An Overview of Spatial Statistics

Bryan Vinyard
Biometrical Consulting Service
USDA, ARS, Beltsville

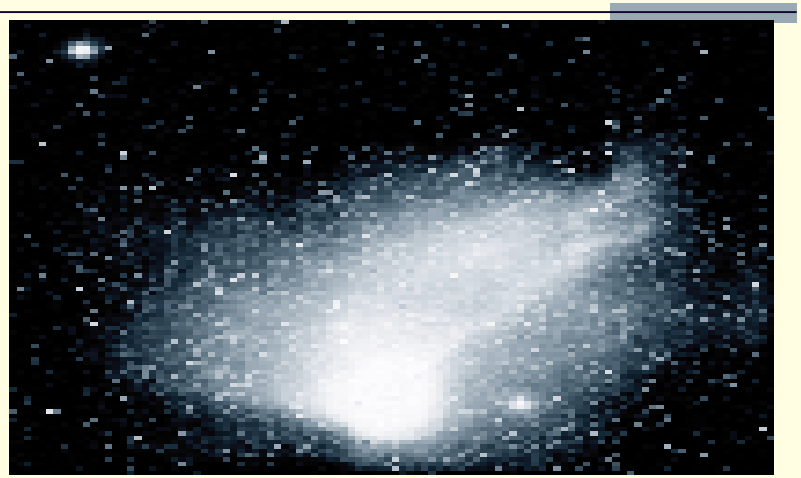
Philosophy of this Workshop

- Focus on Concepts
- Define Terminology
- Illustrate using Graphics
- Avoid excessively Technical Explanations
- Apply the Concepts to Data
- Focus on '*What*?' & '*Why*?' not '*How*?'

"Statistics, the science of uncertainty, attempts to model order in disorder."

Cressie (1991)

Observed Data ('*disorder*' ?)



Characteristics of Interest 'Y'

measured at each observed data point

Examples:

- **Stream Flow**
- **Yield**
- **Muscle Tissue Toughness**
- **CY3, CY5 Image Reflectance**
- **Insect Damage Rating**
- **Bacteria Count**
- **Nitrate Flux**
- **Turbidity**

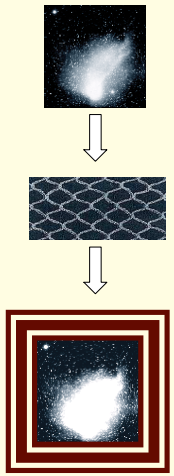
Primary Goal of Applied Statistics

Use observed Y values
together with scientific knowledge

to obtain accurate predictions ($\hat{Y}$)
of unobserved Y values
or to understand a process
(i.e., test the effects of a treatment)

by creating a statistical model:

$\hat{Y}$ =

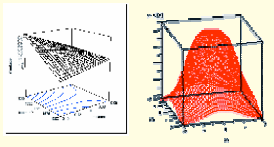


Notes

Fitting a statistical model to data can be viewed as a process of identifying a sequence of "filters" through which the observed data are "sifted".

Initial Attempt to Predict Y
Model the 'Large-Scale' Trend

$\mathbf{Y}$ =



+ ε^*

$\hat{\mathbf{Y}}$

where

$\mathbf{Y}$ is predicted by fitting a 'large-scale' trend to the observed data.

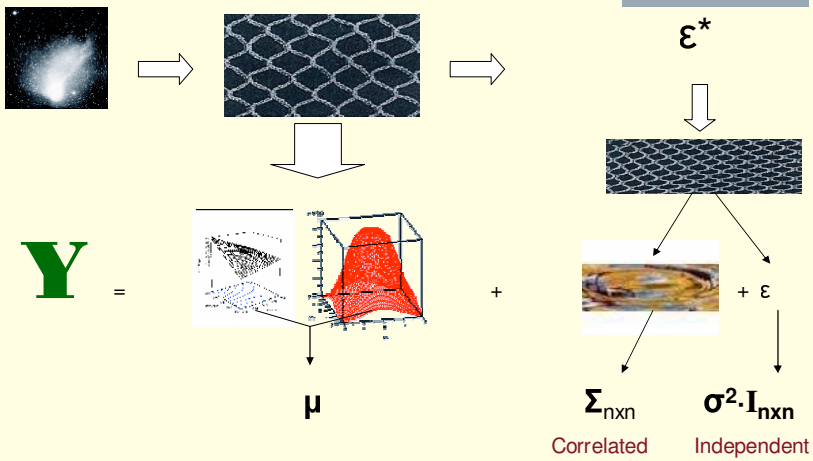
ε^* is data variability remaining after the model is fit.

