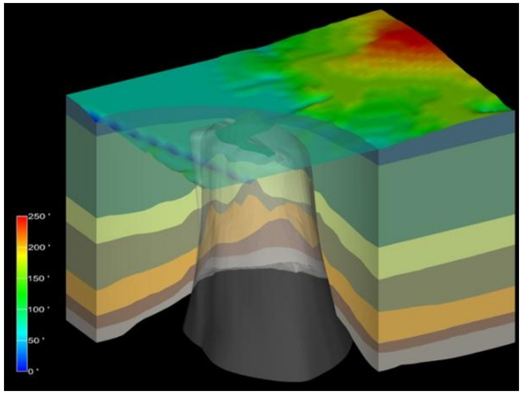
SAND2013-xxxx

SAND2013-4769C

Exceptional service in the national interest



Introduction to Geostatistics for Site Characterization and Safety Assessment

IAEA Training Course – July 1-5, 2013

Bill Arnold

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Acknowledgements

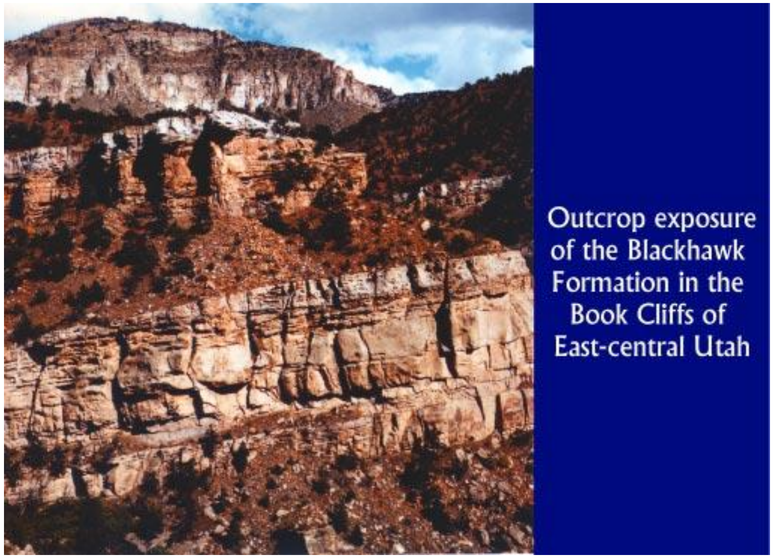
Coarse materials and exercises were developed in collaboration with Sean McKenna, formerly with Sandia National Laboratories. Sean McKenna is currently with IBM Research in the Smarter Cities Technology Centre – Dublin, Ireland

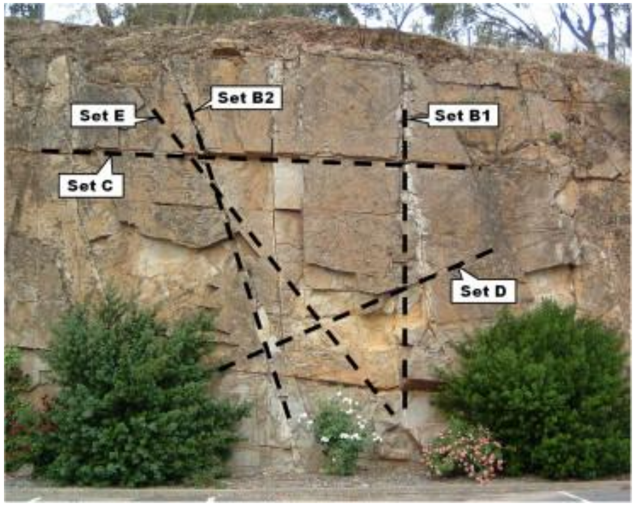
Topics

- **Ubiquitous nature of geological heterogeneity**
- **Relationship of geostatistics to classical statistics**
- **Estimation methods versus simulation methods in geostatistics**
- **Relation between spatial variability and uncertainty**
- **Uses of geostatistics in site characterization**
- **Uses of geostatistics in risk assessment**

Background

Geologic materials are ubiquitously heterogeneous


Outcrop exposure of the Blackhawk Formation in the Book Cliffs of East-central Utah



Background

Limited sampling of subsurface results in uncertainty in spatial distribution of hydraulic and physical parameters

Background

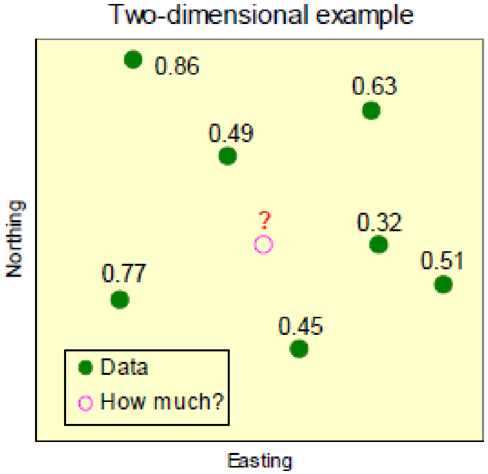
- Many geological media are produced by processes that are very complex and occur on scales ranging from microscopic to 100s of km
- Sampling of subsurface media is extremely limited relative to the volume of material of interest, because of economic and practical limitations
- The challenge for geoscientists and engineers is the characterize and model geological media adequately for predictive analysis and decision making
- Characterization and prediction is necessarily uncertain because exact description of the system is impossible
- The goal is to make full use of available geological information and to understand the limitations of our knowledge and the related uncertainty

Background

- Fortunately, geological media have spatial structure that can be characterized with subjective and quantitative knowledge
- Geological knowledge and interpretation are important sources of information on spatial structure and continuity, particularly for large-scale features
- Material properties also tend to have spatial correlation related to the continuity of processes that operated during formation of the geological media
- Intuitively, the values for a particular property tend to be more similar for locations that are closer together than for locations that are more widely spaced
- This characteristic of geological media forms the basis for the field of geostatistics

Geostatistics - Overview

- Primary objective of geostatistics is the characterization of spatial or temporal systems that are incompletely known
- Classical univariate statistics considers only the population of values for a particular variable
- Geostatistics is an extension of bivariate statistics that uses the sampling location (in space or time) of every measurement
- Geostatistical analysis is only meaningful if the measurements show some spatial (or temporal) correlation

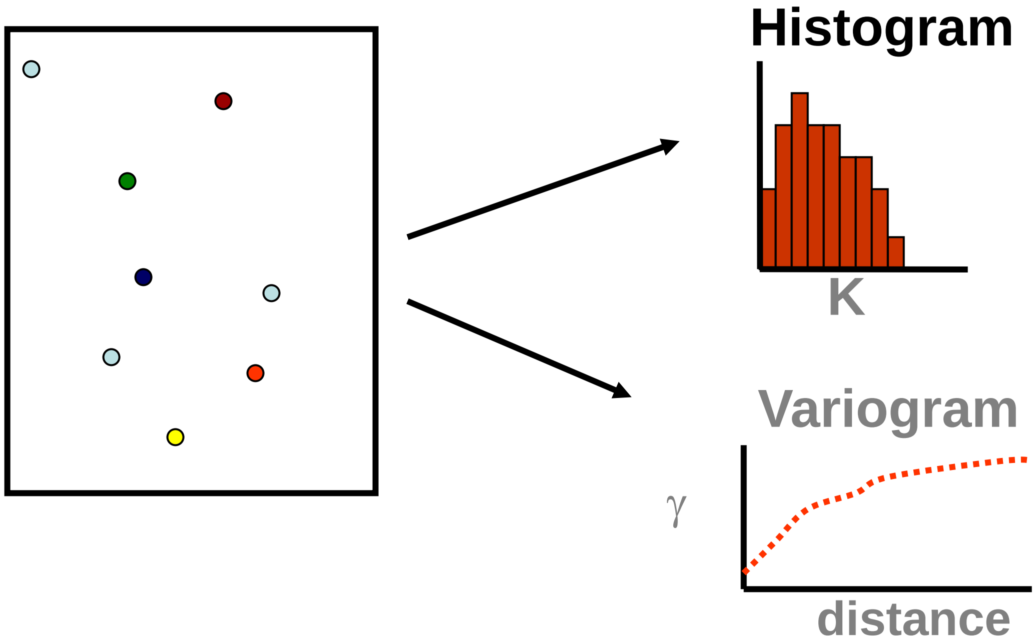


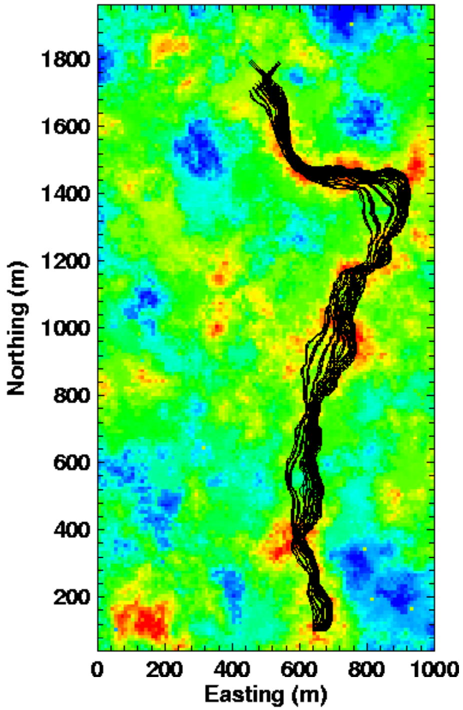
Geostatistics - Background

- Geostatistics is a very broad field; this workshop provides only a brief introduction to the topic
- Geostatistics has developed over a long time frame, starting with theoretical developments in the 1950s and expanded significantly in the era of digital computation
- Original applications included estimation methods applied to calculation of ore reserves in mining
- More recent applications have been more focused on simulation methods, as applied to reservoir modeling in petroleum engineering applications
- Geostatistical estimation methods are generally used for interpolation of measurement, but simulation methods can be used in extrapolation of parameters (with caution!)

Geostatistics - Overview

Geostatistics provides a set of tools for modeling spatial distributions of parameters based on the available data and the two-point spatial covariance



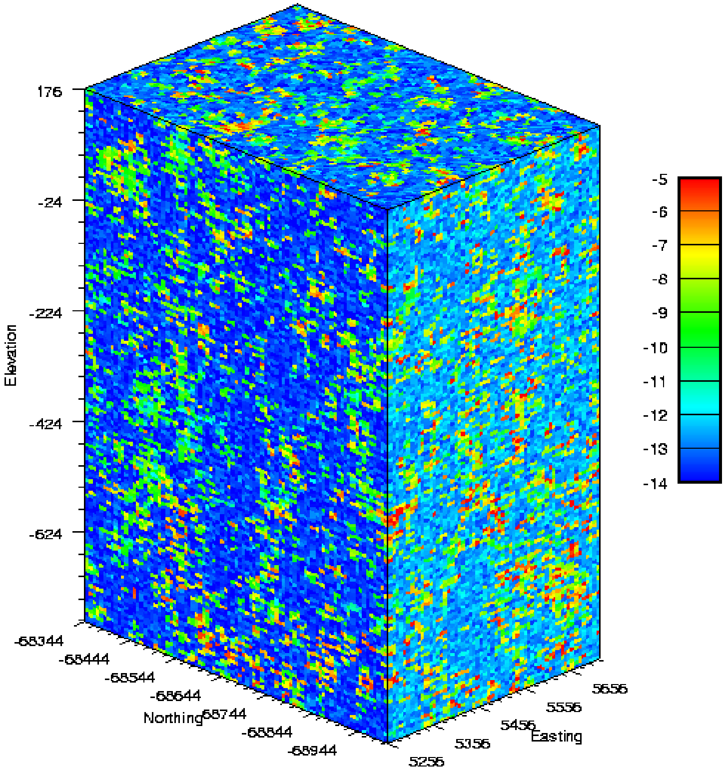


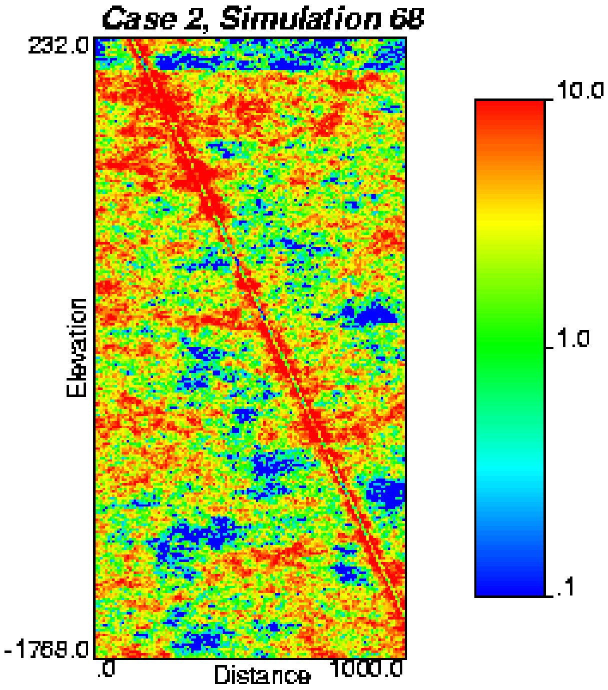
Geostatistics

- The study of spatially and/or temporally correlated data
- Suite of tools for quantifying the amount and style of spatial correlation
- Adaptations to classical regression techniques to take advantage of spatial correlation
- Includes both interpolation (estimation) techniques and Monte-Carlo simulation techniques

Geostatistics Applications

Three-dimensional model of fracture permeability at the JNC MIU site, Japan





Cross-section model of fracture frequency at the JNC MIU site

Geostatistics – Problem Statement

Site Characterization and Performance Assessment Concerns:

- **How can we model spatial variability in hydraulic and physical properties?**
- **How does uncertainty in spatial distribution of properties affect uncertainty in performance assessment results?**
- **If more site characterization boreholes are worthwhile, where are the best locations for them?**

  $\Rightarrow$ Optimal Site Characterization

Geostatistics – Approach

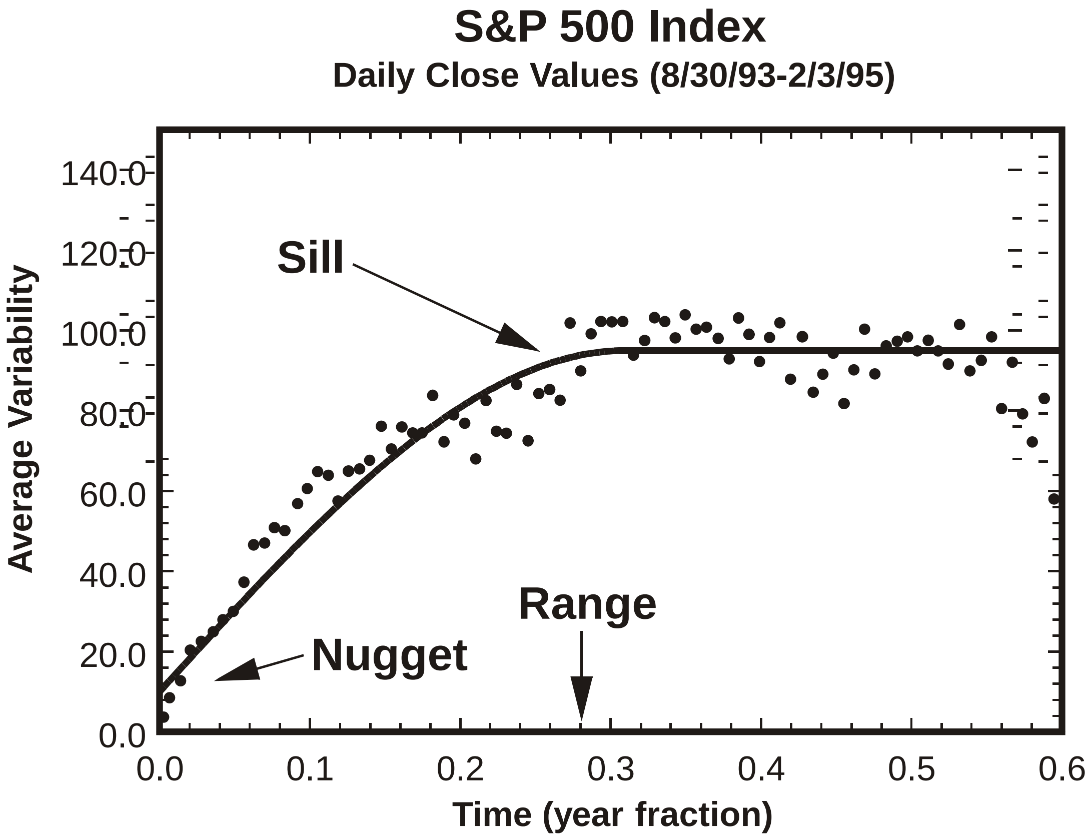
Once the site characterization goals and PA performance measures have been established, Geostatistical Simulation is used to:

- **Accurately estimate the values of properties (K, T, porosity, fracture frequency, thermal conductivity) at unsampled locations**
- **Provide a measure of the uncertainty in that estimate**
- **Create equiprobable maps of the spatial distribution of properties with associated uncertainty levels**
- **Provide sampling locations that will contribute the most to reducing uncertainty**

Geostatistics – Example Variogram

The **sill** is the gamma value corresponding to the total variability in the dataset

The **nugget** accounts for variability at zero lag distance.

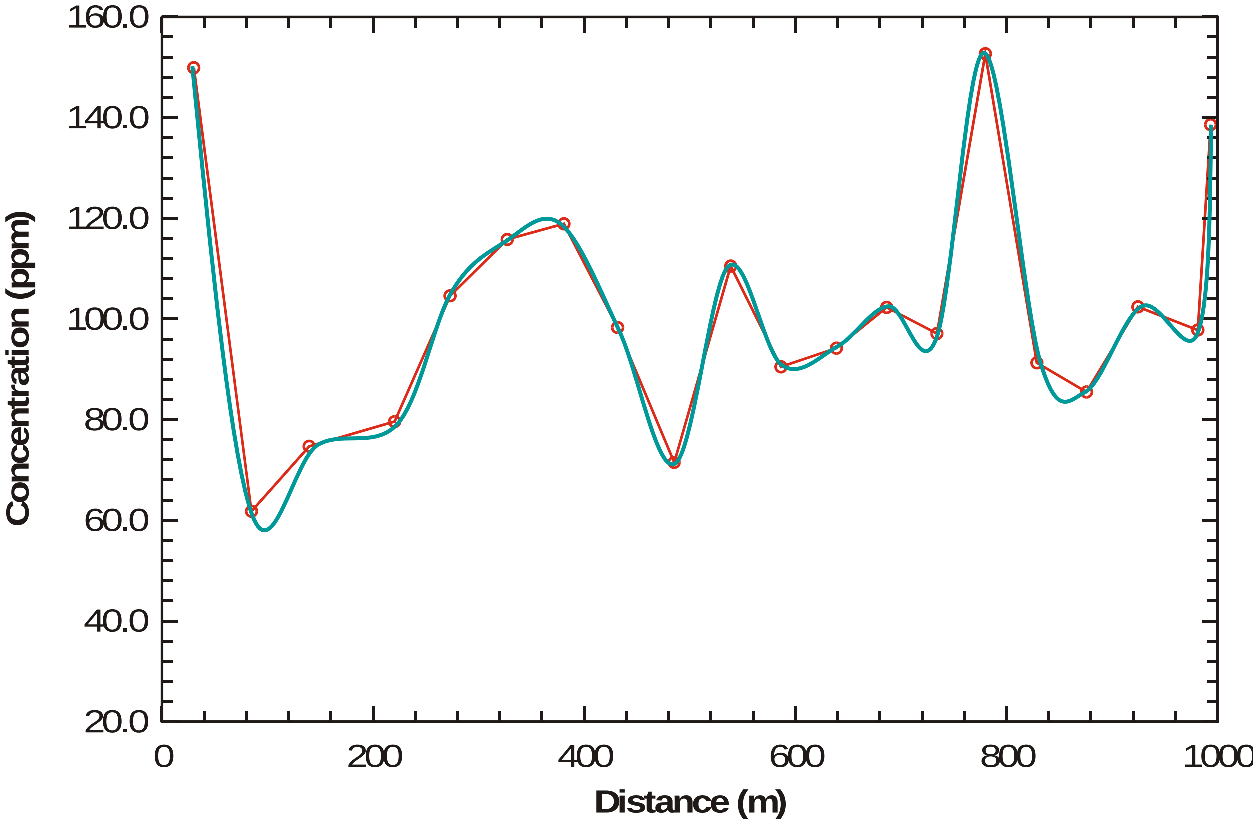


The **range** is the separation distance (time) at which the sill is reached

Geostatistics - Estimation

- **Estimation is an interpolation technique**
- **Weighted combinations of the surrounding data are used to determine an estimate of the value at an unsampled location**
- **Kriging is a geostatistical estimation procedure that uses the information in the variogram to determine the weights used in estimating unsampled locations.**
- **Estimation procedures determine the "best-guess value" at any location**

Geostatistics - Estimation

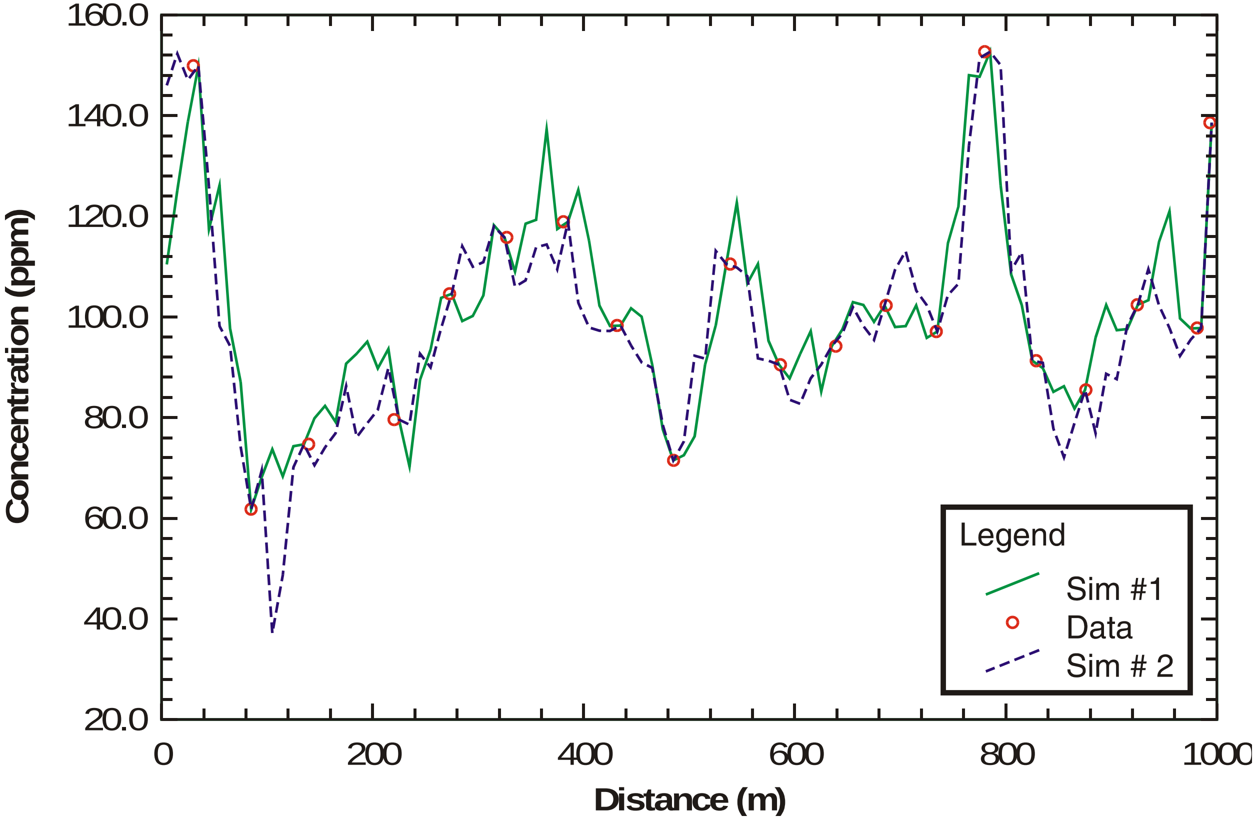


Geostatistics - Simulation

Geostatistical simulation is a Monte-Carlo technique for producing multiple, equally probable, realizations of a sampled variable.

