbf{s}_0)]$$

The algebra is "messier" than with ordinary kriging but the end result is similar: the coefficients in the UK predictor are the first n elements of the vector $\boldsymbol{\lambda}_u$ that satisfies the UK equations

$$\boldsymbol{\Gamma}_U \boldsymbol{\lambda}_U = \boldsymbol{\gamma}_U$$

Here,

$$\boldsymbol{\lambda}_U = (\lambda_1, \ldots, \lambda_n, m_0, m_1, \ldots, m_p)'$$

$$\boldsymbol{\gamma}_U =$$

$$[\gamma(\mathbf{s_0} - \mathbf{s_1}), \ldots, \gamma(\mathbf{s_0} - \mathbf{s_{n1}}), 1, f_1(\mathbf{s}_0), \ldots, f_p(\mathbf{s_0})]'$$

And $\mathbf{\Gamma}_U$ is a symmetric
$(n+p+1) \times (n+p+1)$ matrix

$$\mathbf{\Gamma}_U = \begin{cases} \gamma(\mathbf{s}_i - \mathbf{s}_j) & \text{for } i = 1, \ldots, n; j = 1, \ldots, n \\ f_{j-1-n}(\mathbf{s}_i) & \text{for } i = 1, \ldots, n; \text{ and} \\ & \text{for } j = n+1, \ldots, n+p+1. \\ 0 & \text{for } i = n+1, \ldots, n+p+1; \\ & \text{for } j = n+1, \ldots, n+p+1. \end{cases} \tag{6}$$

3. Median Polish Kriging

- First do a median polish fit of overall, row, and column effects and compute the residuals from this fit.
- Next, treat the residuals as the data and perform ordinary kriging to get $\hat{\epsilon}(\mathbf{s})$,say.
- To get the median polish kriging predictor of $Z(\mathbf{s}_0)$, merely add the planar interpolated median polish fit at $\mathbf{s}_0$ to the kriged residual:

$$\hat{Z}(\mathbf{s}_0) = m(\mathbf{s}_0; \hat{a}, \{\hat{r}_k\}, \{\hat{c}_l\}) + \hat{\epsilon}(\mathbf{s}_0)$$

- The kriging variance of the median-polish kriging predictor is taken (with little justification) to be the ordinary kriging variance based on the median polish residuals.

4. Indicator Kriging

Define the indicator random field

$$I(\mathbf{s}, z) = \begin{cases} 1 & \text{if } Z(\mathbf{s}) \leq z \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The indicator random field is the second stationary if the following two conditions hold:

1. $E[I(\mathbf{s}, z)] \equiv F(z)$ for all $\mathbf{s} \in D$ and all $z \in R$.
2. $\frac{1}{2}\text{var}[I(\mathbf{s}, z) - I(\mathbf{s} + \mathbf{h}, z)] \equiv \gamma_{I,z}(\mathbf{h}, z)$ for all $\mathbf{h}$ and all $z \in R$.

Indicator kriging then proceeds as does ordinary kriging, but with $I(\mathbf{s}_i, z)$ in place of Z_i and $\gamma_{I,z}(\cdot)$ in place of $\gamma(\cdot)$. Prediction is often carried out at K levels $z_1, \ldots, z_K$ which requires the K corresponding semivariograms to be estimated and modeled.

5. Cokriging

When sampling over a spatial domain, measurements are often collected on more than one, say m, variables; and we may want to predict one or more values of the variables at an unsampled location $\mathbf{s}_0$.

Suppose that the data are $m \times 1$ vectors $\mathbf{Z}(\mathbf{s}_1), \ldots, \mathbf{Z}(\mathbf{s}_n)$. Denote the jth element of the ith of these vectors by $Z_j(\mathbf{s}_i)$. Let $\mathbf{s}_0$ denote an unsampled site.

Co-kriging versus Multivariate spatial Prediction

- Suppose that we wish to predict, $Z_1(\mathbf{s}_0)$.We could use the OK predictor based on $Z_1(\mathbf{s}_1), Z_1(\mathbf{s}_2), \ldots, Z_1(\mathbf{s}_n)$ of the first variable. If, however, the other variables are **correlated** with the first variable, then a better predictor can result from basing the prediction on all the elements of $\mathbf{Z}(\mathbf{s}_1), \ldots, \mathbf{Z}(\mathbf{s}_n)$. The BLUP of $Z_1(\mathbf{s}_0)$ based on all of these latter quantities is

called the *(ordinary) co-kriging predictor.*

- Alternatively, we may wish to predict the entire vector of variables at an unsampled site, i.e. $\mathbf{Z}(\mathbf{s}_0)$. This can be accomplished using similar ideas, and is called *multivariable spatial prediction.*

Define $\mathbf{Z}_1 = [Z_1(\mathbf{s}_1), Z_1(\mathbf{s}_2), \ldots, Z_1(\mathbf{s}_n)]'$ and $\mathbf{Z}_2 = [Z_2(\mathbf{s}_2), Z_2(\mathbf{s}_2), \ldots, Z_2(\mathbf{s}_n)]'$

Then the **co-kriging predictor** of $Z_1(\mathbf{s}_0)$ is

$$\boldsymbol{\lambda}_1'\mathbf{Z}_1 + \boldsymbol{\lambda}_2'\mathbf{Z}_2$$

The **multivariate spatial predictor** of $\mathbf{Z}(\mathbf{s}_0)$ is:

$$\boldsymbol{\Lambda}_1'\mathbf{Z}_1 + \boldsymbol{\Lambda}_2'\mathbf{Z}_2$$

where $\boldsymbol{\Lambda}$ is a matrix.