Notes

When modeling a characteristic of interest, Y, there are typically well established large-scale relationships between Y and one or more fixed-effect “covariates” (i.e., regressors) and/or random “block” effects.

These large-scale effects are modeled first so that the remaining (i.e., residual) variability can be examined in detail on a small-scale to model any spatial dependencies that may be present.

Refine the Model to Predict Y
Model ‘Small-Scale’ Variability

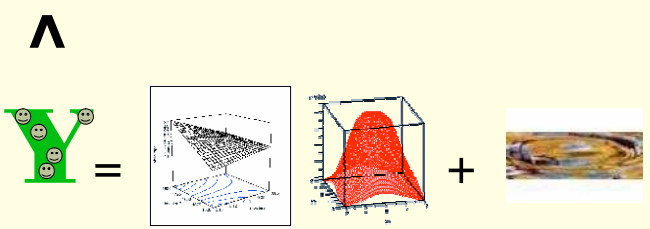


Notes

The initial “filters” capture the large-scale relationships, letting the small-scale relationships remain in the “residual” data to be modeled by a small-scale filter.

Accurate modeling of the small-scale variability, composing the residual data, often requires identification of an appropriate correlation (i.e, covariance) structure.

Refine the Model to Predict Y
Model ‘Small-Scale’ Variability



Notes

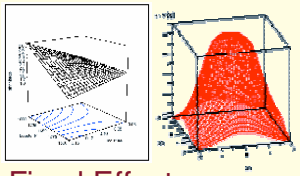
The presence of small-scale dependencies often means that there is correlation among data values located within a certain distance or proximity to one another.

Making use of this "common information" shared by correlated data values improves a model's accuracy.

The primary goal of modeling small-scale dependencies is the identification of an appropriate correlation (i.e., covariance) structure.

Decomposing the Data Variability
ANOVA Terminology

Large-Scale



Fixed Effects

Means
- Deterministic Functions
- Regressors(Covariates)
- Treatments

Small-Scale

All 'residual' variation

(eg., a raindrop on water surface)

Random Effects*

Variance Component
- Variances
- Covariances/Correlations

*Some models may include large-scale random effects (i.e., blocks)

The General Linear Model (GLM)
Perspectives on Model Components

Y	=	**Large-Scale Variation**	+	**Small-Scale Variation**
Y	=	**Fixed Effects**	+	**Random Effects***
Y	=	**Mean** &/or **Covariates**	+	**Variances** & **Covariances***
Y_{nx1}	=	$X_{nxp} \cdot \beta_{px1}$	+	ε_{nx1}
Y_{nx1}	=	μ_{nx1}	+	ε_{nx1}
Y_{nx1}	=	$\hat{y}_{nx1}$	+	ε_{nx1}

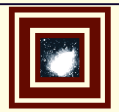
*Some models include random effects (i.e.,blocks) that are considered "large-scale".

Notes

Traditional (General Linear) models typically use only the large-scale effects to model the process; by either predicting Y at various values of regressors or for a collection of experimental "treatments" (i.e., fixed-effects). In traditional GLMs, the small-scale effects do not contribute to improving predictability of Y; rather they are used as precision measures (i.e., root mean-square error) to test hypotheses for the fixed-effects.

Spatial models examine small-scale effects more closely and use the information shared among correlated data values to improve predictability of Y.

The General Linear Model (GLM)
Assumptions – the i.i.d. Mantra


Observed Data - Model Prediction = Model Error

$\mathbf{y}_{nx1}$ - $\hat{\mathbf{y}}_{nx1}$ = $\boldsymbol{\varepsilon}_{nx1}$

Classical GLM assumptions: ε_i are ***i.i.d.***

$\boldsymbol{\varepsilon}_i \sim$ Normal (0, σ^2_ε)

- **i**ndependent → no correlation among the n residual values.
- **i**dentically **d**istributed

Small-Scale Variation
Variances & Covariances that Describe Model Error

For $\boldsymbol{\varepsilon}_i$ ***i.i.d.***

no correlation among the n data values

$\boldsymbol{\Sigma}_{nxn}$ is a diagonal matrix…

$$\boldsymbol{\Sigma}_{nxn} = \sigma^2_\varepsilon \cdot I_{nxn} = \begin{pmatrix} \sigma^2_\varepsilon & 0 & \cdots & 0 \\ 0 & \sigma^2_\varepsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_\varepsilon \end{pmatrix}_{nxn}$$