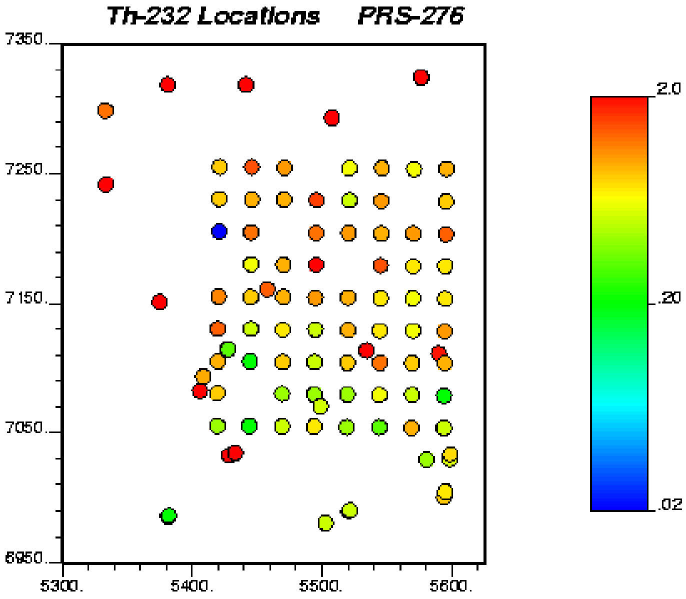
- **Simulation reproduces the variability of the initial data and does not have the smoothing effect of estimation**
- **Simulation reproduces the actual values, histogram, and variogram of the input data**
- **Each realization is a plausible model of the reality from which the samples were obtained**

Geostatistics - Simulation

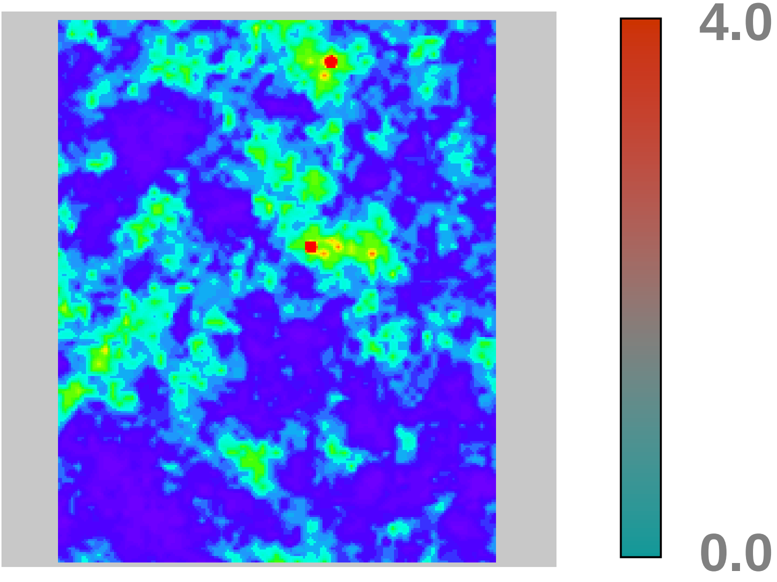


Geostatistics - Simulation Example

Sample Data showing location and Activity of Th-232

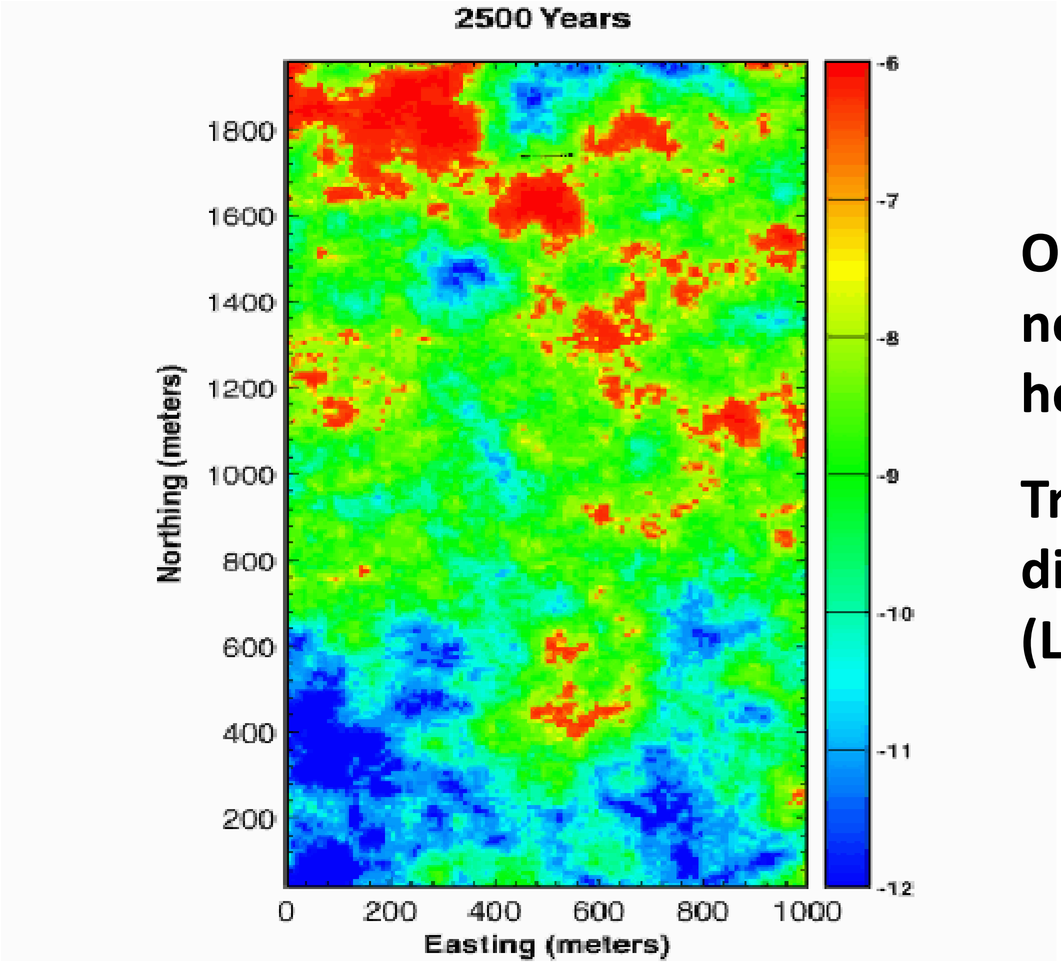


20 Realizations of Th-232 Activity Levels



Example of soil contamination from U.S. DOE Mound Site

Geostatistics – Groundwater Flow

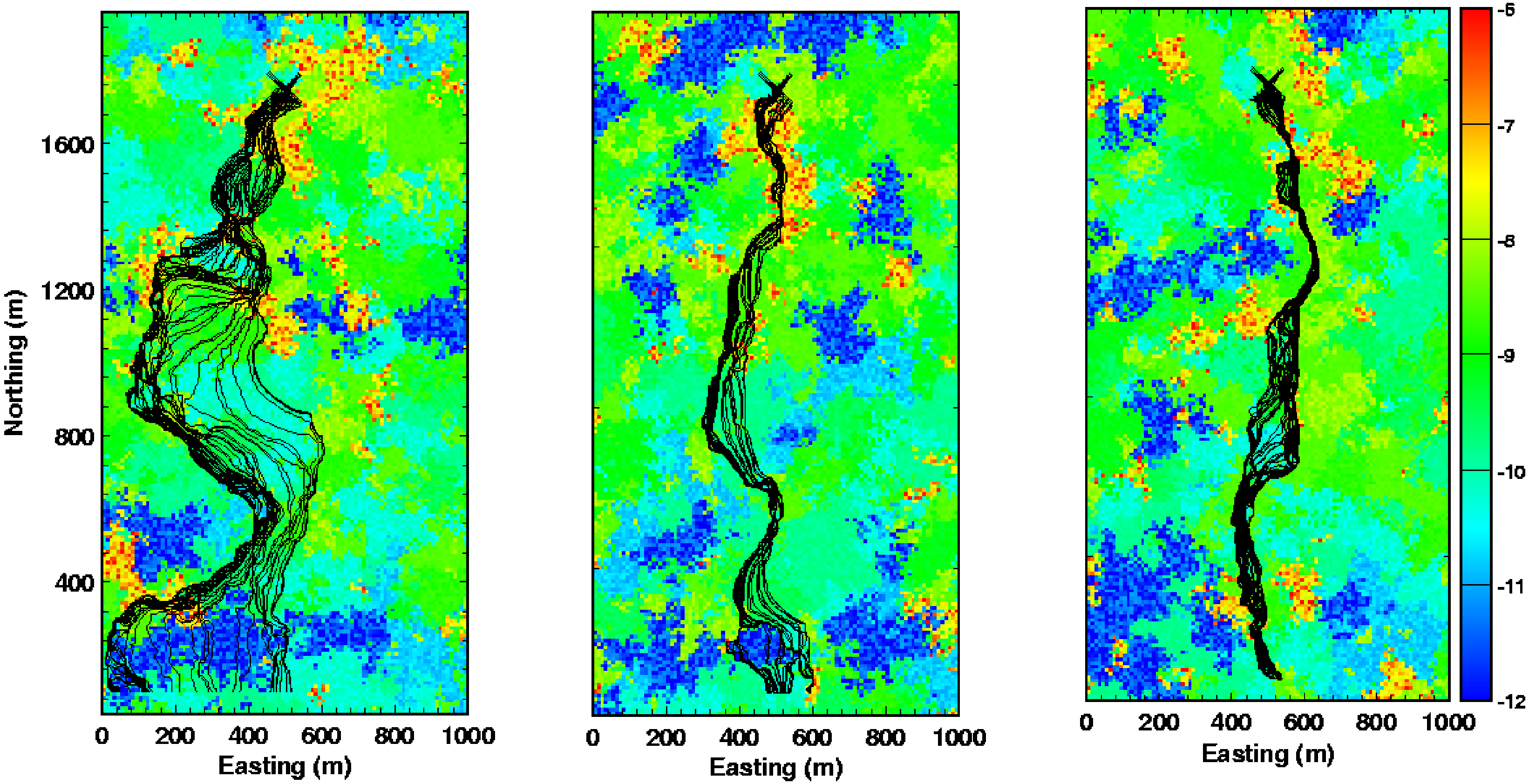


One realization of non-uniform flow through heterogeneous material

Transport is convective, not dispersive (Large Peclet numbers)

Geostatistics – Groundwater Flow

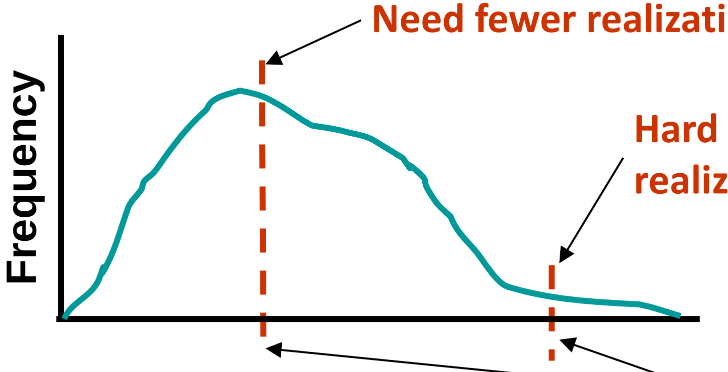
Three Realizations conditioned to same 96 boreholes



Geostatistics – Realizations

How many realizations are enough to make an informed decision?

- **The answer is dependent on the performance metric**
  (performance metric relative to uncertainty distribution)



Need fewer realizations to get accurate picture

Hard to define the tail of the distribution, need more realizations to get an accurate picture

Expected dose

Alternative Performance Measures

- **Is an idea of the best, worst, and expected conditions enough?**
- **Test how many is "enough" with concept of a representative elementary volume (REV) from groundwater hydrology**

Geostatistics – Additional Sampling

Where to locate additional boreholes? (3 Approaches)

Traditional Technique

- Reduce estimation error or kriging variance by putting boreholes in unsampled locations

Decision Based Technique

- Areas of maximum uncertainty defined by probability mapping

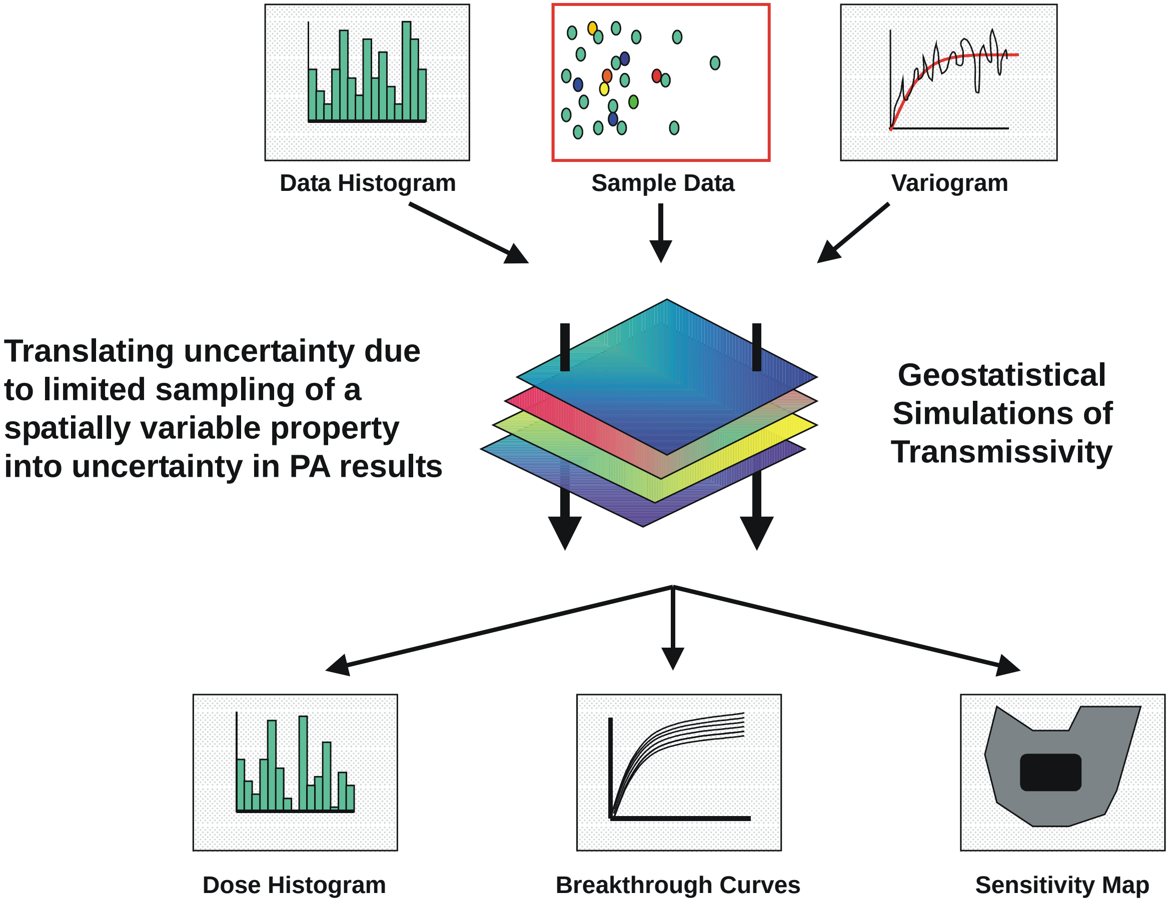
New Idea

- Consider K to be a stochastic input parameter to transport model and use sensitivity analysis

Geostatistics – Summary

- Uncertainty is due to limited sampling of a spatially heterogeneous variable
- Spatial uncertainty creates uncertainty in performance assessment results
- Geostatistical simulation provides a technique for examining and quantifying the amount uncertainty.
- Estimates of uncertainty can be propagated through performance assessment models
- Relationship between PA uncertainty and spatial uncertainty can be used to guide site characterization

Geostatistics



Data Histogram

Sample Data

Variogram

Translating uncertainty due to limited sampling of a spatially variable property into uncertainty in PA results

Geostatistical Simulations of Transmissivity

Dose Histogram

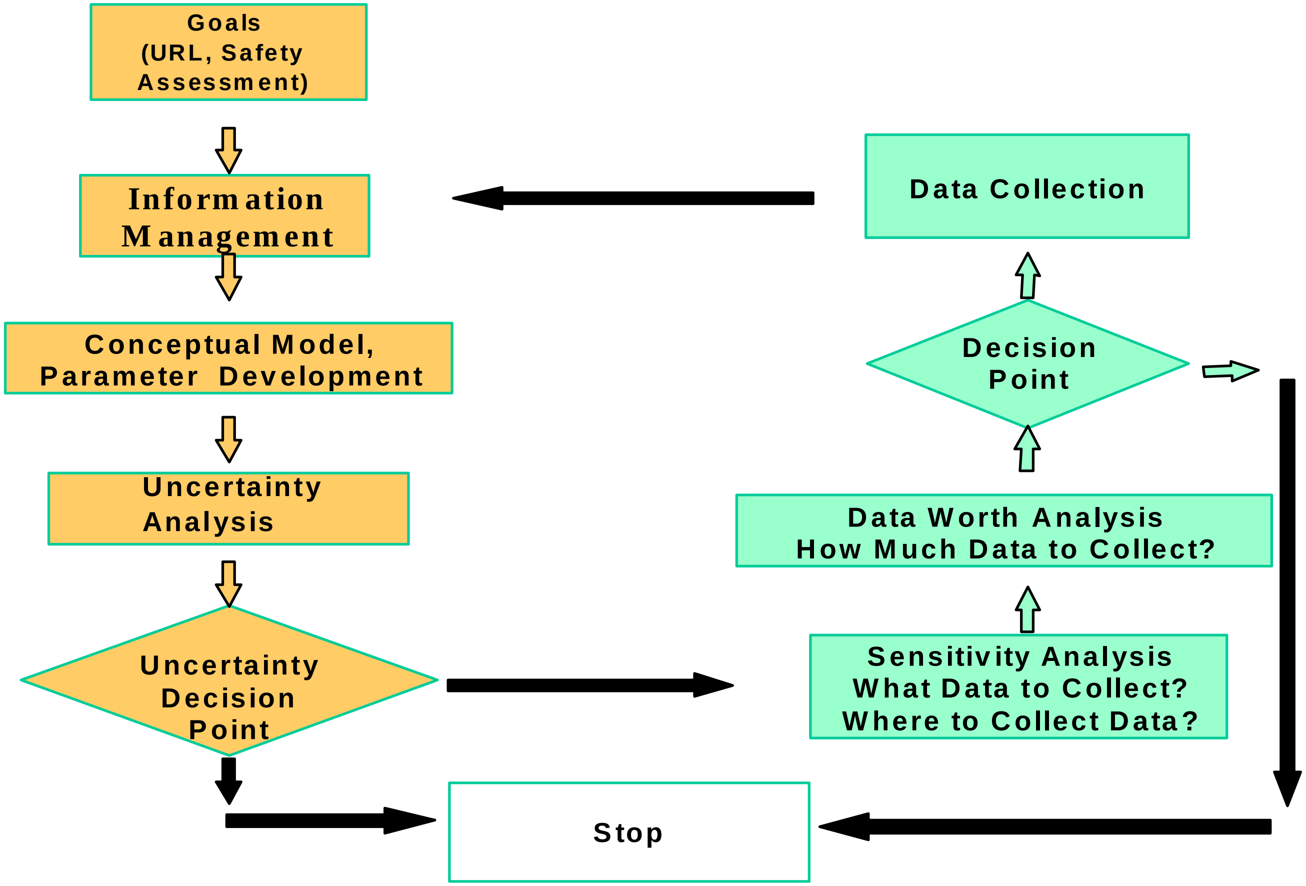
Breakthrough Curves

Sensitivity Map

Geostatistics – Exercises Objectives

- **Become familiar with SGeMS and GSLIB software**
- **Learn the basics of exploratory data analysis**
- **Learn the basics of spatial correlation analysis and variogram analysis**
- **Learn the basics of spatial estimation methods**
- **Learn the basics of spatial simulation methods**
- **Apply these analytical techniques and methods to example 2D and 3D data sets**

Geostatistics – PA Decision Framework



Goals
(URL, Safety Assessment)

Information Management

Conceptual Model, Parameter Development

Uncertainty Analysis

Uncertainty Decision Point

Sensitivity Analysis
What Data to Collect?
Where to Collect Data?

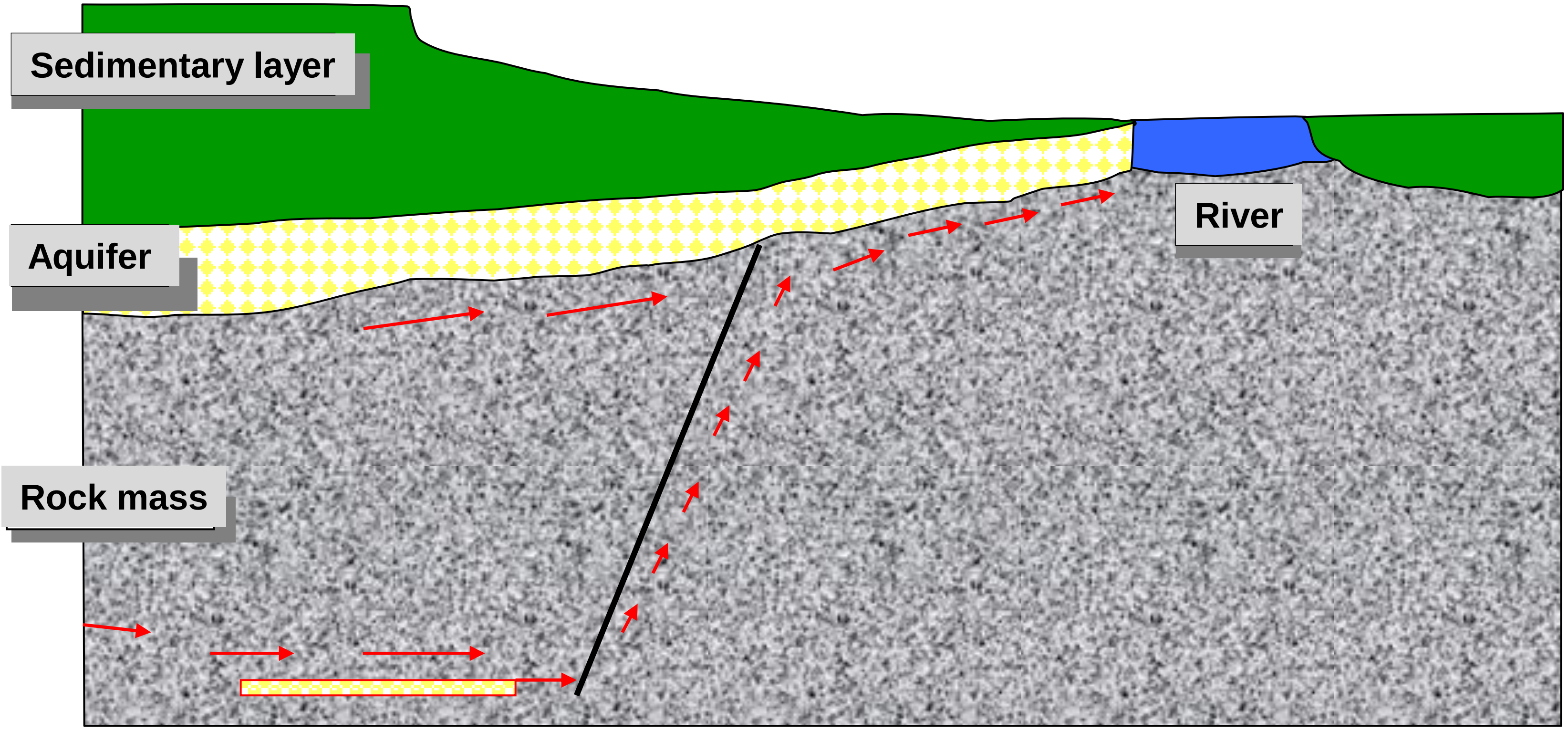
Data Worth Analysis
How Much Data to Collect?