5.1 Characterizing spatial cross-dependence

- Cross-covariance function

  $$C_{ij}(\mathbf{s}, \mathbf{t}) = \text{cov}(Z_i(\mathbf{s}), Z_j(\mathbf{t})) \quad (i, j = 1, 2)$$

  Note that:

  - $C_{ij}(\mathbf{s}, \mathbf{t})$ is an ordinary covariance function $(i = 1, 2)$.
  - For $i \neq j$, $C_{ij}(\mathbf{s}, \mathbf{t}) \neq C_{ij}(\mathbf{t}, \mathbf{s})$, in general.
  - For $i \neq j$, $C_{ij}(\mathbf{s}, \mathbf{t}) \neq C_{ji}(\mathbf{s}, \mathbf{t})$, in general.

- "Traditional" cross-variogram (covariance-based cross-variogram)

  $$2\nu_{ij}(\mathbf{s}, \mathbf{t}) = \text{cov}[Z_i(\mathbf{s}) - Z_i(\mathbf{t}), Z_j(\mathbf{s}) - Z_j(\mathbf{t})]$$

  for $(i, j = 1, 2)$

- "Pseudo" cross-variogram (variance-based cross-variogram)

  $$2\gamma_{ij}(\mathbf{s}, \mathbf{t}) = \text{var}[Z_i(\mathbf{s}) - Z_j(\mathbf{t})] \quad (i, j = 1, 2)$$

Pros/cons of the two cross-variograms:

- $2\nu_{ij}$ requires that data on both **variables must be measured at the same location.** $2\gamma_{ij}$ allows the two variables to be measured at completely different locations.
- $2\gamma_{ij}$ require that the two variables **be measured in the same units** in order to be meaningful; $2\nu_{ij}$ has no such requirement.
- In order for co-kriging using $2\nu_{ij}$ to give the **right answers** (i.e. to give the optimal predictor), the symmetry condition $C_{ij}(\mathbf{s}, \mathbf{t}) = C_{ij}(\mathbf{t}, \mathbf{s})$ must be satisfied. Co-kriging based on $2\gamma_{ij}$ does not require this.
- We could transform the variables via $Y_i(\mathbf{s}) = (Z_i(\mathbf{s}) - \bar{Z}_i)/S_i$ where $\bar{Z}_i$ is the sample mean of the ith variable, and S_i is the sample standard deviation. Then, co-krige the Y-variable(s) and transform

back to the original units at the end.

(Handout with co-kriging equations and example)

Linear Model of Coregionalization

In the case of cokriging the problem is more complicated, the "variogram" or covariance is a matrix valued function and as a function it should satisfy a positive definiteness condition (or conditional). The problem is that we do not have a list of standard valid models as we do in the case of one variable. The **linear coregionalization model** is a simple way on ensuring the validity of the model, one need only check the positive definiteness of the coefficient matrices in the linear coregionalization model, i.e., a small number of constant matrices.

The linear model of coregionalization uses the same structures and same ranges for the covariates U and V, and for the cross-variogram, i.e. same model for γ_U, γ_V and γ_{UV}. The nugget and sill vary for the different variables (i.e. γ_U might have different sill than γ_V).

OVERVIEW OF THE GEOSTATISTICAL MODEL

1. Using exploratory techniques, prior knowledge, posit a model of possibly nonstationary mean plus second-order or intrinsically stationary error for the process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ that generated the data. Geostatistical model:

$$Z(\mathbf{s}) = m(\mathbf{s}; \boldsymbol{\beta}) + \epsilon(\mathbf{s})$$

2. Estimate $\boldsymbol{\beta}$ e.g. by ordinary least squares or median polish.

3. Ussing fitted residuals, $\hat{\epsilon}(\mathbf{s})$, estimate $\gamma(\mathbf{h})$ empirically and plot it (in several directions)

4. Select a valid semivariogram model $\gamma(\mathbf{h}; \boldsymbol{\theta})$ that is compatible with the plot from the previous step.

5. Fit the chosen model to the estimated semivariogram to estimate the model's parameters.

6. Using the fitted semivariogram model, re-estimate $\boldsymbol{\beta}$ by generalized least squares (or some other method which accounts for correlation among observations).

7. Repeat Steps 3-6, if needed.

8. "Krige" (i.e. predict) unobserved values at sites (or over regions) of your choosing and estimate the corresponding variances of prediction error.

9. Determine optimal locations to take additional observations, and repeat Steps 1-8, it needed.

FINAL THOUGHTS ON GEOSTATISTICS

A partial list of the guesswork and/or non-optimal procedures involved in the geostatistical method is as follows:

1. The model decomposition

$$Z(\mathbf{s}) = m(\mathbf{s}) + \epsilon(\mathbf{s})$$

   cannot be uniquely identified on the basis of one realization.

2. It may not be reasonable to assume global second-order or intrinsic stationarity, even of just the error process.

3. The nonparametric estimator $\hat{\gamma}(\cdot)$ is unbiased but little else is known about its properties.

4. The most popular method of estimating semivariogram parameters, namely weighted least squares, does not actually use the exactly correct weights, and it ignores correlations.

5. Ordinary kriging, though optimal when the semivariogram is known, is not known to be optimal when the semivariogram is replaced by any particular estimate.

Nevertheless, the geostatistical method has enjoyed great success in the mining industry and in environmental science applications.

1. Kriging is insensitive to the particular model decomposition chosen;
2. provided that only those observations in a neighborhood of $\mathbf{s}_0$ are used, Kriging relies only on local intrinsic stationarity so it is insensitive to large-scale trends in mean and variance which might exist in the data;
3. Kriging is relatively insensitive to misspecification of the semivariogram.