When the $\boldsymbol{\varepsilon}_i$ are correlated,

correlations appear as non-zero 'covariances' in the off-diagonals of $\boldsymbol{\Sigma}_{nxn}$ …

$$\boldsymbol{\Sigma}_{nxn} = \begin{pmatrix} \sigma^2_\varepsilon & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma^2_\varepsilon & \cdots & \sigma_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma^2_\varepsilon \end{pmatrix}_{nxn}$$

Notes: The Covariance Matrix

To assist in the visualization of how n observed data points are correlated with one another, statisticians use an n x n "covariance" matrix, denoted as $\boldsymbol{\Sigma}_{nxn}$

The element in row i and column j of $\boldsymbol{\Sigma}_{nxn}$ is the "covariance" between data observation i and data observation j. This covariance is typically denoted as σ_{ij}

Correlation, ρ, is defined to be a standardized covariance, $\rho = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$

By definition, when $\sigma_{ij} = 0$, observation i and data observation j are independent (i.e., not correlated, $\rho = 0$).

By definition, a covariance matrix, $\boldsymbol{\Sigma}_{nxn}$, is symmetric about the main (northwest to southeast) diagonal because $\sigma_{ij} = \sigma_{ji}$

Covariances on the main diagonal are more commonly referred to as variances, $\sigma_{ii} = \sigma_i^2$ (for data observation i).

General Spatial Model

Focus is on modeling Small-Scale Variability

when there is dependence or correlation among observed residual values.

Correlation

implies

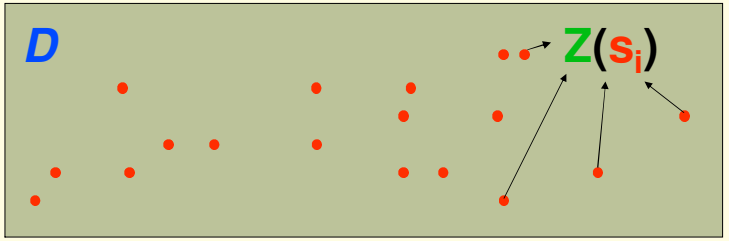
$\boldsymbol{\Sigma}_{nxn}$ is not diagonal

$$\boldsymbol{\Sigma}_{nxn} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{21} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}_{nxn}$$

Notes

The key to successfully modeling the small-scale variance for spatially-correlated data is to accurately identify the relationship between the proximity or distance among data points and their correlation to one another.

General Spatial Model
Terminology



D is the spatial domain or area of interest

s_i notates the spatial coordinates

Z is a characteristic of interest measured or observed at the spatial coordinates

Notes

We make a shift in notation here from that used by the traditional general linear model to that used by spatial models.

Characteristic of Interest:

Y	for traditional general linear models
Z(s)	for spatial models; s denotes the location of the measurement.

Spatial Auto-Correlation
A Definition

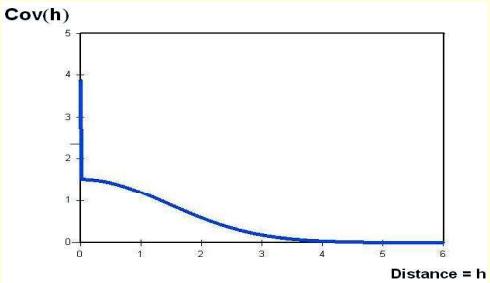
A measure's **correlation with itself** relative to proximity/location.

Data values observed at n locations are **auto-correlated**
when values $Z(s_i)$ and $Z(s_j)$ in close proximity to one another $|s_i - s_j| < h$
are more alike than values located at a further distance $|s_i - s_j| > h$.

As the distance, **h**,

increases between 2 data observations, s_i and s_j,

the correlation between $Z(s_i)$ and $Z(s_j)$ decreases.