Decision Point

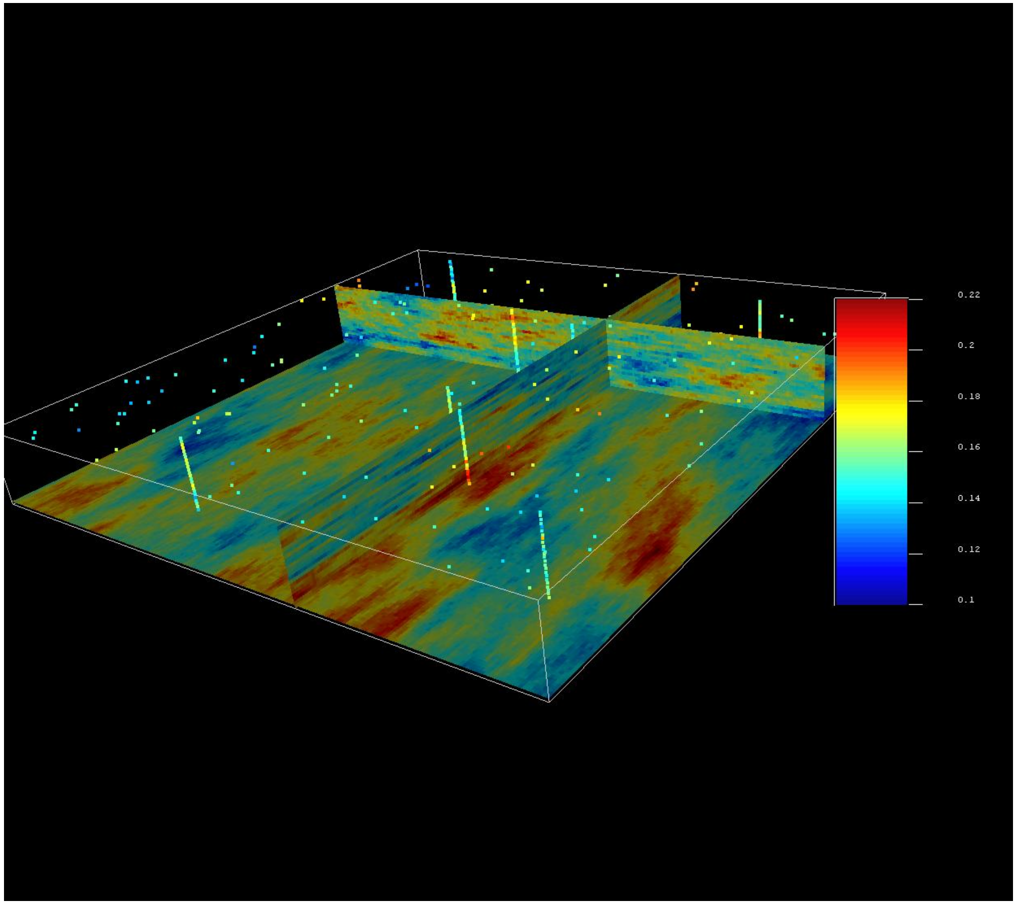
Data Collection

Stop

Geostatistics – PA Decision Framework



Geostatistics – Exercise Preview

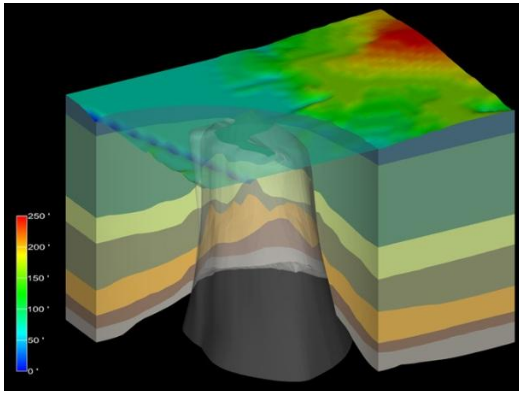


- **Exploratory Data Analysis of these data**
- **Analyze spatial correlation of parameters by constructing and modeling variograms**
- **Create estimate of porosity and permeability fields**
- **Create multiple realizations of porosity and permeability fields**

SAND2013-xxxx

Exceptional service in the national interest

Sandia National Laboratories



Spatial Variability: Exploratory Data Analysis and Spatial Correlation Analysis

IAEA Training Course – July 1-5, 2013

Bill Arnold

U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration
IAEA International Atomic Energy Agency

Outline

- Introduction
- Exploratory data analysis
    - Visualization
    - Univariate statistics
    - Data correlations
    - Clustering, transformations, and trends
- Spatial correlation analysis
    - Experimental variograms
    - Correlation anisotropy
    - Variogram models

Introduction

- Spatial correlation and heterogeneity are important characteristics of geological system with potentially important consequences for repository performance
- Objectives of spatial correlation analysis include:
  - Data checking
  - Conceptual understanding of heterogeneity
  - Univariate statistical characterization
  - Understand complicating factors in spatial correlation (e.g., data transformation, anisotropy, trends
  - Quantitative analysis of spatial correlation
  - Modeling of spatial correlation

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is everything you do to understand your data. It includes both objective and subjective analyses.

Three essential functions of EDA:

- **error checking**
- **understanding physical processes for use in modeling**
- **statistical validation of results**

Topics

- **mapping the data**
- **histogram techniques**
- **probability-plotting techniques**
- **correlations among multivariate data**
- **data transformations**

EDA - Mapping the Data

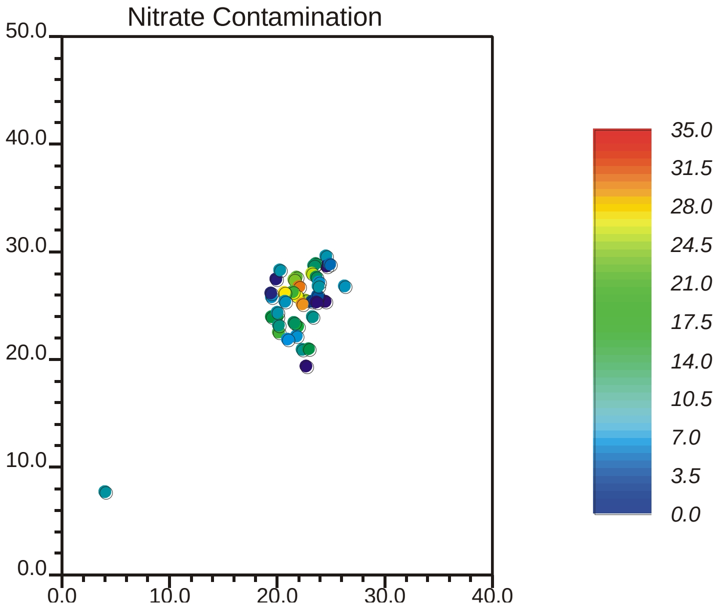
Plot it! Humans are very good at processing visual data.

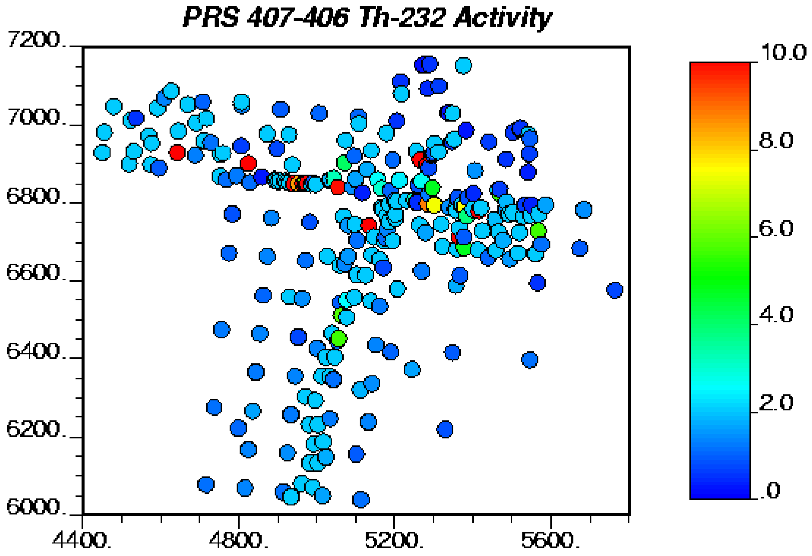
- **Look for spatial patterns, correlations with other variables**
- **Spot problems in data: keypunch errors,**
  transposed coordinates
- **Beware of pitfalls: data clustering, preferential sampling**
- **Use:**

  - Plots with colors or symbols proportional to value
  - Indicator plots
  - Quick contouring with a simple program

EDA - Data Posting

Data posting shows locations and values in single image



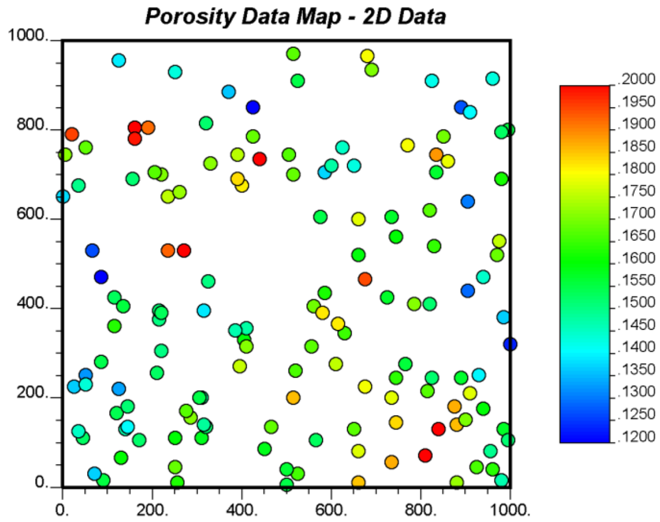


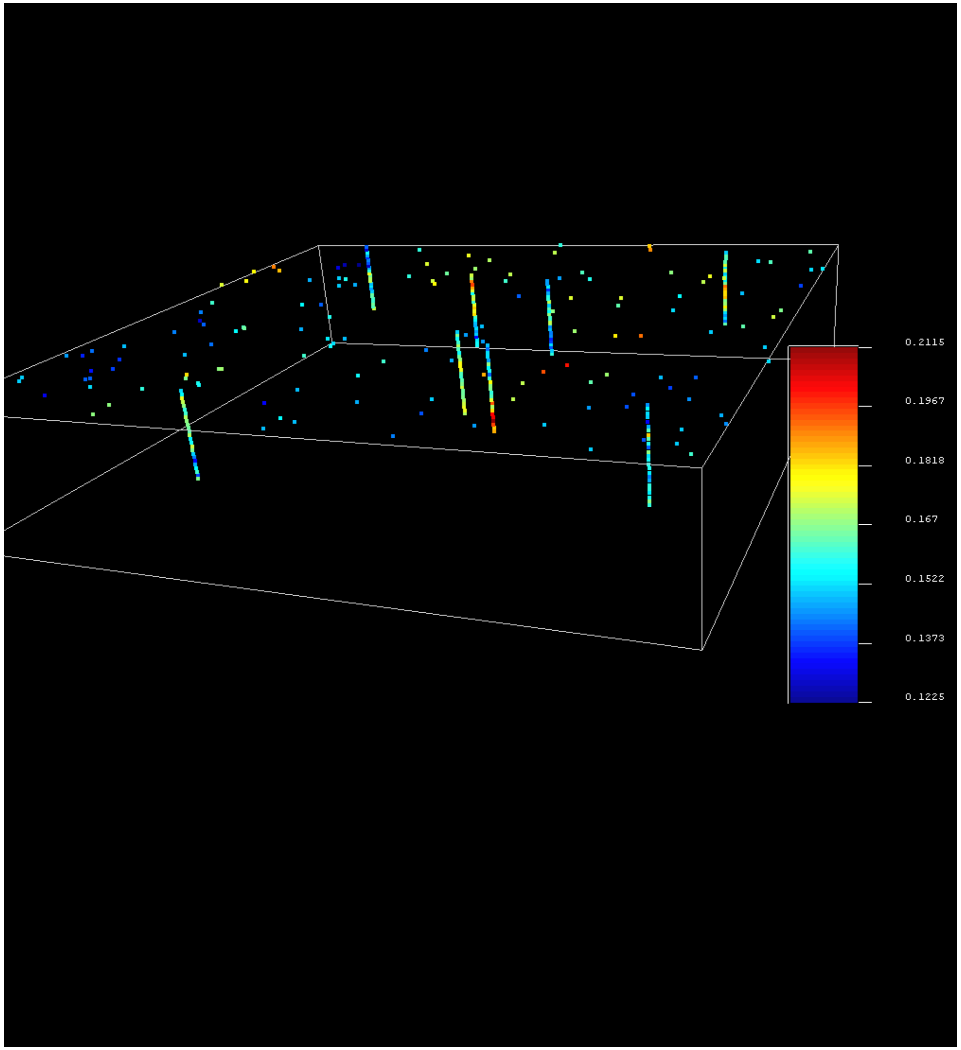
Coordinates of lower left data point may contain an error

Data posting tool should be used more often

Data Posting

Data posting examples from workshop exercises



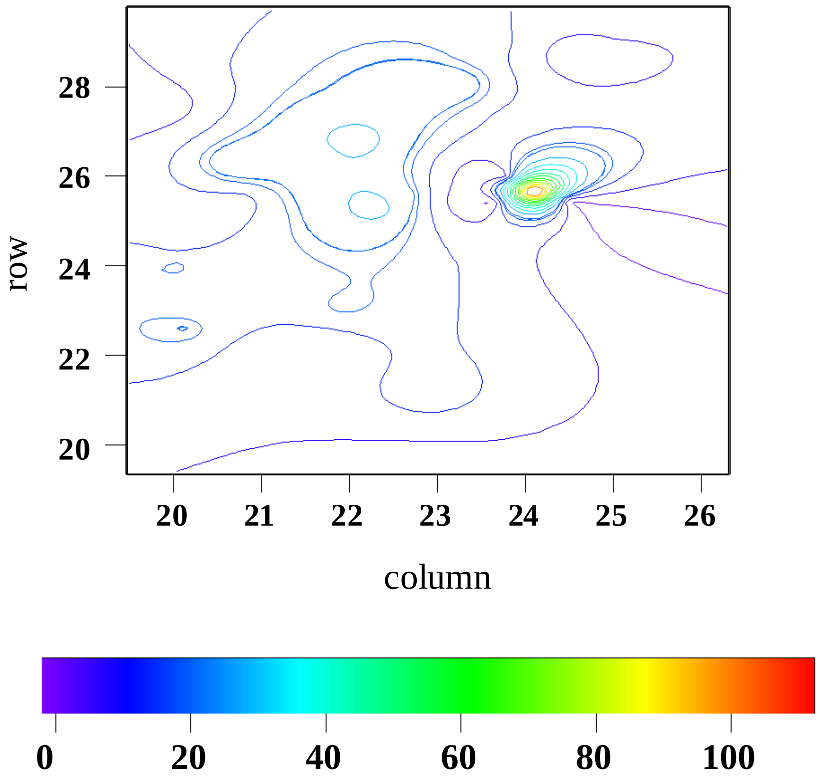


EDA – Contour Mapping

Contour mapping provides a quick way to map the data.

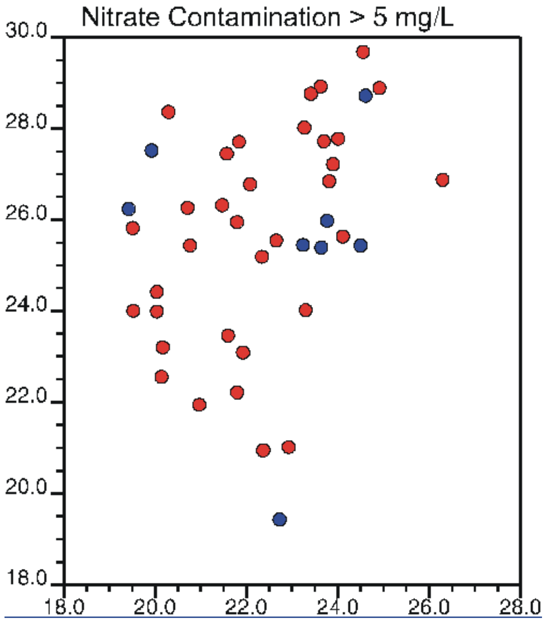
Sharp high or low values lead to steep gradients in contour map and may indicate bad data

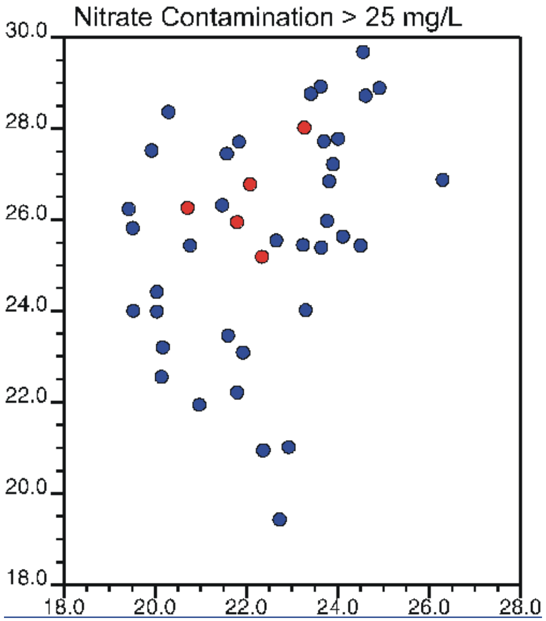
Different contouring algorithms give different maps. Which one is best?



EDA – Indicator Plots

$I(x)=1$ if $Z(x) > Z^*$; $I(x) = 0$ otherwise



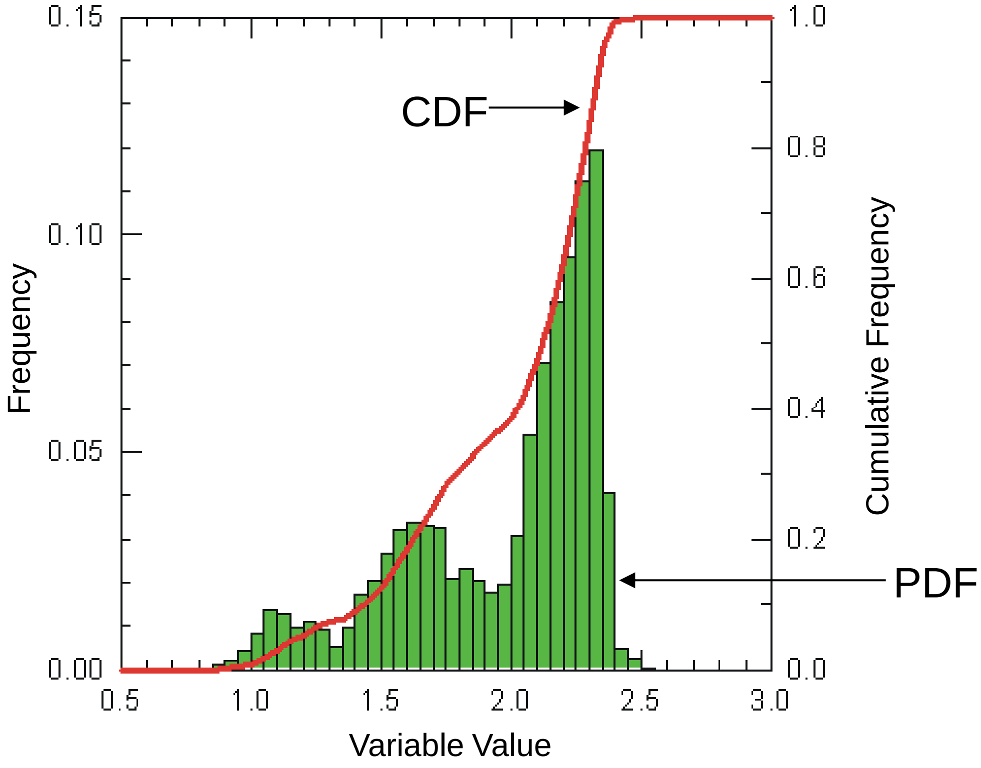


EDA – Histograms

- Simple Histograms
  - Probability density function (PDF) is *histogram*
  - Cumulative density function (CDF)
  - Check for outliers (and their cause)
  - Multimodality: evidence for multiple processes
  - Clustering of data or preferential sampling

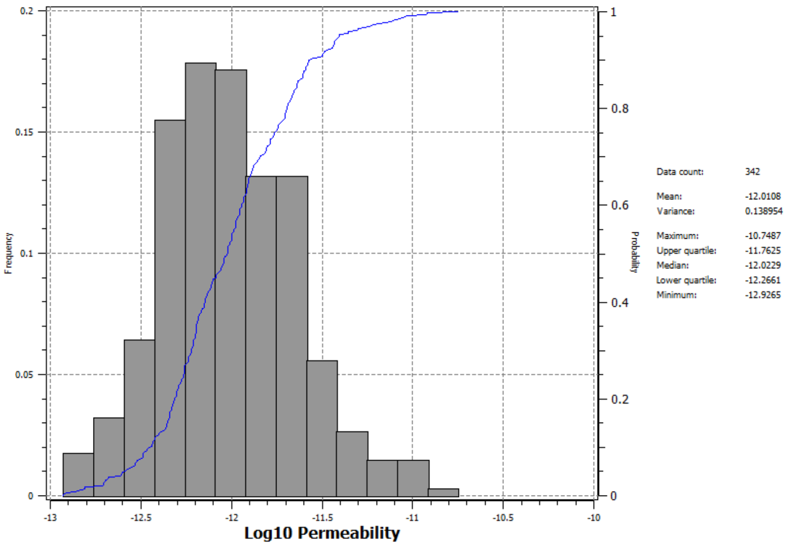
- CDF format provides an important conceptual link to downstream modeling (Monte Carlo Models)
- Relationship of the histogram and other "ensemble statistics" to model output
- Need to decluster data to remove effects of preferential sampling (need an unbiased estimate of sampled property)