Semivariance – A Statistic for Measuring Autocorrelation

Semivariance Formula:

$$\gamma(s_i, s_j) = \tfrac{1}{2}\,\mathbf{Var}[Z(s_i) - Z(s_j)]$$

$$= \tfrac{1}{2}\{\ \mathbf{Var}[Z(s_i)] + \mathbf{Var}[Z(s_j)] - 2\cdot\mathbf{Cov}[Z(s_i), Z(s_j)]\ \}$$

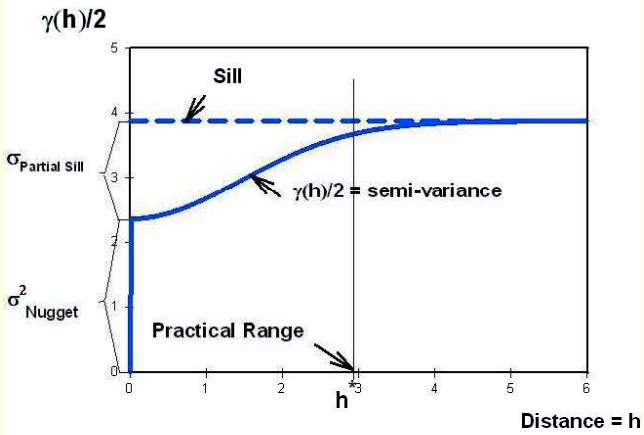
Notes: Semivariance

Semivariance is a statistic and a function that facilitates examining the relationship between the covariance (i.e., correlation) between the characteristic of interest, Z, and the locations where it was measured, s_i and s_j.

The "Auto-Correlation Definition" slide, above, clearly exhibits decreased covariance (and correlation) with increased distance between s_i and s_j.

We will see in a few subsequent slides ("Required Assumptions for Modeling Spatial Data") that under the assumptions of "stationarity", only the distance between observed data points is important to allow accurate estimation of the semivariance (and hence, covariance and correlation).

Semivariogram – A Tool for Measuring Autocorrelation



Notes: Semivariogram

A semivariogram, upon initial consideration, may not be as intuitively interpretable as the covariance plot on the "Spatial Auto-Correlation" slide above.

Most readily, a semivariogram provides the Practical Range, h*, which indicates the distance between any two points in the observed process beyond which those two points are independent of one another (i.e., not correlated).

Also, for any distance ($h=s_i-s_j$), $Cov[Z(s_i), Z(s_j)] = Cov[h] = sill - \gamma(h)/2$.

The component parts of a semivariogram can be interpreted as:

$sill = \sigma^2_{nugget} + \sigma_{partial\ sill} = Var[Z(s_i)]$

where $\sigma_{partial\ sill}$ is the portion of $Var[Z(s_i)]$ due to variation in Z

σ^2_{nugget} is the portion of $Var[Z(s_i)]$ due to measurement error or small-scale variation in the process

Effective Sample Size in Presence of Autocorrelation

"…*positive autocorrelation results in 'loss of information'.*

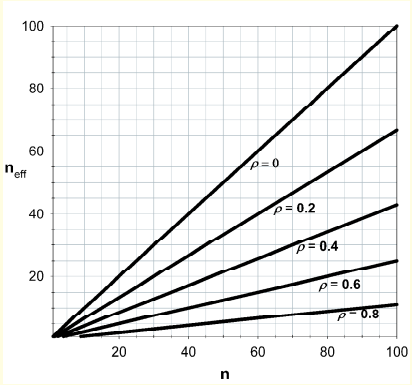
$$n_{effective} = \frac{n_{corr} \cdot (1-\rho)}{(1+\rho)}$$

$n_{effective}$ = uncorrelated (independent) samples

n_{corr} = correlated (dependent) samples

where ρ is autocorrelation

with $0 \le \rho \le 1$.



Notes

The more strongly spatial data are correlated, the less "unique" information is provided by each individual observed data point.

Information shared by data points in closer proximity can improve the ability to accurately model the characteristic of interest, Z.

Simultaneously, strongly correlated data points can reduce the statistical power of inferences (i.e., hypothesis tests).

The effective sample size formula (on the previous slide) results from the assumption (Cressie 1991, p.14-15) that

$$Cov[Z(s_i),Z(s_j)] = \sigma^2 \cdot \rho^{|si-sj|}$$

or equivalently $Cov[h] = \sigma^2 \cdot \rho^h$ where $h=|s_i-s_j|$

Notes

The below table illustrates how correlated data contains less unique information than independent data. For example, if two data points are located at a distance from one another that causes them to have a correlation of ρ=0.2, observing n_{corr} data points provides information equivalent to the amount provided by two-thirds fewer independent (i.e., uncorrelated data points.

ρ	$n_{effective}$
0	n_{corr}
0.2	⅔ n_{corr}
0.5	⅓ n_{corr}
0.8	1/9 n_{corr}
1	0

Autocorrelation Influences Statistical Inference

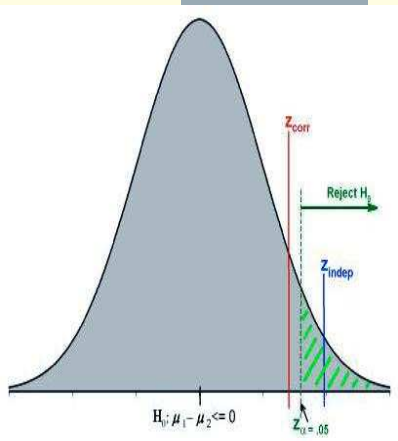
$$Z_{indep} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \cdot \sqrt{\frac{2}{n}}}$$

$$Z_{corr} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \cdot \sqrt{\frac{2 \cdot (1+\rho)}{n \cdot (1-\rho)}}}$$

If positive autocorrelation is present and ignored, a treatment effect can be incorrectly declared significant.

Divisor: n for z_{indep}

$n_{effective}$ for Z_{corr}



Notes

Example:

On the previous slide, a hypothesis test for the equality of 2 treatment means has a divisor of n when the data values for the 2 treatments were independently replicated (i.e., not correlated). In this case, the test statistic $z_{indep} > z_{\alpha=.05}$ and there is sufficient evidence to reject H_0 and declare a significant difference between the treatment means.

However, if the data values observed for the 2 treatments are correlated with one another, the divisor for the test statistic, z_{corr}, is $n_{effective} = \frac{n_{corr} \cdot (1-\rho)}{(1+\rho)}$

which is smaller than the divisor in the independent case.

Hence, $z_{corr} < z_{\alpha=.05}$ so there is an insufficient amount of data (i.e., statistical power) to reject H_0.

Spatial Data

- has **no** independent replications
- consists of a **single** n-dimensional observation:

  $\{ Z(s_1),\ldots, Z(s_n) \}$ at locations $s_1,\ldots,s_n$
- estimates dependency, $\sum$, via semivariance=$\gamma(s_i,s_j)$ using:

  1) the observed $\{ Z(s_1),\ldots, Z(s_n) \}$

  and 2) distances, h, between the $s_1,\ldots,s_n$
- predicts $Z(s_0)$ at an unobserved location, s_0, using the observed $\{ Z(s_1),\ldots, Z(s_n) \}$ and the estimated semivariance=$\gamma(s_i,s_j)$

Required Assumptions for Modeling Spatial Data

Stationary Process

Constant Mean: $Z(s_i)=\mu$ for all s_i in D

Covariance is function of distance ($h=s_i-s_j$), NOT location (s_i):

$Cov(s_i - s_j)$ NOT $Cov(s_i)$

Notes

The validity of all statistical models requires that the data meet some basic assumptions. Typical spatial models require that the data possess characteristics of a "Stationary Process", as defined on the previous slide. If the data do not represent a stationary process, the fitted spatial model will produce incorrect predictions and/or inferences.

Spatial models require that the characteristic of interest, Z, have a constant mean value over the entire domain. This can typically be achieved by modeling the large-scale effects and use the residual variability as the spatial data to which a spatial model is fit.

Required Assumptions for Spatial Data Modeling

The water level of
a calm pond
during
a light rain shower
is an example
of a stationary process:

Photo "Raindrops on the Pond"
by Mark Schretlen 11-May-2003

Notes

The water level on the surface of a pond in a light rain shower is a natural phenomenon that illustrates a stationary process:

1. The average water level is constant over the entire pond surface

2. The water level within a radius from the point where the rain drop strikes the surface depends on the water level at all other locations within that radius. Since the intensity of rain is similar across the entire surface of the pond, the correlation of water levels within the radius is the same regardless of where the rain drop hits the surface of the pond and the strength of correlation within the radius depends only upon the distance from the raindrops point of impact.

Definition: Kriging

Predict unobserved $z(s_0)$ as a weighted average of the observed $z(s_1),\ldots,z(s_n)$ spatially-correlated data

Σ and h (i.e., distance) determine the kriging weights assigned to each of the observed $z(s_1),\ldots,z(s_n)$ in the kriged estimate, $\hat{z}(s_0)$

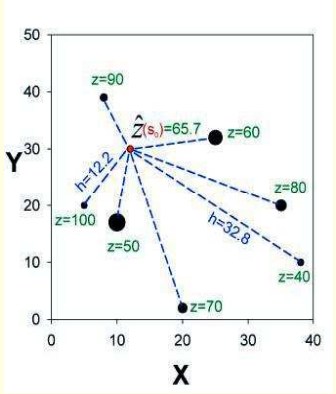
The term Kriging was coined by G. Matheron(1963) in honor of South African mining engineer D.G. Krige, whose work (1951) laid preliminary groundwork for the field of "geostatistics".