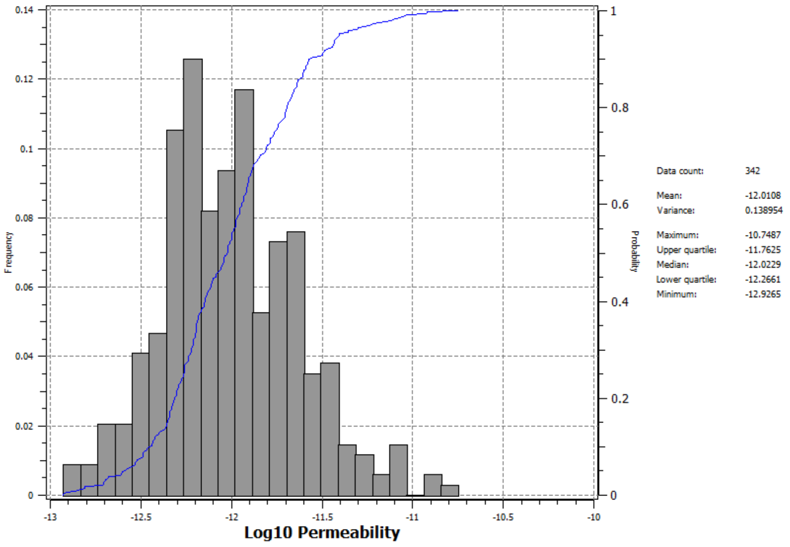
EDA – PDF and CDF plots



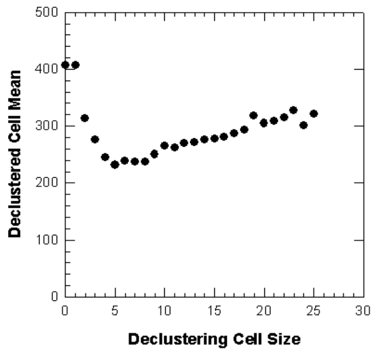
EDA – PDF and CDF plots

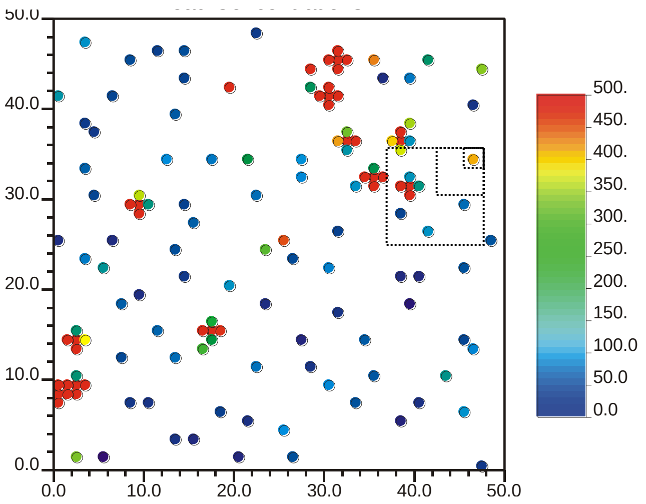
- PDF plots of the same data may appear significantly different depending on the binning of the data
- Care should be taken to examine PDF plots for different numbers of bins to avoid misinterpretation
- CDF plot is unique for a given data set and not prone to misinterpretation from plotting preferences

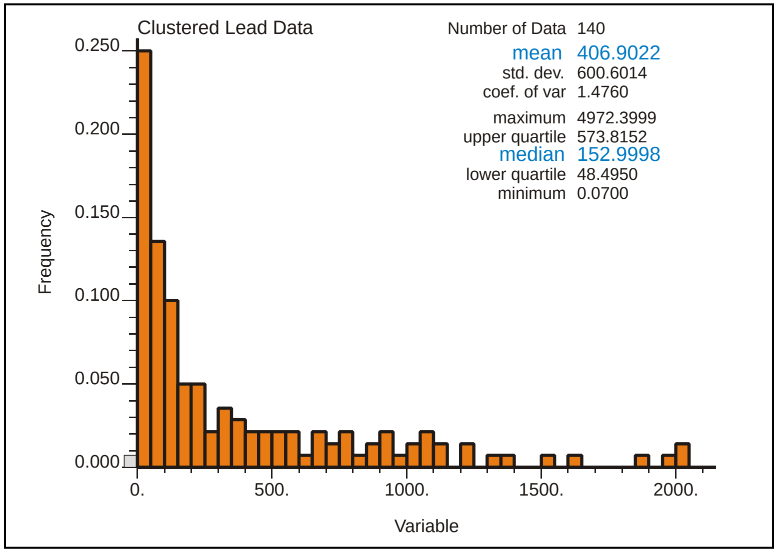


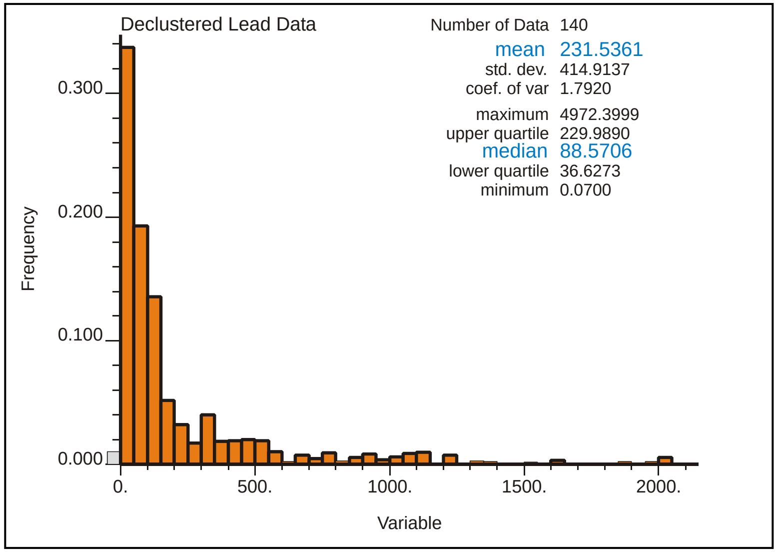


EDA – Sample Clustering and Plots









EDA – Probability Plotting

Simple Probability Plots

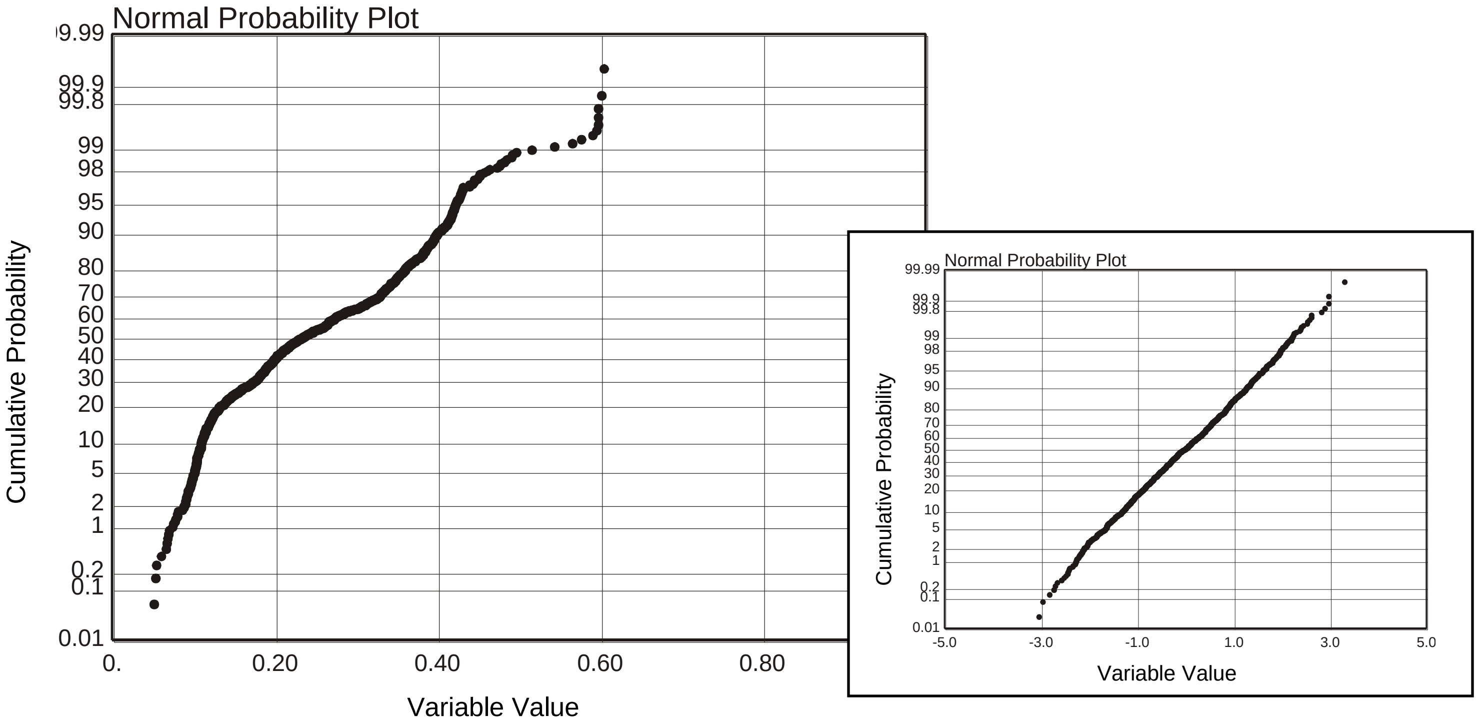
- plot sample value against probability value rather than against simple frequency
- implicit comparison to theoretical Gaussian distribution
- different underlying populations will plot as straight line segments
- can compare to populations other than Gaussian

Quantile-Quantile Plots

- plot corresponding quantiles of any two populations
- use in understanding cross-correlated variables
- use in “validating” output models

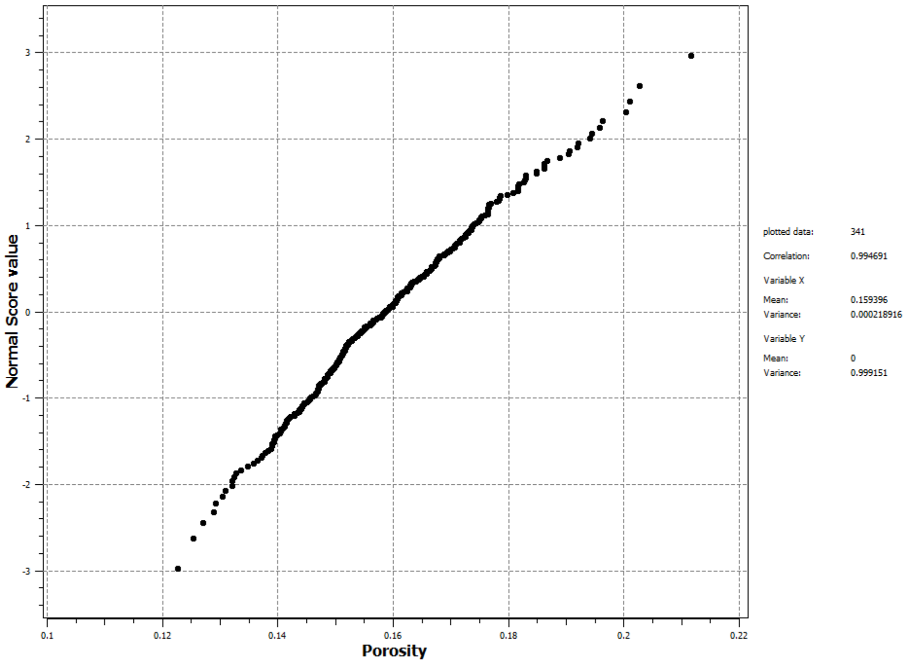
EDA – Normal Probability Plot

Plots sample values vs. Gaussian probability



EDA – Normal Probability Plot

Example probability plot from workshop exercises – using normal score transform value for the Gaussian probability axis



plotted data:	341
Correlation:	0.994691
Variable X	
Mean:	0.159396
Variance:	0.000218916
Variable Y	
Mean:	0
Variance:	0.999151

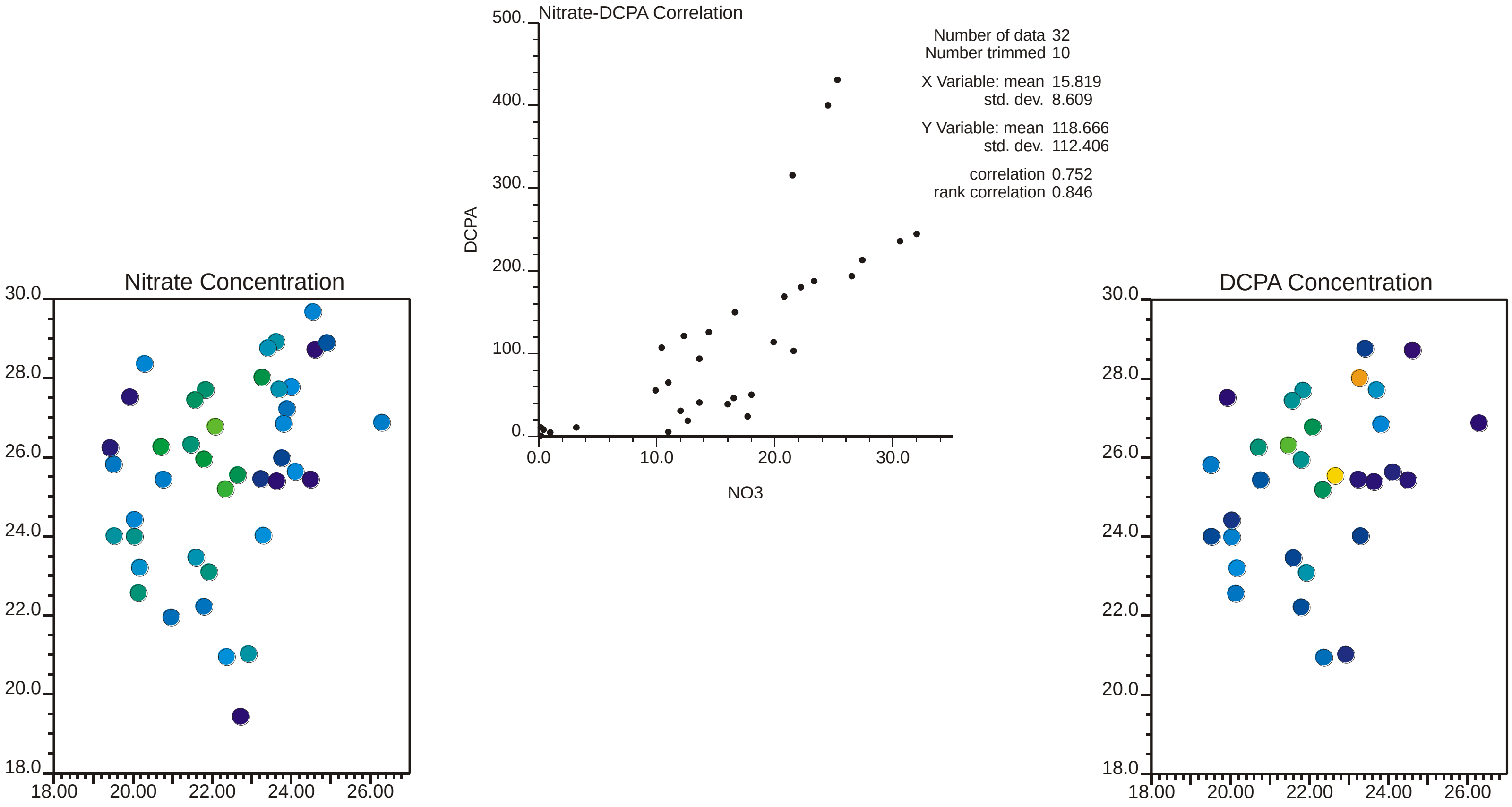
EDA – Correlation of Multivariate Data

Multiple properties of interest

Multiple measurement methods

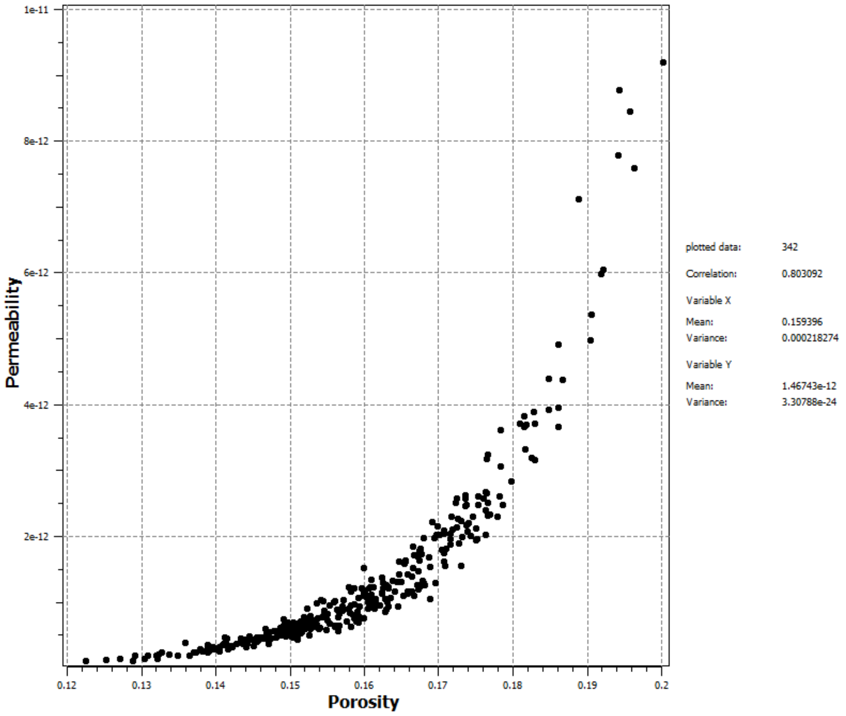
- For our purposes: Scatterplot Analysis
  - direct and inverse correlation
  - strength of correlation
    - correlation coefficient (r)
    - coefficient of determination (r^2)
    - rank-order correlation coefficient
- Concept of conditional expectation

EDA – Correlation of Multivariate Data

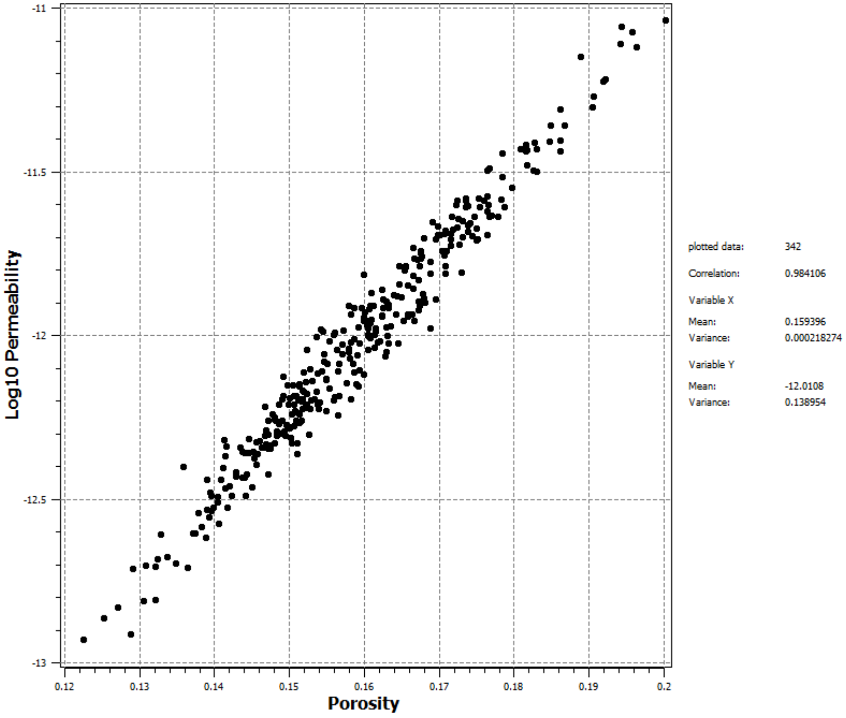


EDA – Correlation of Multivariate Data

Example scatter plots from workshop exercises

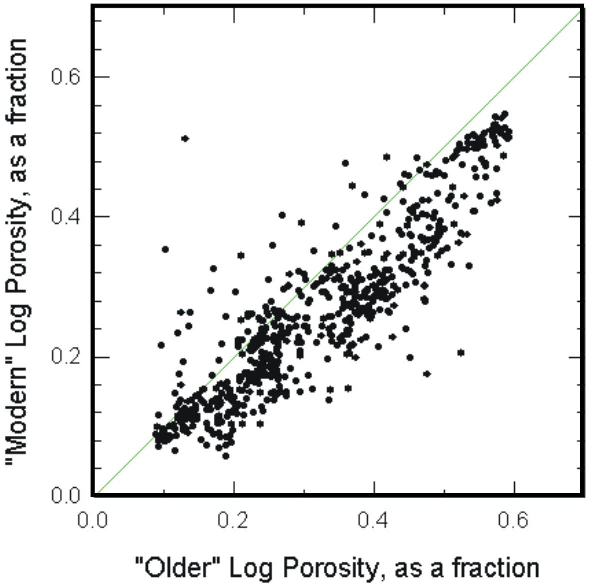


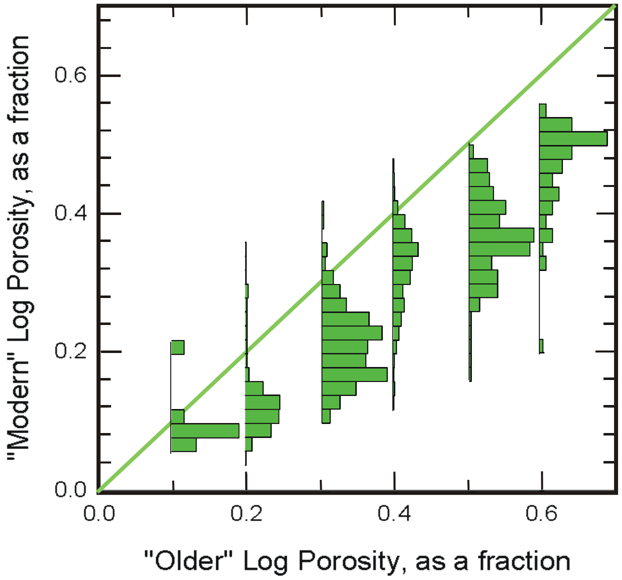
```
plotted data:       342

Correlation:        0.803092

Variable X

Mean:               0.159396
Variance:           0.000218274

Variable Y

Mean:               1.46743e-12
Variance:           3.30788e-24
```



```
plotted data:       342

Correlation:        0.984106

Variable X

Mean:               0.159396
Variance:           0.000218274

Variable Y

Mean:               -12.0108
Variance:           0.138954
```