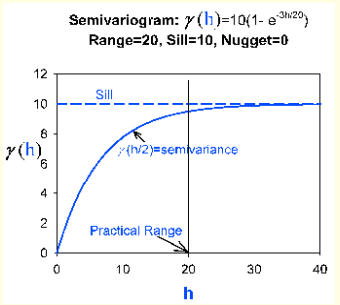
Semivariance determines Kriging Weights

Range=20, Sill=10, Nugget=0

Kriged Estimate, $\hat{z}(s_0)$, at s_0 = (x=12,y=30) is 65.7

$\hat{z}(s_0)$ is a weighted average of the observed $z(s_i)$. The weights sum to 1.
Each point on the graph is sized proportionately to its weight.



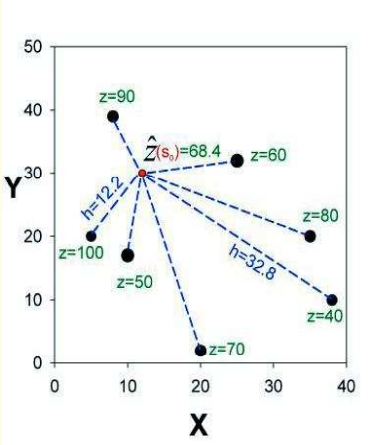


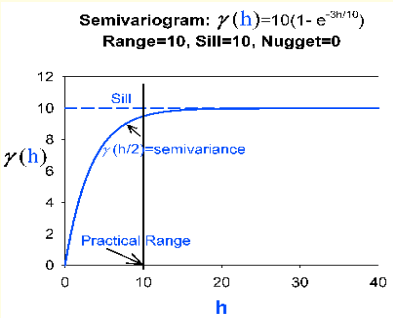
Notes

The previous and next several slides use Isaaks' and Srivastava's (1989, pp. 291, 301-307) small data set of seven observations and one prediction location to examine the effect of semivariogram parameter on ordinary kriging predictions. This example was also given as Example 5.5 in Schabenberger & Gotway (2005).

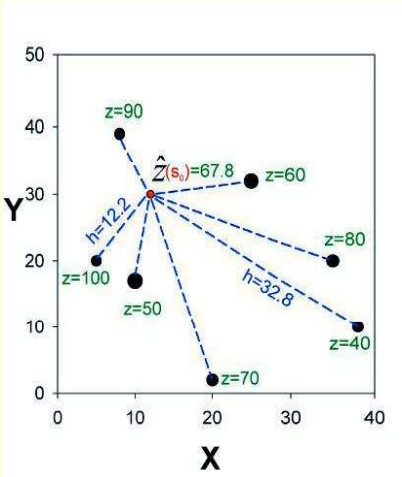
The only difference between the previous and the next semivariogram is the range. The larger practical range in the previous slide causes greater "short-distance" correlations, which results in greater heterogeneity in the weights used to obtain the kriged estimate.

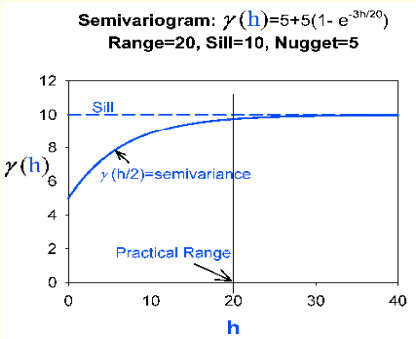
Practical Range Changes from 20 to 10
Kriged Estimate, $\hat{z}(s_0)$, at s_0 = (x=12,y=30)
Changes from 65.7 to 68.4





Nugget changes from 0 to 5
Kriged Estimate, $\hat{z}(s_0)$, at s_0 = (x=12,y=30)
Changes from 65.7 to 67.8

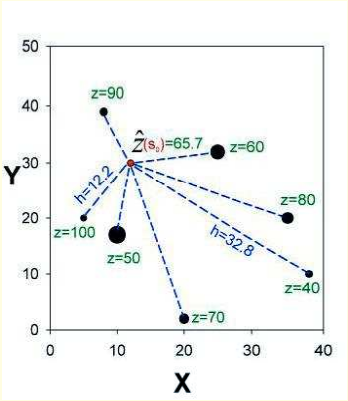


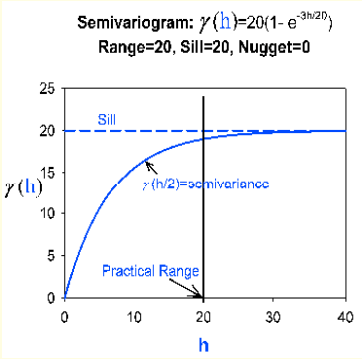


Notes

Introduction of a nugget effect yields more homogeneous kriging weights, similar to the kriging weights resulting from the doubling of the practical range.

Sill Doubles from 10 to 20
Kriged Estimate, $\hat{z}(s_0)$, at s_0 = (x=12,y=30)
Remains Unchanged at 65.7





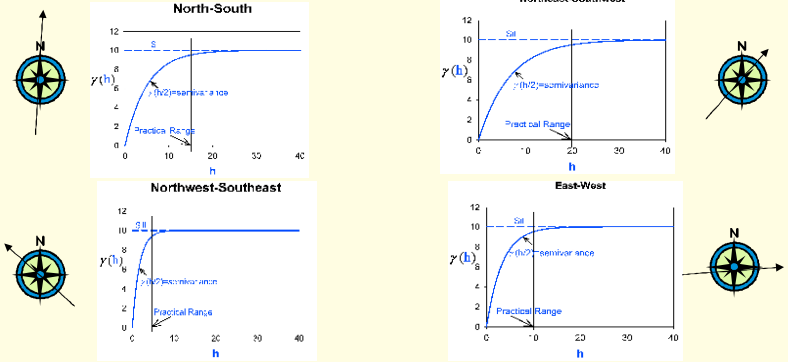
Notes

Compared to the first kriged estimate that used an exponential semivariogram with Range=20, Sill=10, Nugget=0; the previous slide used an exponential semivarigram with Sill=20. Doubling of the sill did not change the kriging weights at all. The larger sill caused only a larger the kriging variance (i.e., variance of Z(s))

Direction in Spatial Modeling

Isotrophy – autocorrelation is equivalent in all directions

Anisotrophy – autocorrelation is direction dependent.



Note

If distance correlations change depending on direction, the appropriate semivariogram for the spatial model also changes with direction. In this case, direction must be considered when fitting the model.

3 Types of Spatial Models

- Geostatistical / Point-referenced
- Lattice / Areal
- Point-Process / Point-Pattern

Notes

The majority of the information presented thusfar most readily lends itself to geostatistical data. However, the general concepts apply (with appropriate adjustments or modifications) to all 3 types of spatial models.

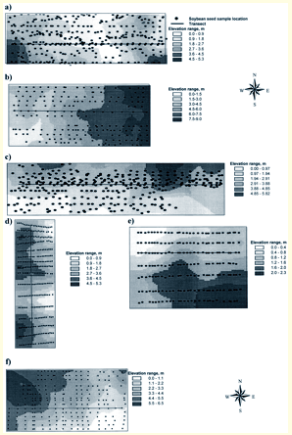
Geostatistical / Point-referenced Models

Specific locations $s_1,\ldots,s_n$ in the domain **D** are selected.

The characteristic of interest, $z(s_1),\ldots,z(s_n)$, is observed.

Example: Six fields, each planted in a different soybean cultivar.
- Locations $s_1,\ldots,s_n$ are n individual soybean plants.
- $z(s_1),\ldots,z(s_n)$ are protein concentration of the plant's yield.

Crop Science 42:804-815 (2002), A. N. Kravchenko and D. G. Bullock



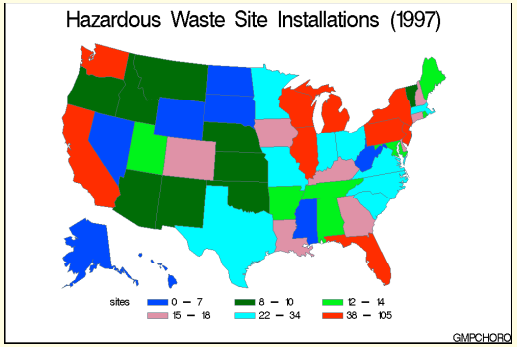
Notes

The figures on the right of the previous slide illustrate how kriging can produce prediction maps.