EDA – Conditional Expectation





EDA – Data Transformations

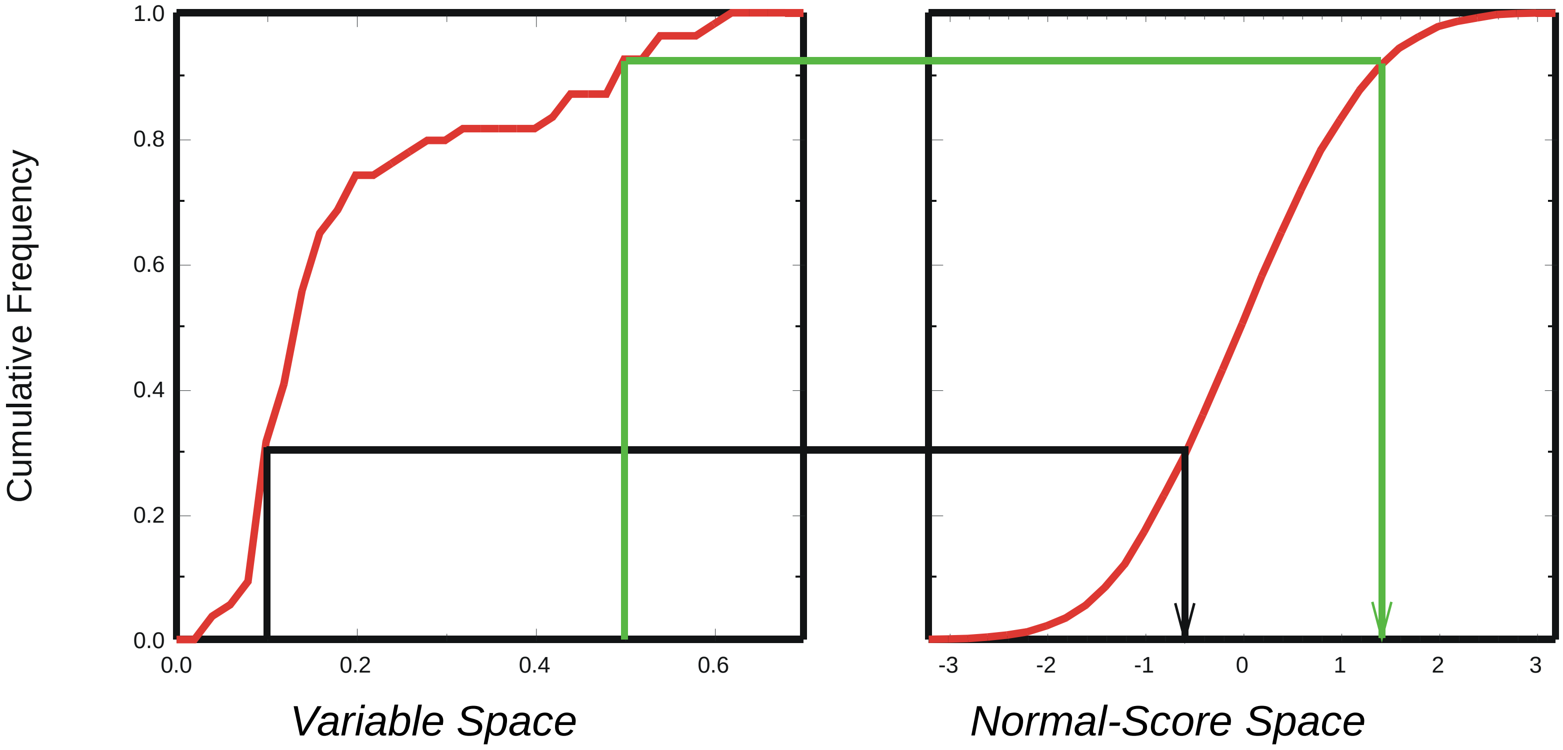
- A powerful tool for understanding data.
  - reduce numerical artifacts that can obscure relationships
  - simplify portions of numerical modeling
- A “two-edged sword” (benefits and difficulties)
  - back-transformation may have negative implications

Examples:

- logarithmic $U = \log (Z)$
- indicator $I = 1$ if $Z < Z^*$; $I = 0$ Otherwise
- rank-order Z in order 1,2,3, N
- normal-scores $\mu = 0$ $\sigma^2 = 1$
- uniform-scores $[0, 1]$

EDA – Normal-score Transform

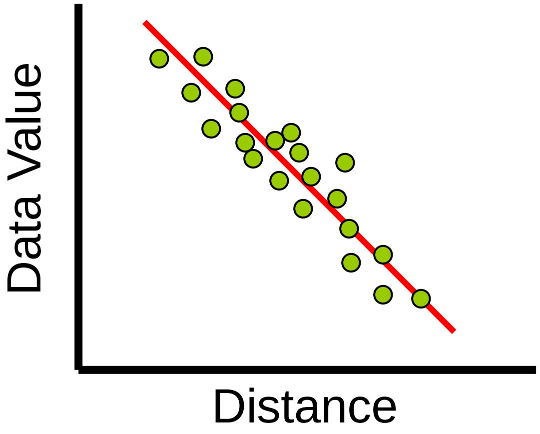
Graphical conceptualization of quantile-preserving process

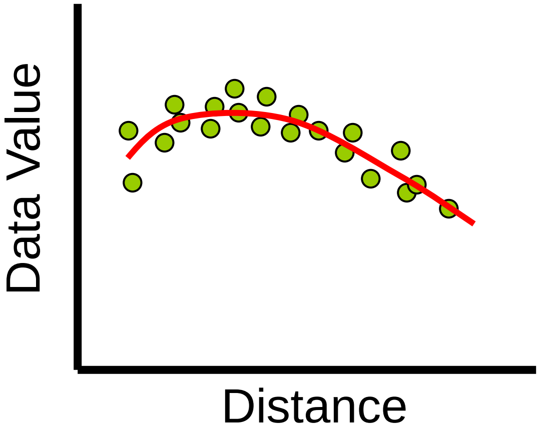


EDA – Trends

What is a trend?

- **Geostatistics assumes second-order stationarity**
- **“Deterministic Geologic Processes”**
  - **Trend analysis and modeling must make geologic sense**
- **Removing a trend - Analysis of residuals**





EDA – Summary

Focused principally on understanding the data

- Error checking
- Physical process responsible for deposition
- Development of reasonable target characteristics for models (also used for validation)

Techniques

Mapping

Transformations - pros/cons

Histograms

Distributional analysis

Declustering

Trend removal

Spatial Correlation Analysis

- Measurements of an earth science variable are rarely independent.

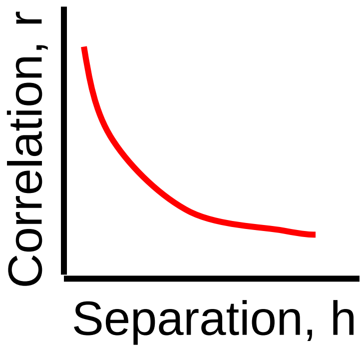
  Independence is the premise underlying sampling theory based on traditional statistics.

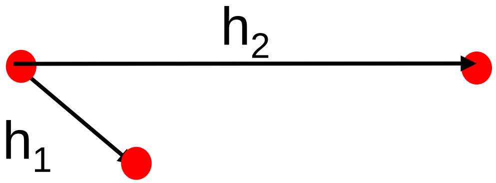
- It is this emphasis on spatial correlation that sets geostatistics apart from traditional statistics

- The traditional measurement of spatial correlation within geostatistics is the semi-variogram, commonly called the variogram.

Spatial Correlation

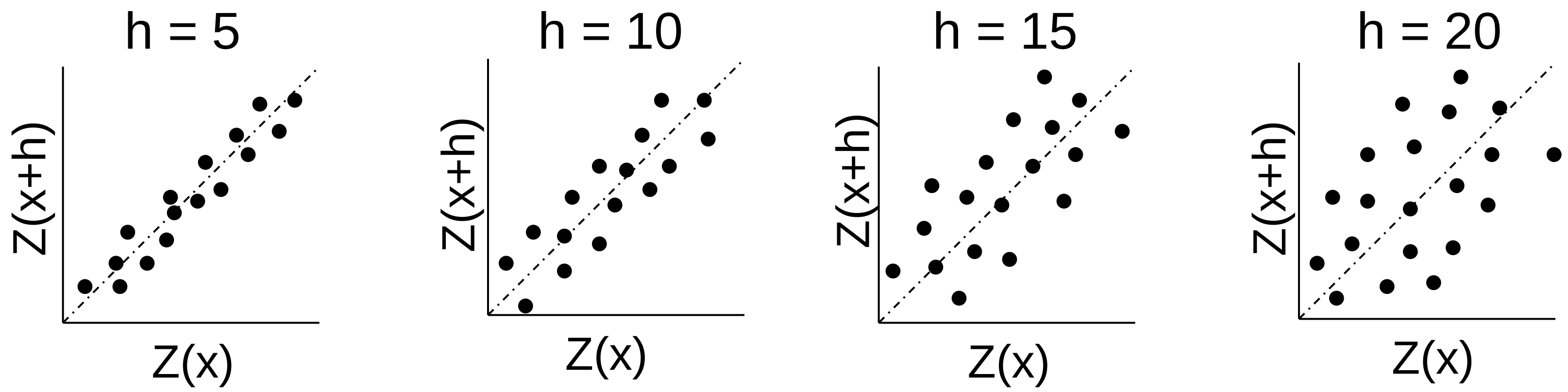
How does the correlation coefficient, *r*, behave as *h* increases?





- **The greater the distance between points, the less correlated the values.**
- **Since h is a vector, direction matters. Separation and differences may be different in different directions.**

Scatterplot Example



A scatter plot is a means of seeing the variability of sample values for all sample points separated by a distance *h*.

- **At small separations between any pair, correlation is strong**
- **As the separations between samples increase, correlation decreases**

Stationarity

Stationarity is the invariance of a property (e.g., the mean) across space or time.

A statistically homogeneous field is the result of a stationary process.

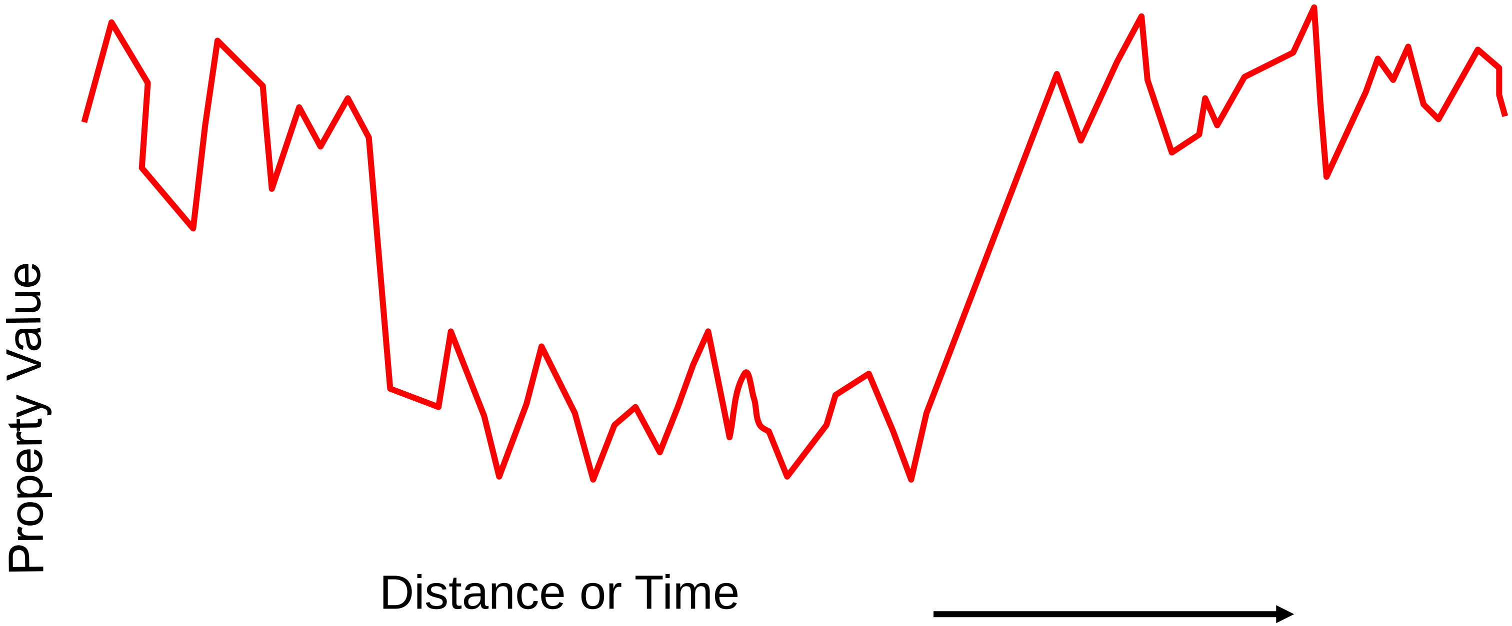
- ***First order stationarity*** **refers to the mean value remaining constant in space**
- ***Second order stationarity*** **refers to the mean and variance being constant in space**

A Non-Stationary data set will show a trend in space or time of the mean or variance of the data set.

Stationarity

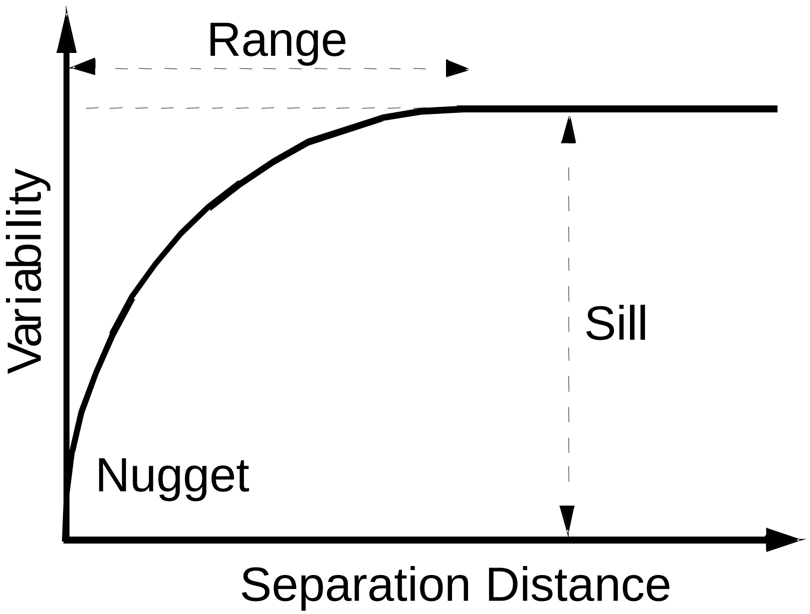
- Measurements collected in a small area should be strongly correlated because there is a relatively small separation distance between samples
- Measurements collected in another area a couple of miles away should also be strongly correlated to each other because of small separation distances.
- But if the two sample groups are compared and do not show correlation this may be evidence of non-stationarity.

1st Order Stationarity



Analyzing Spatial Correlation

- **In geostatistics we tend to look at the opposite of correlation, which is variability.**
- **At very close distances variability is low, and as the separation distance increases, so does variability.**

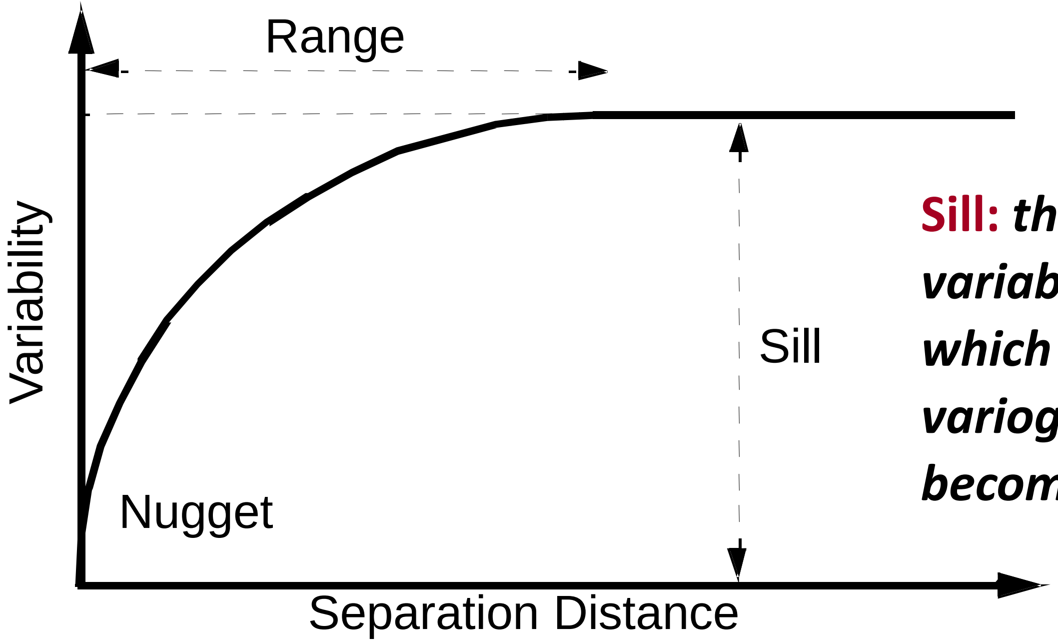


Variogram

The variogram is a measure of variability as a function of separation distance *h*.

Nugget: **some amount of variability at zero separation: a representation of measurement error or variability at separations smaller than the sample distance.**

Range: ***distance at which we reach the total amount of variability***



Sill: ***the total variability level at which the variogram value becomes constant***

Variogram Equation

1/2 the average squared difference between all values separated by distance *h*.

$$\gamma(h)=\frac{1}{2n(h)}\sum_{i=1}^{n(h)}\left(z_i(x)-z_i(x+h)\right)^2$$

Where: γ **is the variability**

z(x) **is the value at location *x***

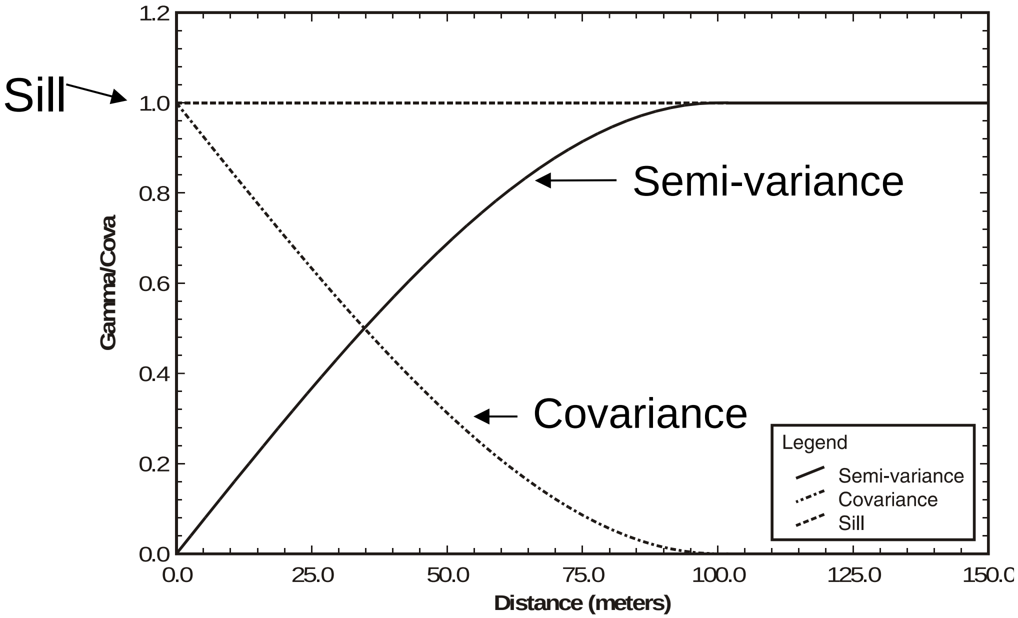
z(x+h) **is the value *h* away from location *x***

n **is (the number of values that are separated by *h*)**

This gives a value for variability at the given h, and the value is a point on the experimental variogram. Repeat for each value of h.

Variogram – Covariance Relationship

$$C(h) = Sill - \gamma(h)$$

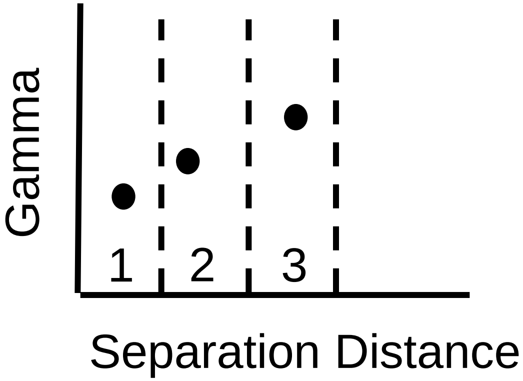


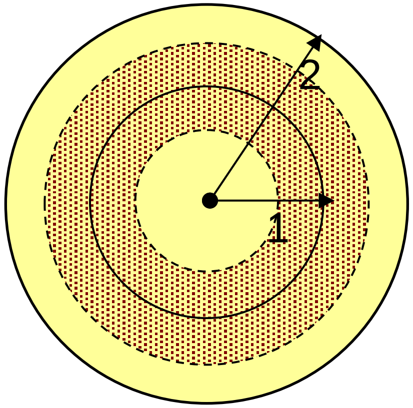
Covariance is the inverse of the variogram

This simple relationship between variogram and covariance is true under the assumption of second order stationarity

Variogram – Search Neighborhood

- To determine how many samples are a given h away from a certain location, a search neighborhood is used.
- The simplest search neighborhood (isotropic) includes all locations in a specified concentric ring away from the current location.
- Determine the average spacing of all values lying between *h-1/2* and *h+1/2*. This average spacing is the x coordinate of the point on the experimental variogram.

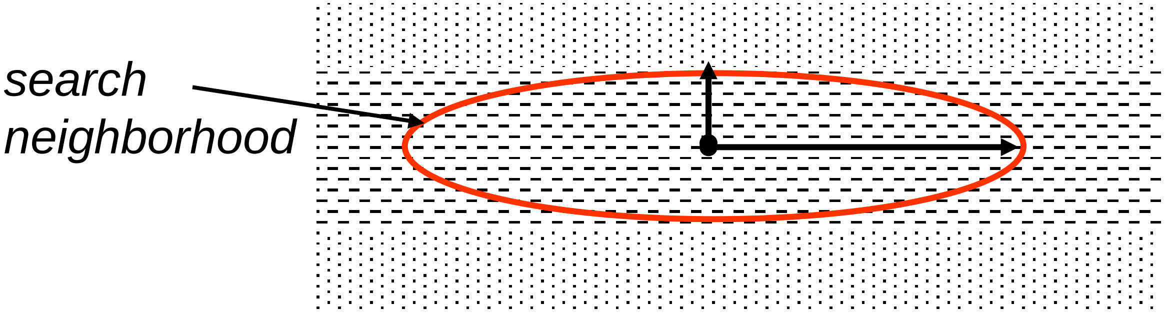




Variogram – Search Neighborhood

Properties in the earth and environmental sciences are often deposited/produced in anisotropic patterns.

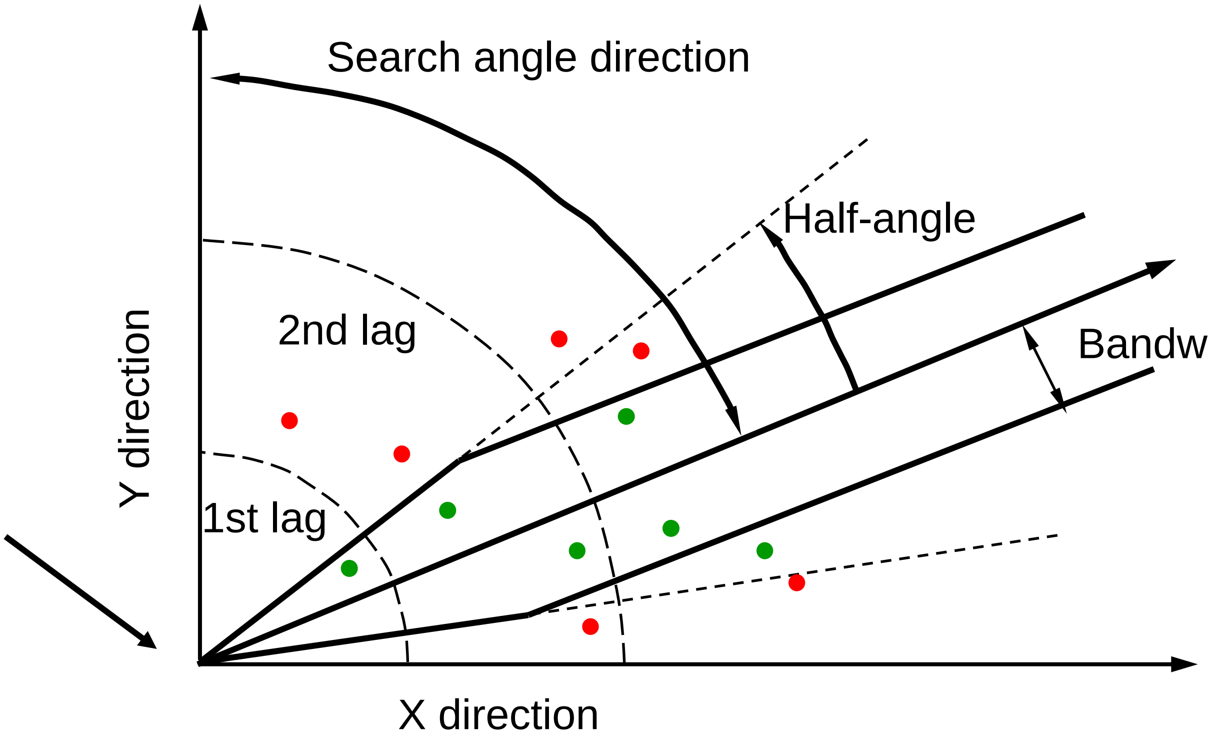
- **Rather than using a circular search neighborhood, may want to use an ellipse oriented along the principle direction of correlation.**
- **For example with sedimentary layers: vertically there are changes in types of rock and large sample variations, horizontally the beds are very similar even at large distances**



Variogram – Search Neighborhood

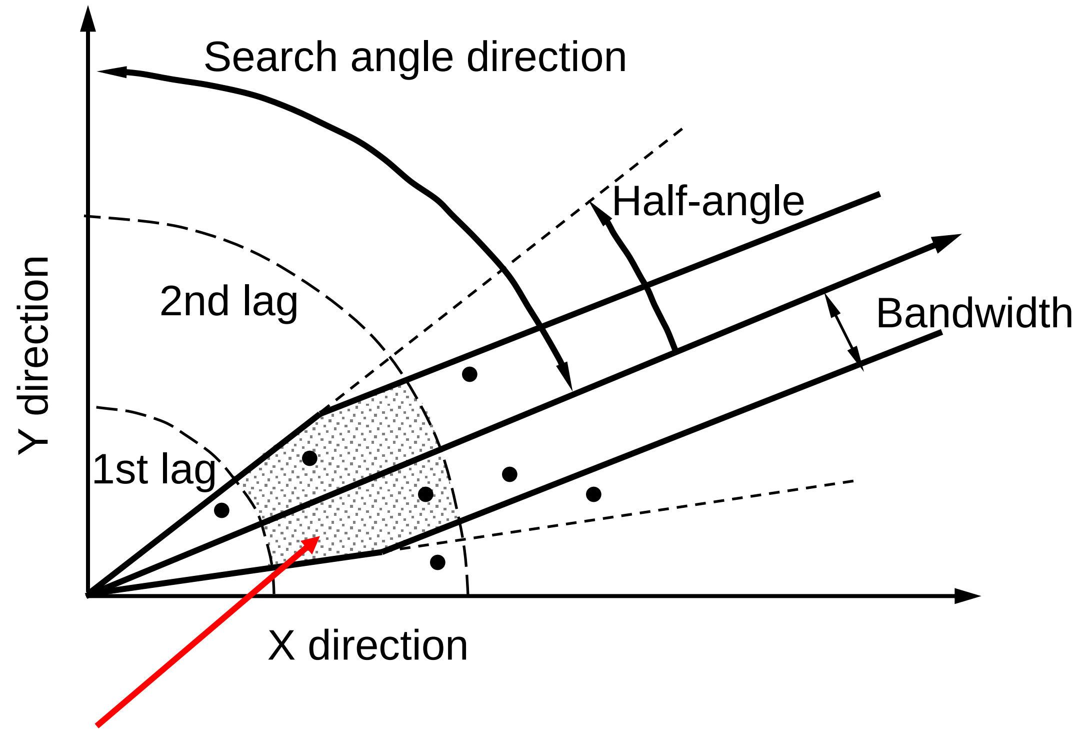
Use geological knowledge of genetic processes to customize search along a preferred orientation. Orient search along this direction, the search direction.

We have a point here, at the origin, and want to search for nearby points.



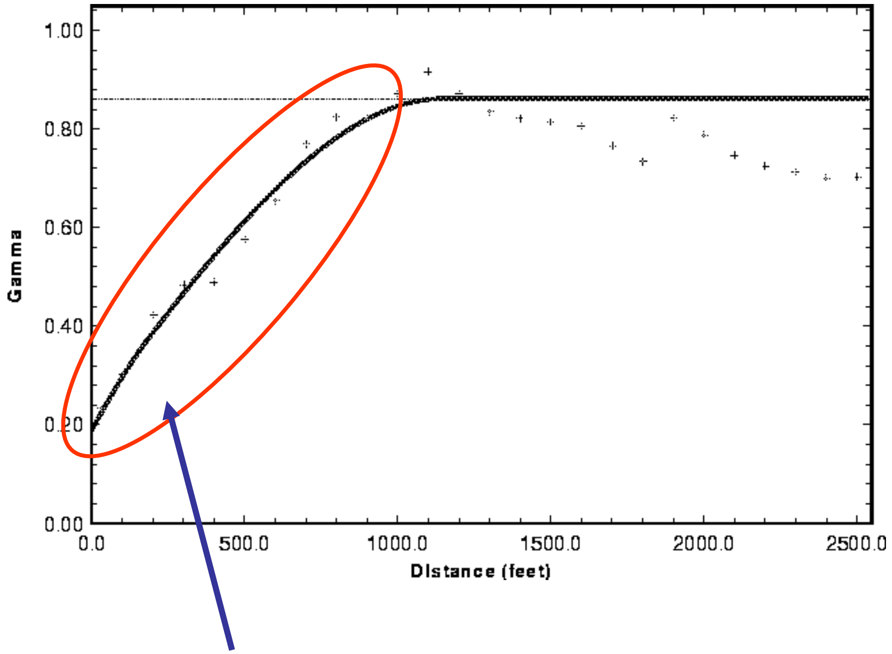
Variogram – Search Neighborhood

The search neighborhood diagram is a template which can be moved to different points and different directions.



One point on the variogram will be generated for each lag

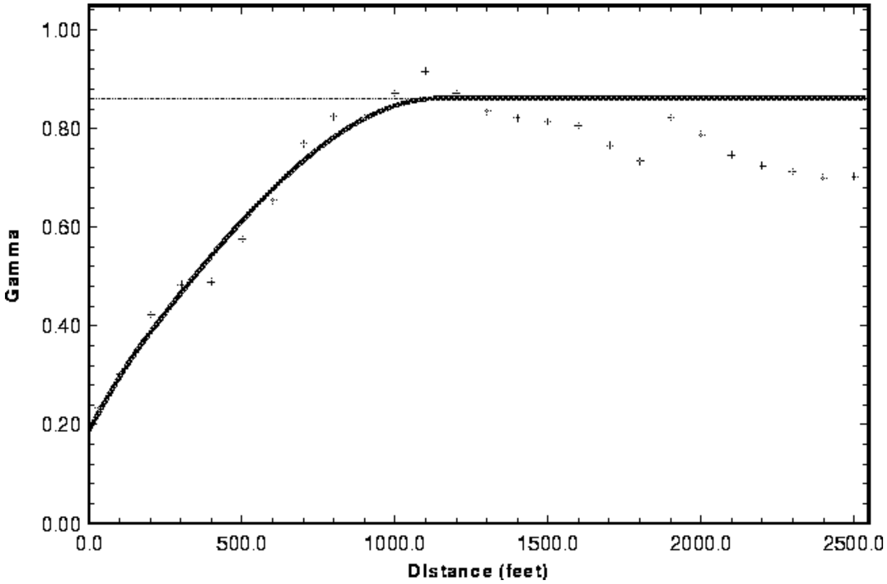
Experimental Variogram



After employing the search neighborhood and entering the points into the variogram equation, the experimental variogram (shown by crosses) is produced.

Typically the experimental variogram starts at small values, increases, then follows, on average, the sill. In modeling, the emphasis is placed on fitting the experimental variogram prior to the sill.

Experimental Variogram



Generally, employ one of a limited number of theoretical models that always yield positive definite matrices

- Fitting a model is still an art

Usually emphasize the model fit to the experimental variogram at smaller *h*.

Variogram Models: Spherical

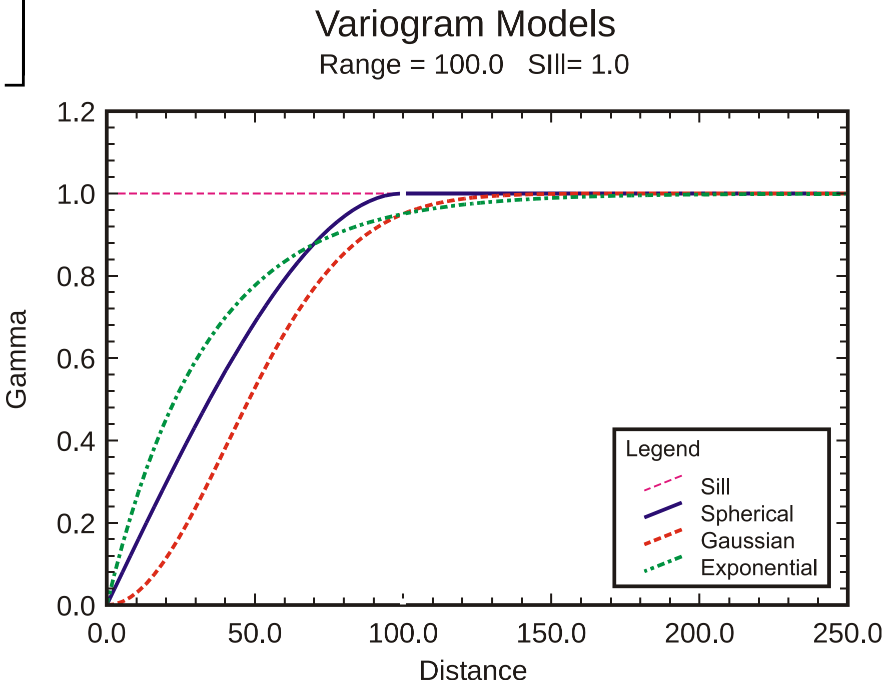
h<a: $$\gamma(h) = C \cdot \left[1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3 \right]$$

h≥a: $$\gamma(h) = C$$

Where C = sill value

a = range

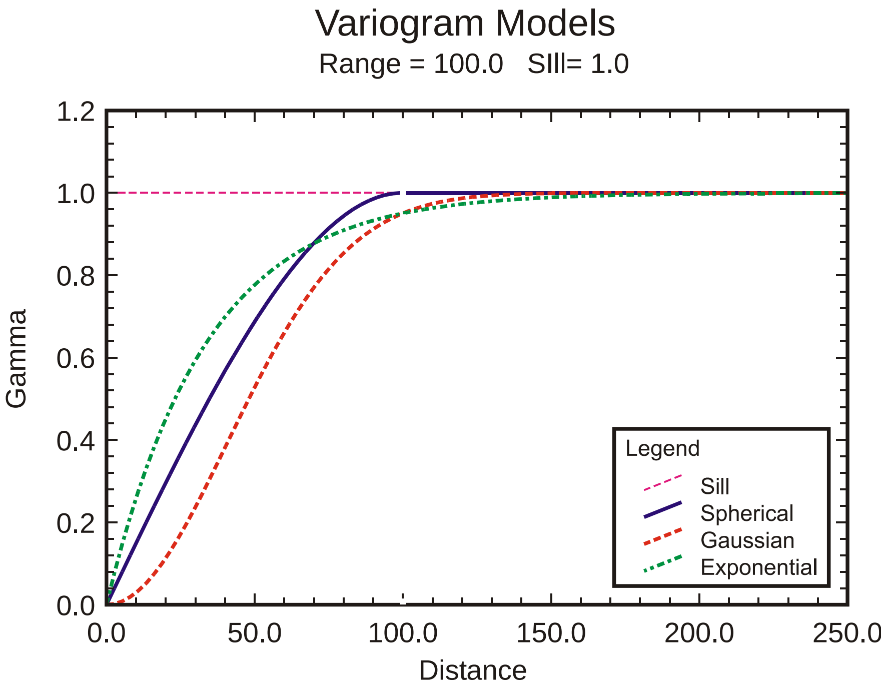
h = lag distance



The spherical model is linear at the origin and the range parameter is exactly the correlation length

Variogram Models: Gaussian

$$\gamma(h) = C \cdot \left[1 - e^{\left(-\frac{(3h)^2}{a^2} \right)} \right]$$



Where **C = sill value**

a = effective range

(95% of sill)

h = lag distance

The **Gaussian model** **has a very slow increase near the origin (low variability at small values of *h*).**

Variogram Models: Exponential

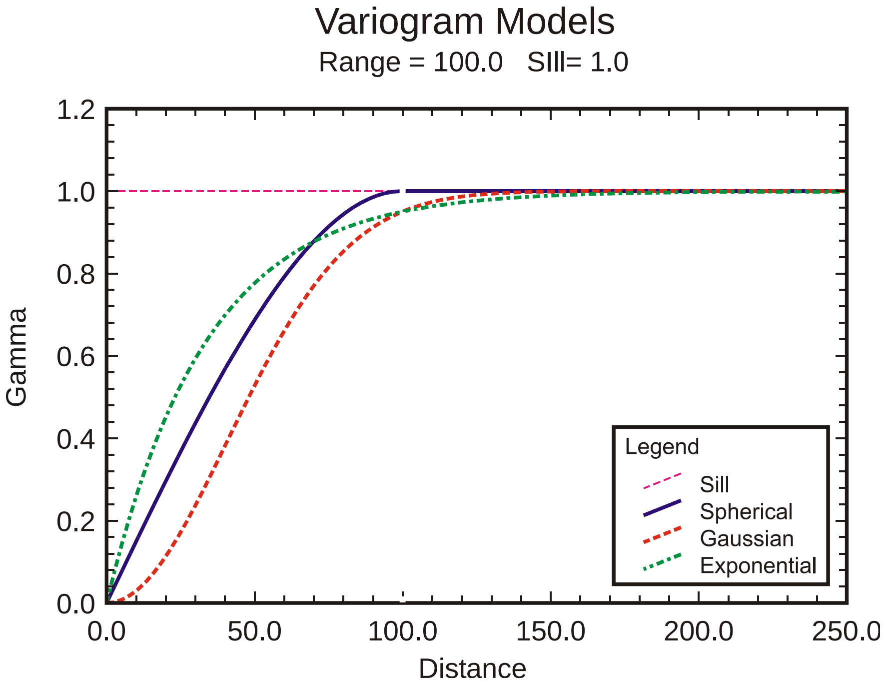
$$\gamma(h) = C \cdot \left[1 - e^{-\frac{3h}{a}} \right]$$

Where C = sill value

a = effective range

(95% of sill)

h = lag distance

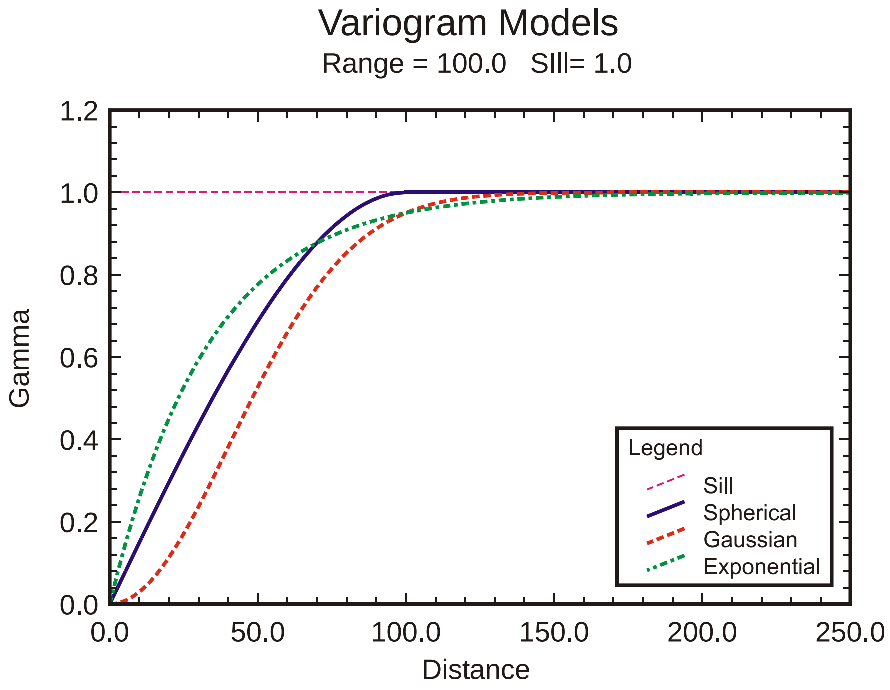


The exponential model displays a continuously increasing level of variability and an asymptotic approach to the sill

Common Variogram Models

Often difficult to select a unique variogram model. Typically experimental variograms show variability, so exact choice of model may be difficult.

Exponential and Gaussian have a practical range where the model hits the sill, but often use a different definition of the range: the point at which 95% of the sill is reached (effective range)



Generally the spherical model is used most often and is most straightforward.

Nested Variogram Models

- Sometimes it is necessary to fit complex structures that may be caused by a combination of processes.
- You can add models together to capture a particular curve that you may want to interpret

$$\gamma(h)_{total} = \gamma_1(h) + \gamma_2(h) + \ldots + \gamma_n(h)$$

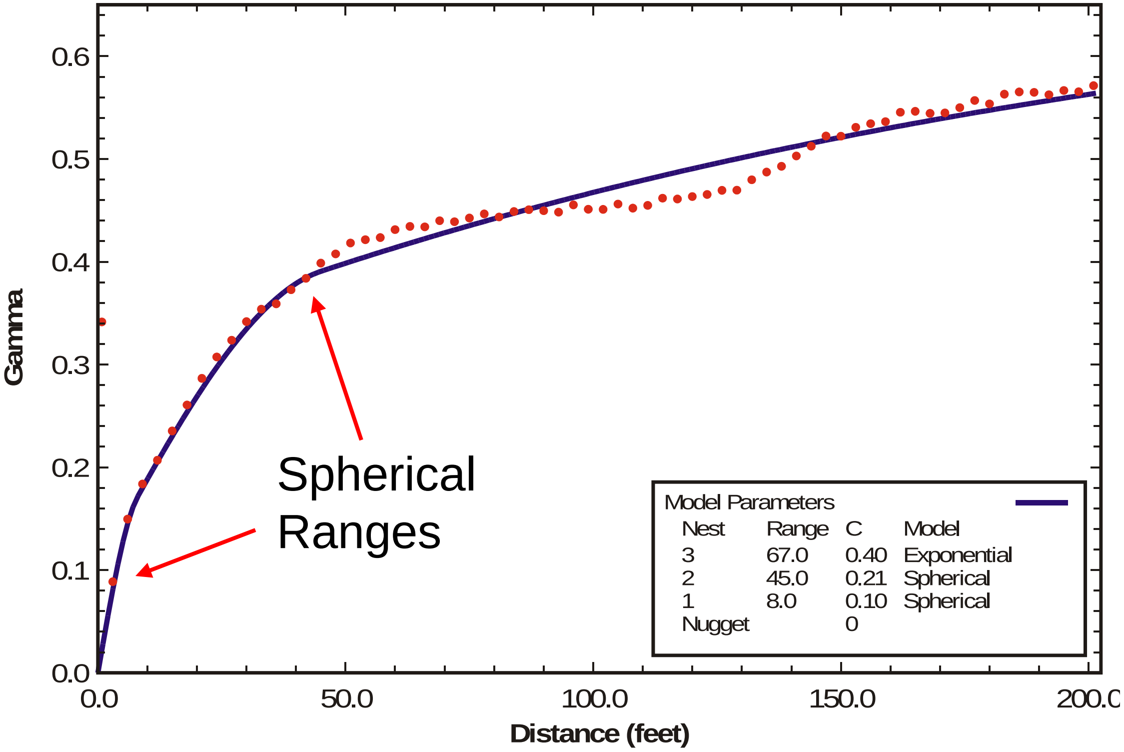
Nested semivariogram models can be created using any linear combination of admissible models. (After Olea, 1994)

Any linear combination will produce a model with a positive definite covariance matrix

Nested Variogram Models

Example shows three nested models:

two spherical
and
one exponential.



Spherical Ranges

Nest	Range	C	Model
3	67.0	0.40	Exponential
2	45.0	0.21	Spherical
1	8.0	0.10	Spherical
Nugget		0	

Anisotropy

- Variograms that show variation as a function of search direction are anisotropic
- Anisotropy in the variable requires fine tuning of search neighborhood

The general types of anisotropy are:

Geometric: ***<u>Constant Sill, Range changes with direction</u>***

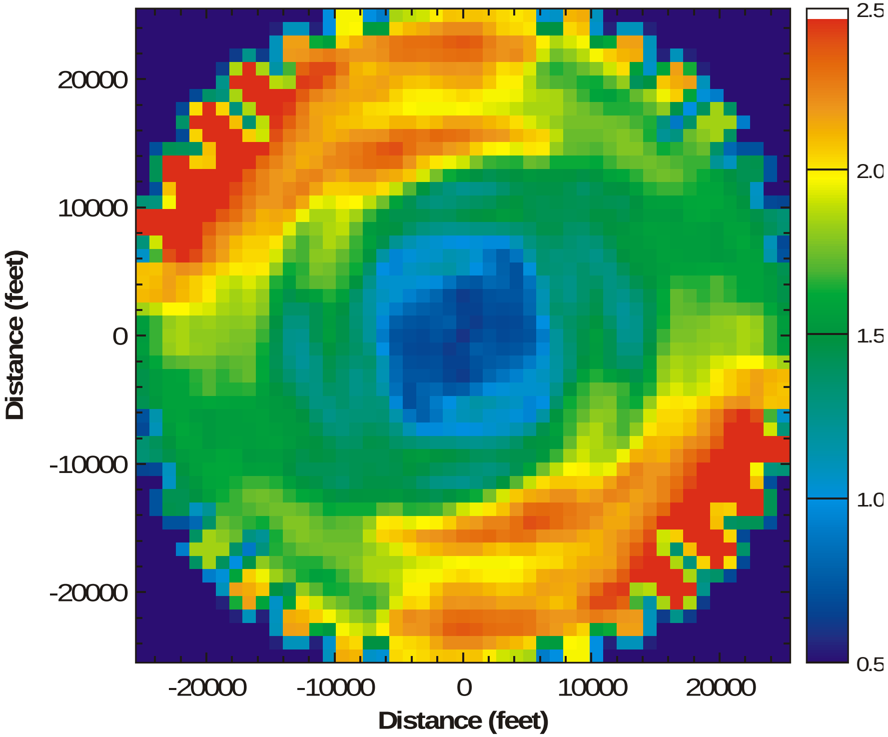
Software almost always requires anisotropy to be geometric.

Zonal: ***<u>Constant Range, Sill changes with direction</u>***

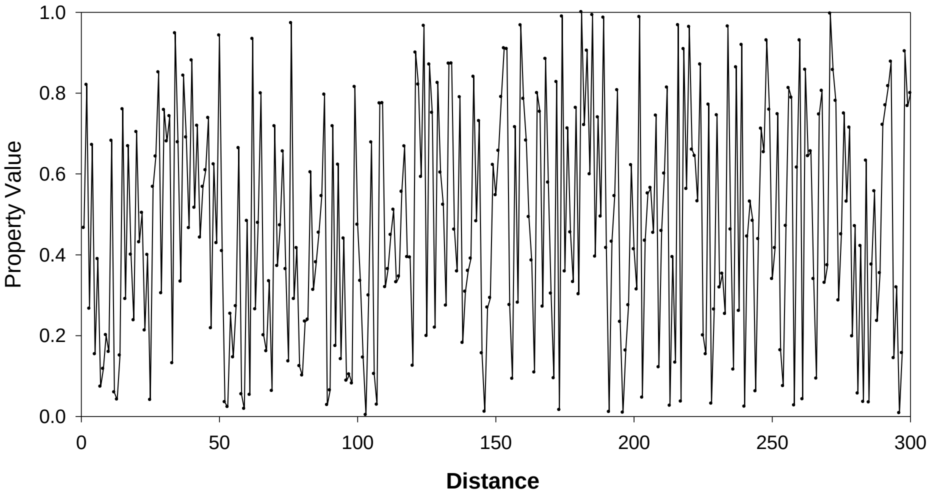
The level of variability is different in different directions.