Lattice / Areal Models

Specific locations $s_1,\ldots,s_n$ represent 'contiguous areas' in the domain **D**. The characteristic of interest, $z(s_1),\ldots,z(s_n)$, is observed for each 'area'.

Example: # of hazardous waste sites in each U.S. state.



SAS/Graph Online Documentation – Proc GMAP

Notes

Lattice or Areal models have the objective of predicting Z(s) where s is an "area" rather than a "point", as in the Geostatistical/point-referenced model case.

Two Methods of Modeling Lattice Data

- Simultaneously Autoregessive
  - Likelihood methodology
- Conditionally Autoregressive
  - Gibbs sampling (Bayesian) methodology

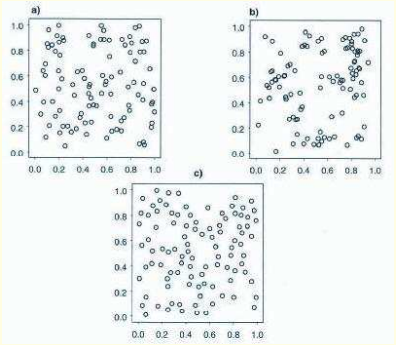
Schabenberger & Gotway (2005) pg.7

Point-Pattern Models

Objective:

Model the 'process' that generated the spatial data.

Fig a) completely random pattern
Fig b) Poisson cluster process
Fig c) process with sequential inhibition regularity

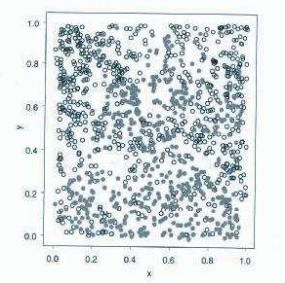


Schabenberger & Gotway (2005), p. 82

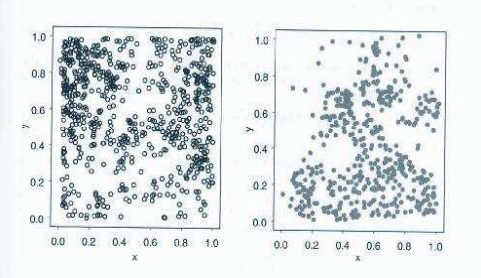
Point-Pattern Data
Example 1: A Marked Process

Distribution of hickory(○) and maple trees(●).

Overlaid



Separate Plots



Schabenberger & Gotway (2005), p.119,121

Notes

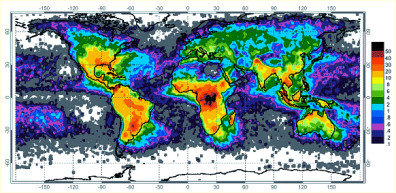
The "mark" in this marked process is whether the species of tree is hickory or maple.

Point-Pattern Data
Example 2: Lightning Strikes

Lightning strikes within 200 miles of the U.S. east coast April 17-20, 2003.

Schabenberger & Gotway (2005) p.13

Kriged predictions can also be obtained for point-pattern data, as shown by the NASA map of global lightning strikes.



Summary of the 3 Model Types

Analogy: A desktop is the domain **D** of locations s_i
Experiment → pour sand on the desktop.

Geostatistical & Lattice Data:

- locations s_i do not change from one pouring (i.e., experiment) to the next
- $z(s_i)$= observed sand depth varies at s_i

Point-Pattern:

- specify a sand depth of interest
- observe all locations s_i in **D** where sand has this depth.

Schabenberger & Gotway (2005)

Data Measured at Multiple Scales

"Even when the disorder is discovered to have a perfectly rational explanation at one scale, there is very often a smaller scale where the data do not fit the theory *exactly*, and the need arises to investigate the new, residual uncertainty."

Cressie (1991)

Hierarchical Models

- Estimate 'parameters' of an experiment using the observed data $z(s_1),\ldots,z(s_n)$
- Assume and impose statistical distributions on the parameters to be estimated
  - distribution choices rely on theory and/or scientific knowledge
  - modeling of distributions uses Bayesian methods
    - GEOBUGS freeware

References

- Schabenberger & Gotway (2005)
  Statistical Methods for Spatial Data Analysis
- Banerjee, Carlin, & Gelfand (2004)
  Hierarchical Modeling and Analysis for Spatial Data
- Cressie (1991)
  Statistics for Spatial Data