Variogram Map

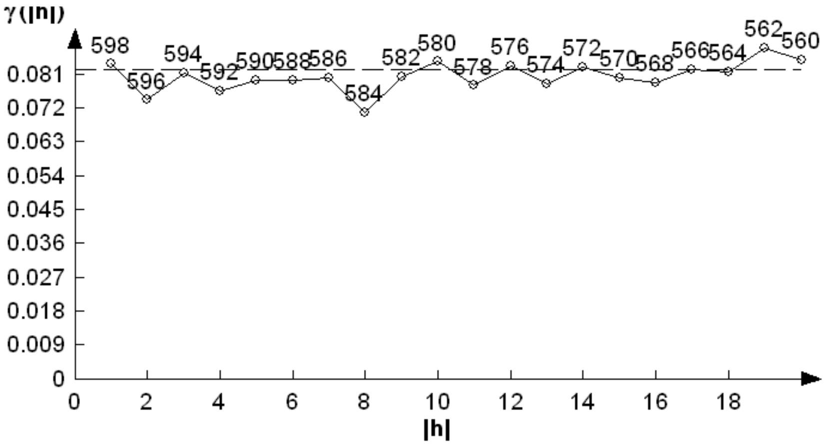
A variogram map can provide a visual check for variogram anisotropy



Nugget Effect Variogram



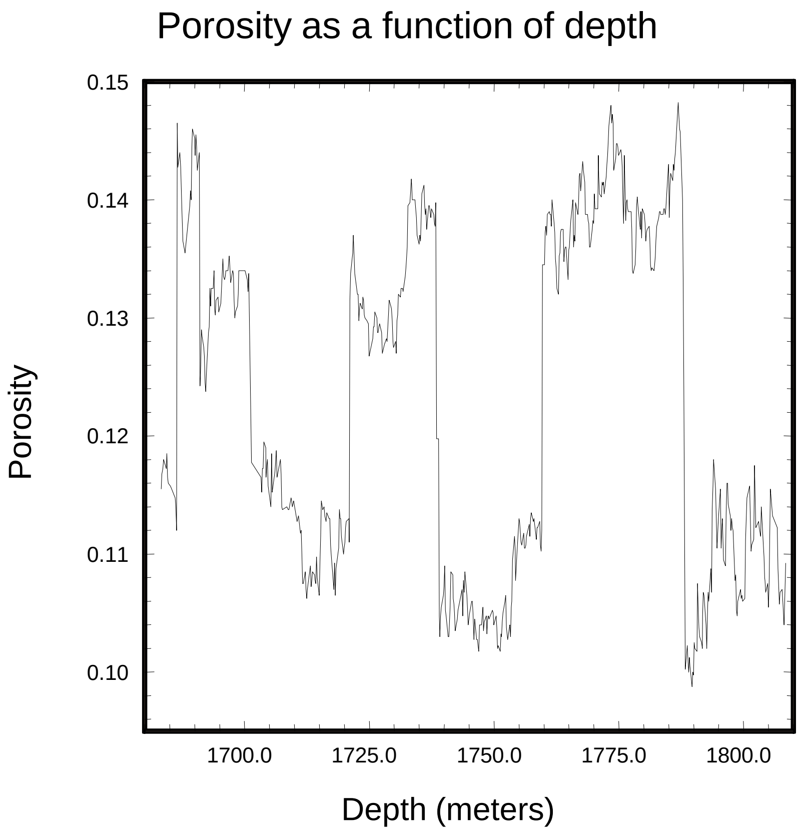
Sampling transect of random, uncorrelated data



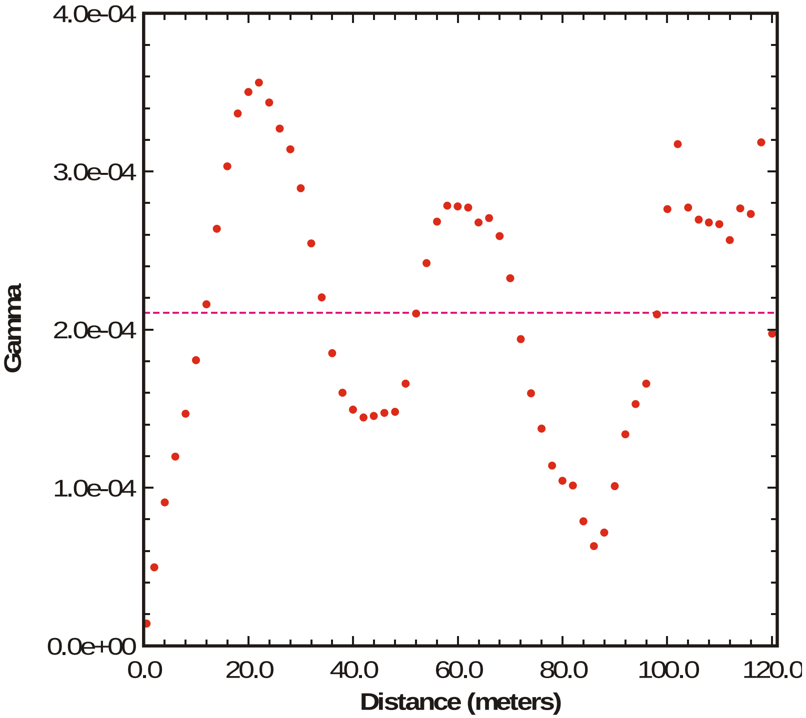
Resulting "nugget effect" variogram

No spatial correlation

Hole Effect Variogram

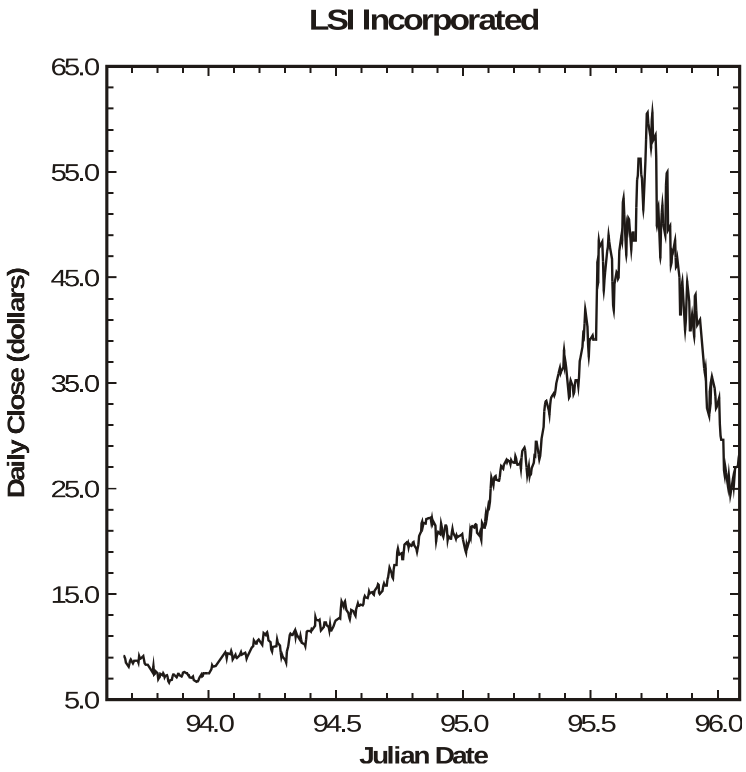


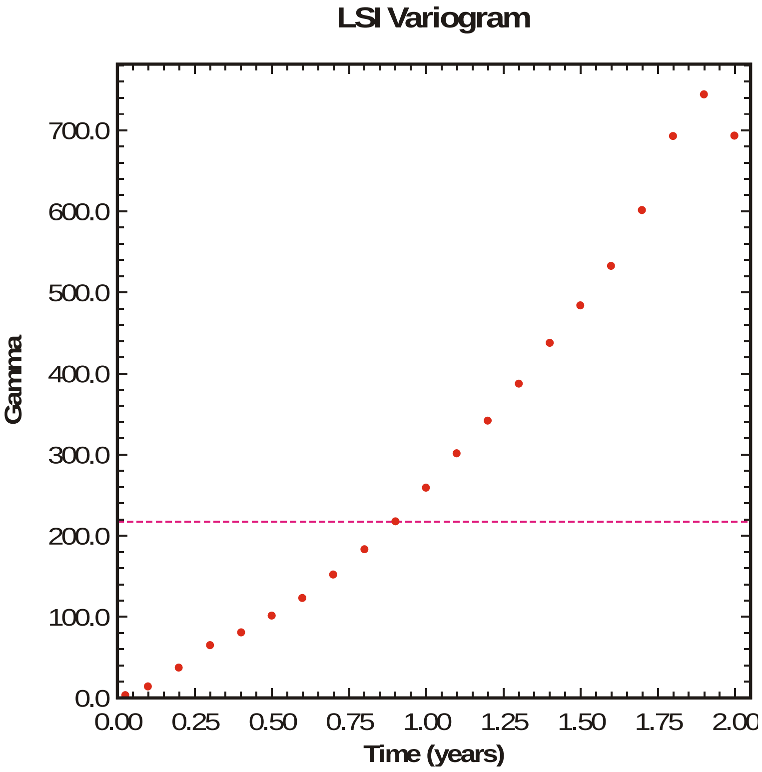
Periodic Data



Hole Effect Variogram

Trend Effect Variogram





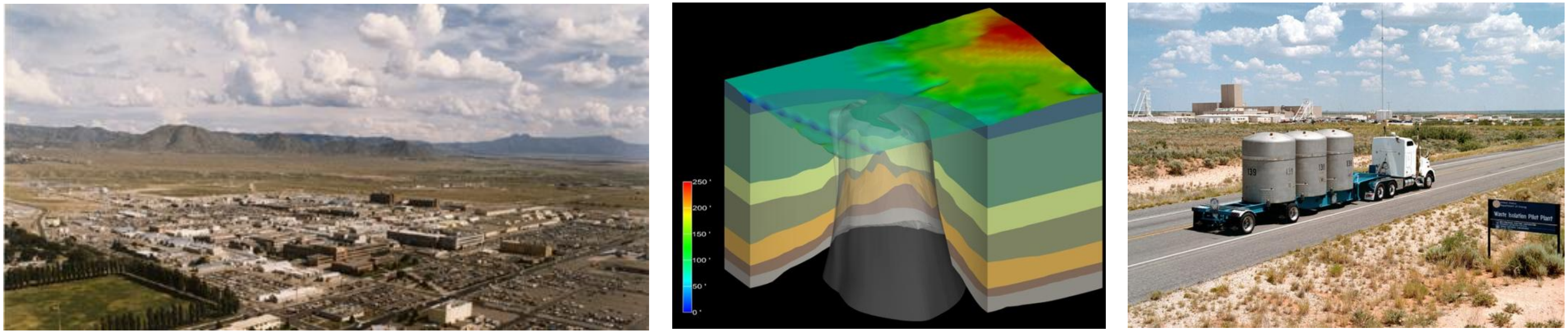
Spatial Correlation Summary

- Determination of spatial correlation
- Relationship between variogram and covariance
- Calculation and modeling of experimental variograms including definition of search neighborhoods
- Differences between several permissible variogram models
- Concept of anisotropy and how to find it with a variogram map
- Review of special variograms (nugget-effect, hole-effect and trend)

SAND2013-xxxx

Exceptional service in the national interest

Sandia National Laboratories



Exercise 1: Exploratory Data Analysis

IAEA Training Course – July 1-5, 2013

Bill Arnold

U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration
IAEA International Atomic Energy Agency

Outline

- Software
- Exercise data files
- Exploratory data analysis exercise tasks

Software

- SGeMS (Stanford Geostatistical Modeling Software)
  - Windows-based software with graphical users interface (GUI)
  - Open source software from Stanford University
  - Based on the original GSLIB suite of DOS-based software
  - Available for download from website http://sgems.sourceforge.net/
- Two GSLIB software codes – NSCORE and LOCMAP
  - DOS-based codes run using parameter input files
  - Available for download from website http://scrf.stanford.edu/resources.software.gslib.php
- GSview software
  - Used for viewing postscript output files generate by GSLIB codes
  - Available for download from website http://pages.cs.wisc.edu/~ghost/

Exercise Data Files

- **Four data sets are provided for use in the geostatistics exercises**
    - **2-D data set 1**
    - **2-D data set 2**
    - **2-D data set 2 - exhaustive**
    - **3-D data set 3**
- **Data sets were randomly extracted from a hypothetical synthetic geological system**
- **Data are given for rock porosity and permeability**
- **Data sets differ in spatial correlation structure in ways that the students will explore and discover**

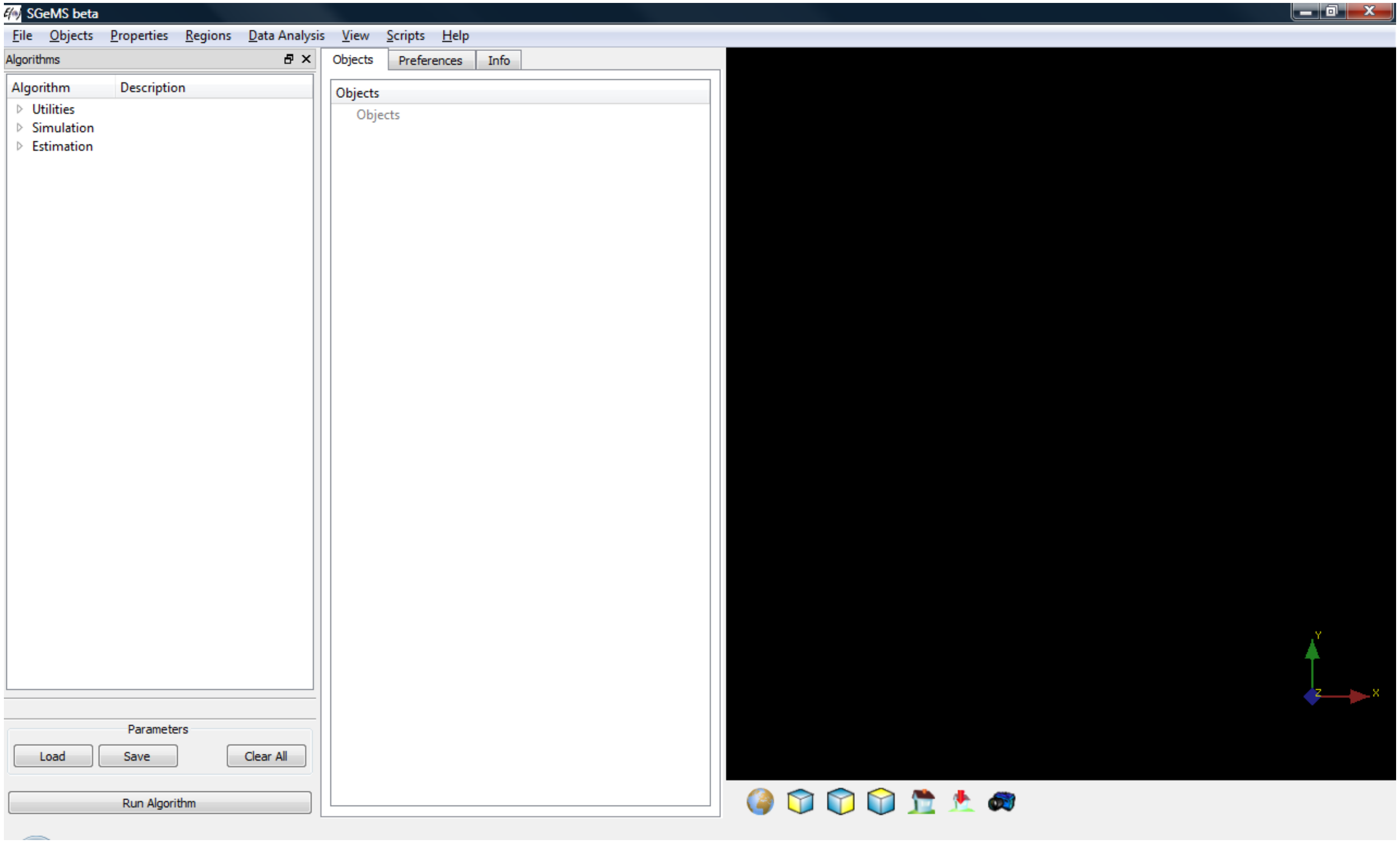
Data File Format

- **SGeMS can read data from the GSLIB format. Exercise data files are provided in this format.**

```
surface data set 1
6
X coordinate
Y Coordinate
Z Coordinate
Porosity
Permeability
Log10 Permeability
25.0        225.0       200.0   0.1362   2.904E-13   -12.537
940.0       175.0       200.0   0.1596   8.615E-13   -12.065
630.0       345.0       200.0   0.1647   1.223E-12   -11.912
835.0       705.0       200.0   0.1562   7.652E-13   -12.116
980.0       690.0       200.0   0.1617   9.353E-13   -12.029
170.0       105.0       200.0   0.1517   7.856E-13   -12.105
```

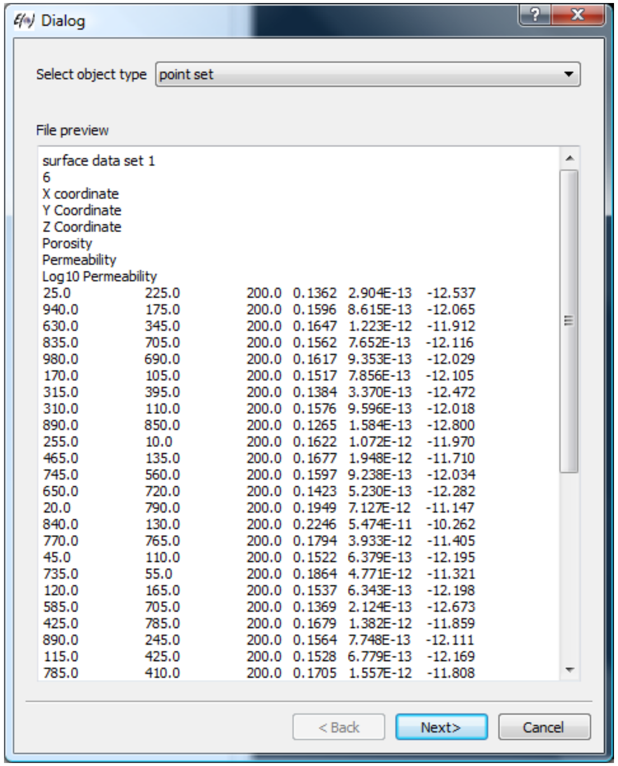
SGeMS Software Introduction

- **The main screen of SGeMS contains the algorithm window, objects window, and the visualization window**



SGeMS Software Introduction

- The first task is to open a data file by selecting “Load Objects” from the “Objects” pulldown menu
- Navigate to the data file, open it, select object type “point set”, go to next screen, enter a Pointset name in the dialog box, and confirm that x, y, and z columns are correctly identified

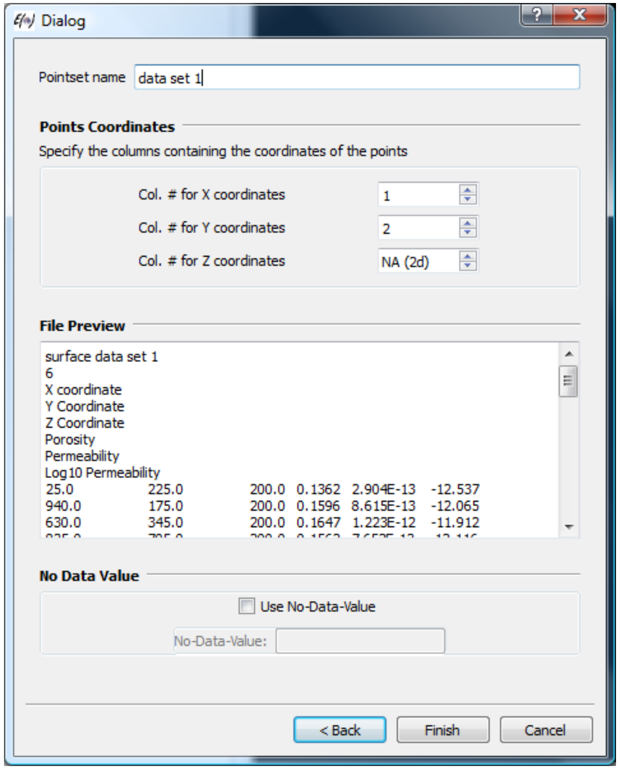


Select object type: point set

File preview

```
surface data set 1
6
X coordinate
Y Coordinate
Z Coordinate
Porosity
Permeability
Log10 Permeability
25.0          225.0         200.0  0.1362  2.904E-13  -12.537
940.0         175.0         200.0  0.1596  8.615E-13  -12.065
630.0         345.0         200.0  0.1647  1.223E-12  -11.912
835.0         705.0         200.0  0.1562  7.652E-13  -12.116
980.0         690.0         200.0  0.1617  9.353E-13  -12.029
170.0         105.0         200.0  0.1517  7.856E-13  -12.105
315.0         395.0         200.0  0.1384  3.370E-13  -12.472
310.0         110.0         200.0  0.1576  9.596E-13  -12.018
890.0         850.0         200.0  0.1265  1.584E-13  -12.800
255.0         10.0          200.0  0.1622  1.072E-12  -11.970
465.0         135.0         200.0  0.1677  1.948E-12  -11.710
745.0         560.0         200.0  0.1597  9.238E-13  -12.034
650.0         720.0         200.0  0.1423  5.230E-13  -12.282
20.0          790.0         200.0  0.1949  7.127E-12  -11.147
840.0         130.0         200.0  0.2246  5.474E-11  -10.262
770.0         765.0         200.0  0.1794  3.933E-12  -11.405
45.0          110.0         200.0  0.1522  6.379E-13  -12.195
735.0         55.0          200.0  0.1864  4.771E-12  -11.321
120.0         165.0         200.0  0.1537  6.343E-13  -12.198
585.0         705.0         200.0  0.1369  2.124E-13  -12.673
425.0         785.0         200.0  0.1679  1.382E-12  -11.859
890.0         245.0         200.0  0.1564  7.748E-13  -12.111
115.0         425.0         200.0  0.1528  6.779E-13  -12.169
785.0         410.0         200.0  0.1705  1.557E-12  -11.808
```

→



Pointset name: data set 1

Points Coordinates

Specify the columns containing the coordinates of the points

Col. # for X coordinates: 1

Col. # for Y coordinates: 2

Col. # for Z coordinates: NA (2d)

File Preview

```
surface data set 1
6
X coordinate
Y Coordinate
Z Coordinate
Porosity
Permeability
Log10 Permeability
25.0          225.0         200.0  0.1362  2.904E-13  -12.537
940.0         175.0         200.0  0.1596  8.615E-13  -12.065
630.0         345.0         200.0  0.1647  1.223E-12  -11.912
```

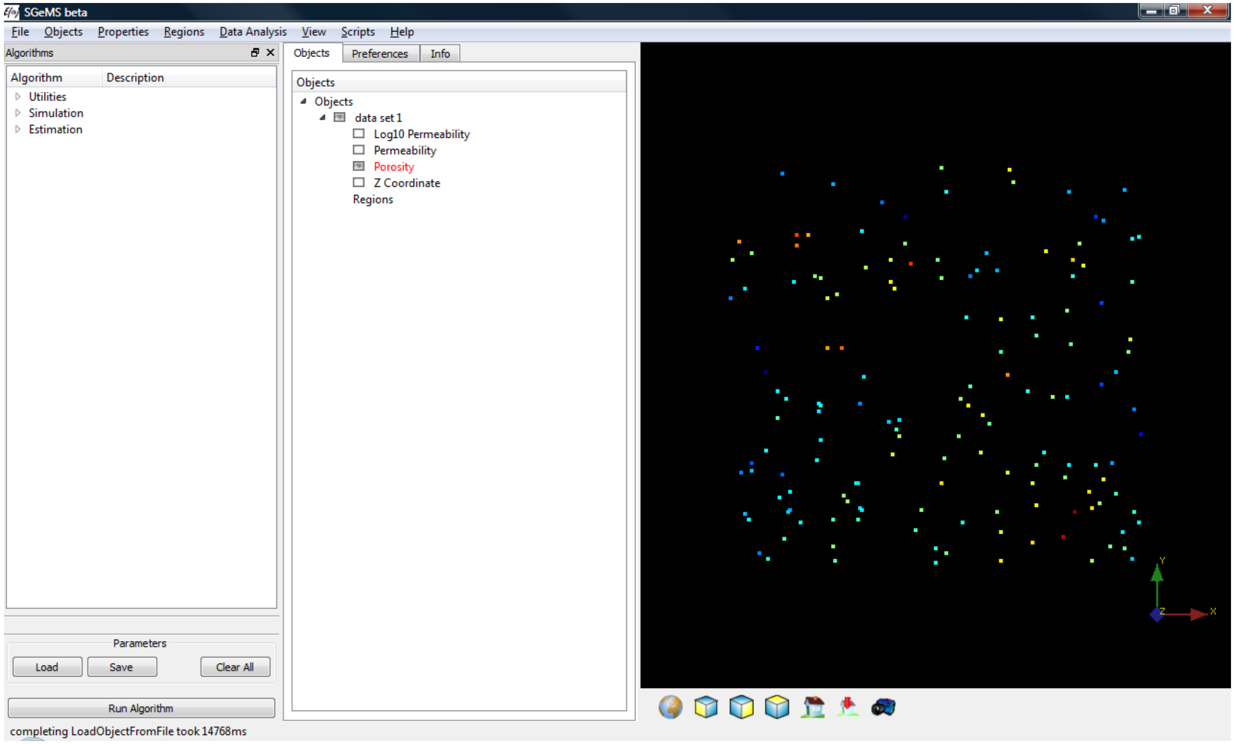
No Data Value

☐ Use No-Data-Value

No-Data-Value:

SGeMS Software Introduction

- Now that the data set is loaded, various actions can be taken, including data visualization (shown below), data analysis tasks, and algorithms.
- Practice manipulating the visualization, adding a colorbar, setting the color scale, and exporting an image of the plot (use the camera icon)
- Note that SGeMS does not manage screen real estate very well and windows may need to be resized to show what you want to see

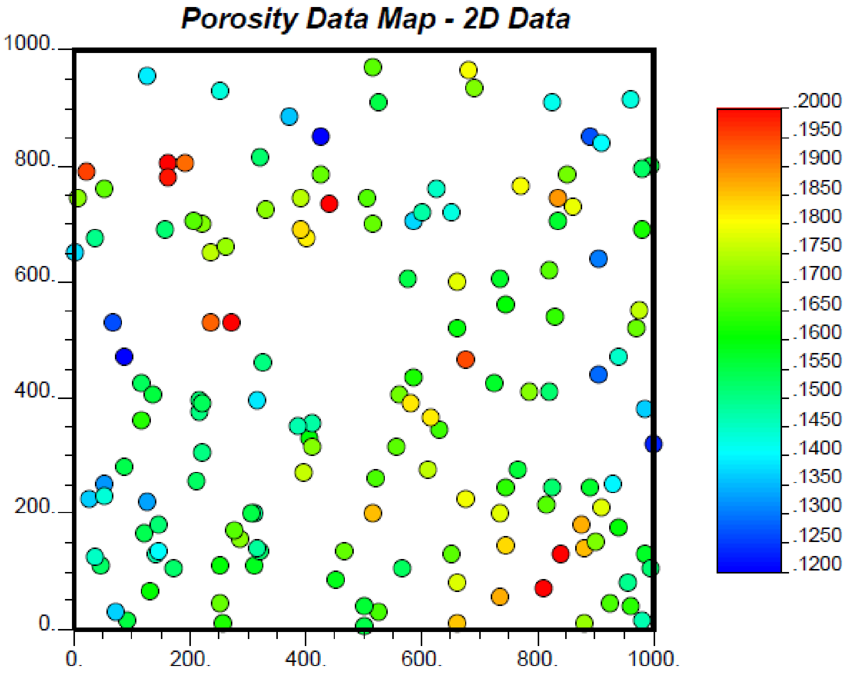


completing LoadObjectFromFile took 14768ms

GSLIB Software - LOCMAP

- Use the LOCMAP DOS program to generate a 2D plot of the sample data
- The LOCMAP program is executed with a parameter control file named locmap.par

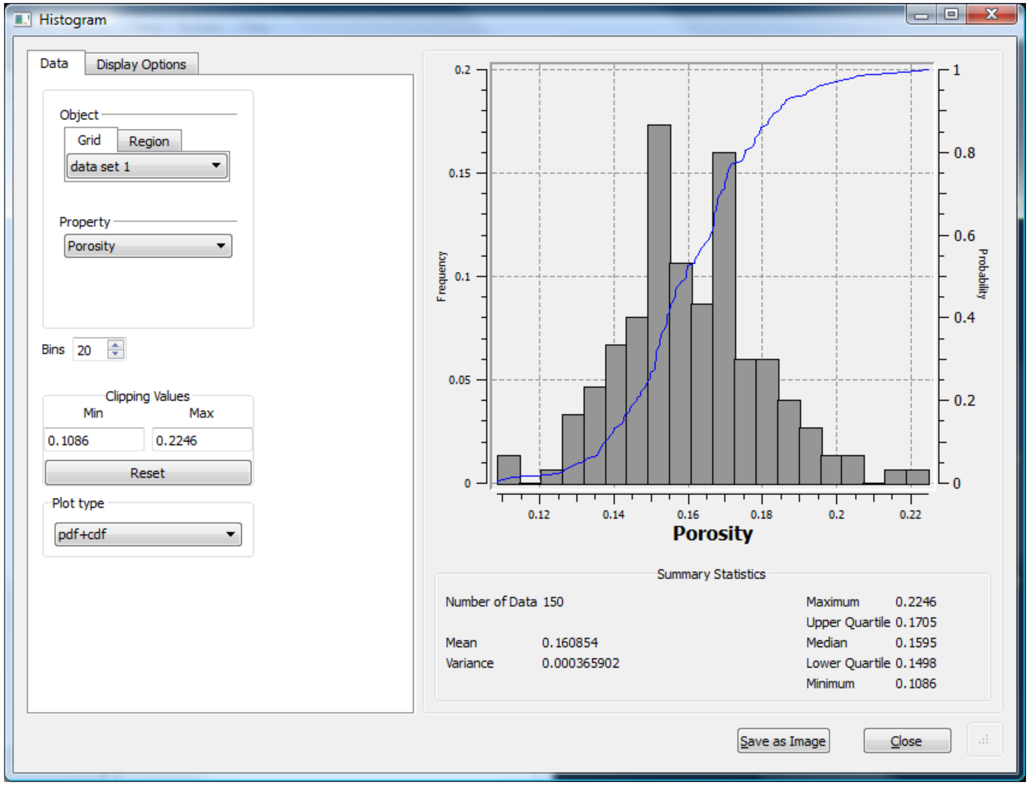
```
          Parameters for LOCMAP
             *********************
START OF PARAMETERS:
data_set_1_2D.dat          \file with data
1   2   4                  \   columns for X, Y, variable
-1.0e21   1.0e21           \   trimming limits
locmap.ps                  \file for PostScript output
0.0     1000.              \xmn,xmx
0.0     1000.              \ymn,ymx
0                          \0=data values, 1=cross validation
0                          \0=arithmetic,  1=log scaling
1                          \0=gray scale,  1=color scale
0                          \0=no labels,   1=label each location
0.12   0.20    0.005       \gray/color scale: min, max, increm
0.5                        \label size: 0.1(sml)-1(reg)-10(big)
Porosity Data Map - 2D Data          \Title
```



- Plot other parameter and data files using the LOCMAP program

SGeMS Software Introduction

- Use the “Data Analysis” dropdown menu to choose the “Histogram” option to examine the distributions of parameters in the data set
- Choose the data set you want to examine from the “Grid” dropdown menu
- Vary the number of bins in the plot, plot the various parameters, change the axis limits and type, and practice exporting plot images



Grid: data set 1

Property: Porosity

Bins: 20

Clipping Values

Min: 0.1086

Max: 0.2246

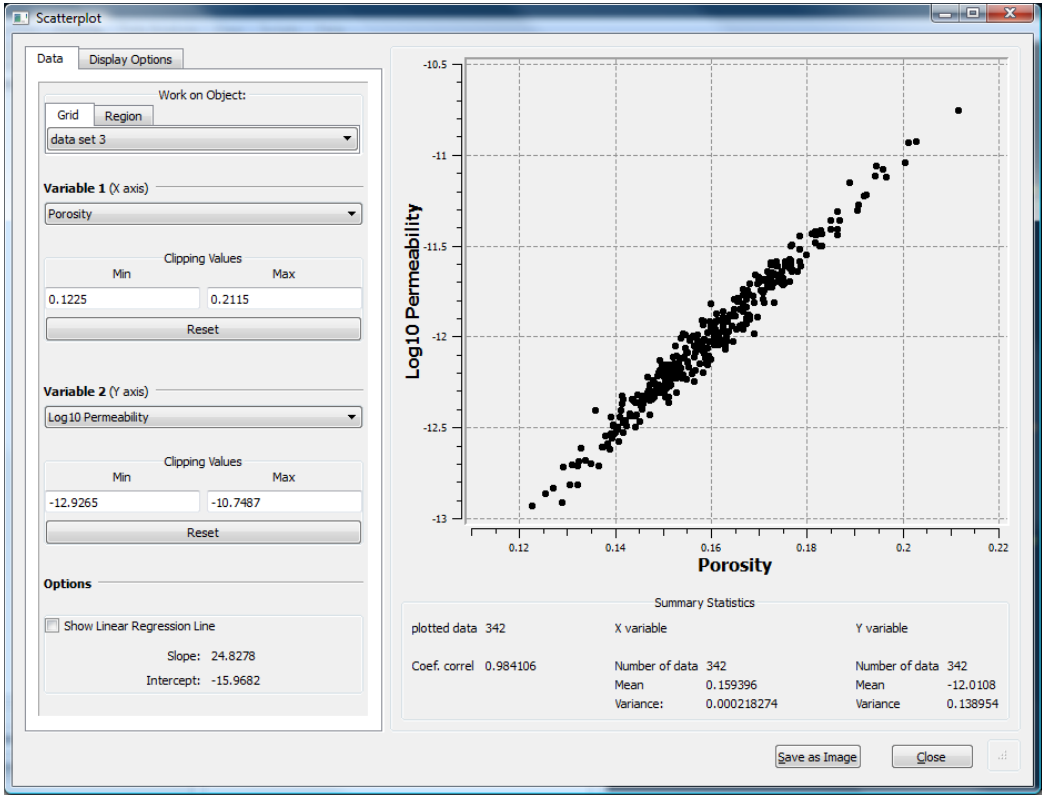
Plot type: pdf+cdf

Summary Statistics

Number of Data	150	Maximum	0.2246
		Upper Quartile	0.1705
Mean	0.160854	Median	0.1595
Variance	0.000365902	Lower Quartile	0.1498
		Minimum	0.1086

SGeMS – Exploratory Data Analysis

- Use the "Data Analysis" dropdown menu to choose the "Scatter-plot" option to examine the correlation of parameters in the data set
- Choose the data set you want to examine from the "Grid" dropdown menu
- Plot the various parameters, change the axis limits and type, and practice exporting plot images



Grid: data set 3

Variable 1 (X axis): Porosity

Clipping Values: Min 0.1225, Max 0.2115

Variable 2 (Y axis): Log10 Permeability

Clipping Values: Min -12.9265, Max -10.7487

☐ Show Linear Regression Line

Slope: 24.8278

Intercept: -15.9682

plotted data 342

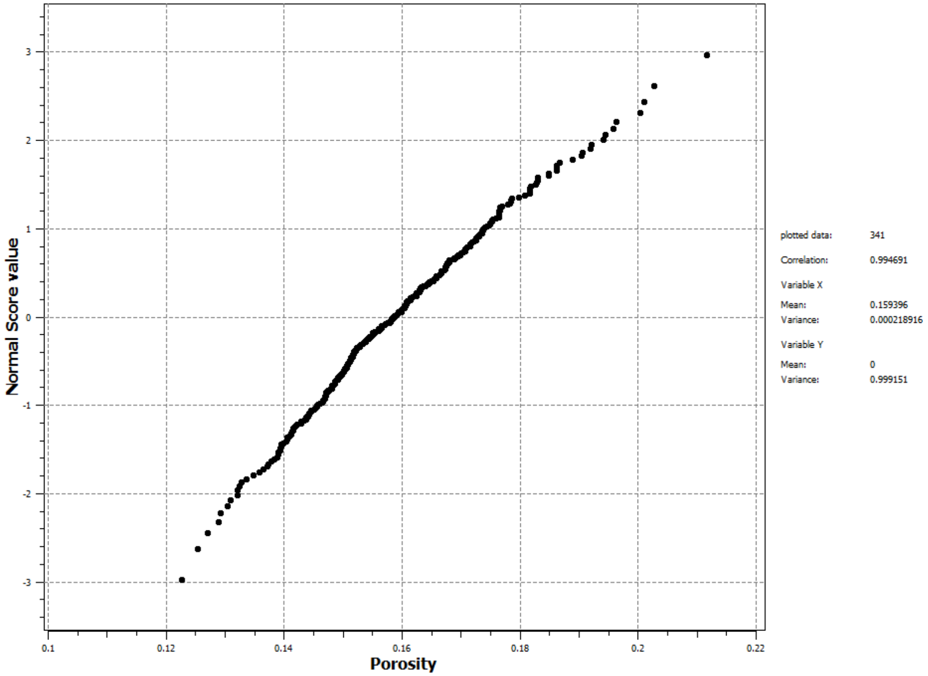
Coef. correl 0.984106

X variable		Y variable	
Number of data	342	Number of data	342
Mean	0.159396	Mean	-12.0108
Variance:	0.000218274	Variance	0.138954

GSLIB Software - NSCORE

- **Use the NSCORE DOS program to calculate the normal score transform of one parameter in the data set**
- **The NSCORE program is executed with a parameter control file named nscore.par**

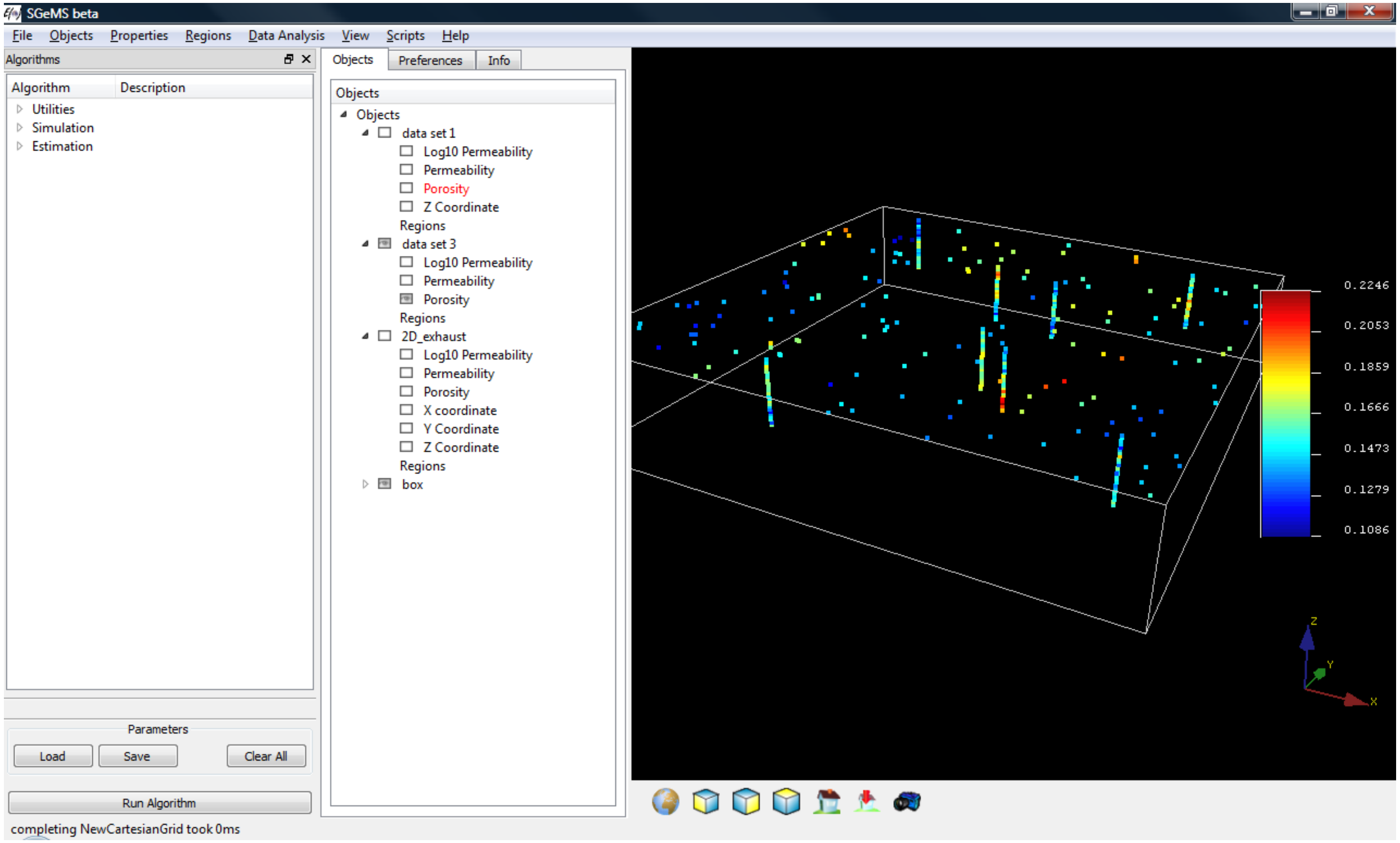
```
             Parameters for NSCORE
              *********************

START OF PARAMETERS:
data_set_3_3D.dat      \file with data
4   0                   \  columns for variable and weight
-1.0e21   1.0e21        \  trimming limits
0                       \1=transform according to specified ref. dist.
unknown.out             \  file with reference dist.
1   2                   \  columns for variable and weight
nscore.out              \file for output
nscore.trn              \file for output transformation table
```



- **The NSCORE program writes the normal score transform values to a new column**
- **Make a probability plot with the scatterplot option and check for a Gaussian distribution**

SGeMS – Exploratory Data Analysis

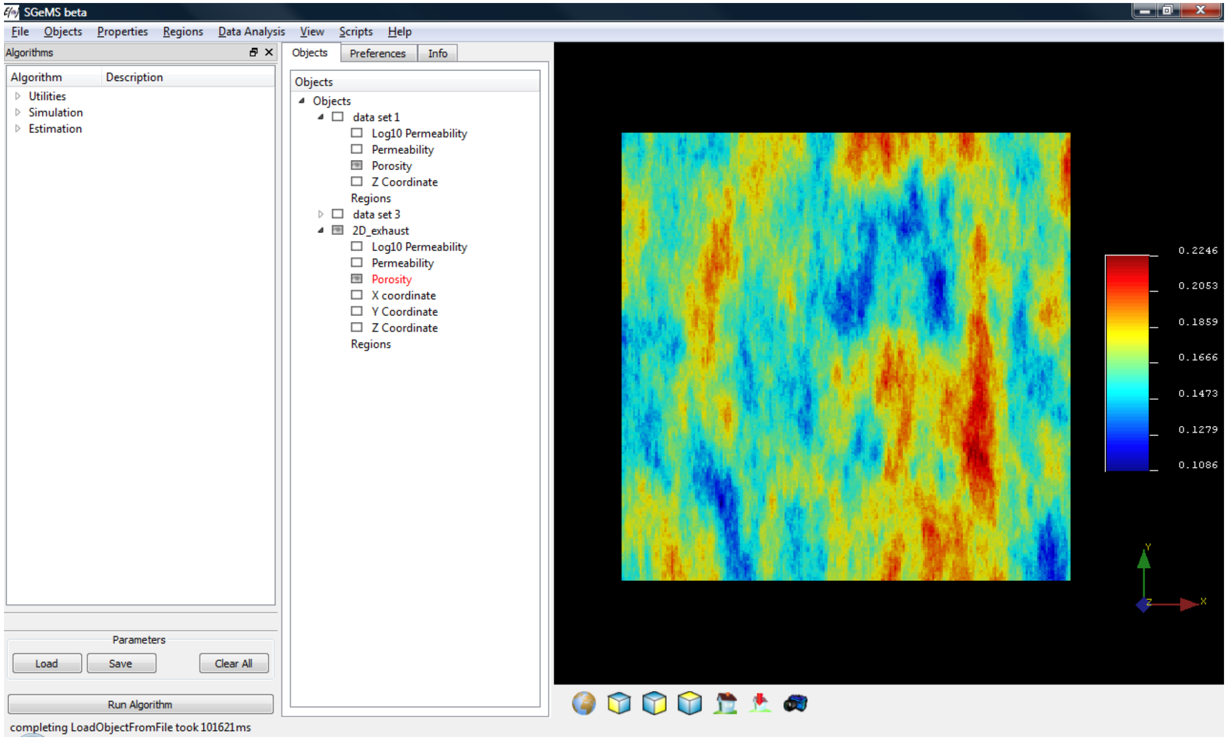
- Now load the data set contained in the file “data_set_3_3D.dat”.
- This is a 3D point data set that contains parameter values at many surface locations and in 8 boreholes
- Practice manipulating the visualization image and plot options



completing NewCartesianGrid took 0ms

SGeMS – Exploratory Data Analysis

- Now load the data set contained in the file “data_set_2_2D_exhaustive.dat”.
- This is an exhaustive 2D data set that must be loaded in as an object using the “cartesian grid” option. By looking at the data file you can see that it consists of parameter values on a regular 201 X 201 grid, which must be entered to load the file.
- Practice changing the various options presented under the “Preferences” tab to control the plot

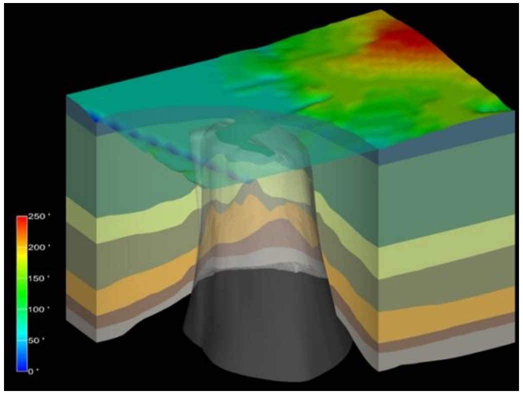


completing LoadObjectFromFile took 101621ms

SAND2013-xxxx

Exceptional service in the national interest

Sandia National Laboratories



Exercise 2: Spatial Correlation Analysis

IAEA Training Course – July 1-5, 2013

Bill Arnold

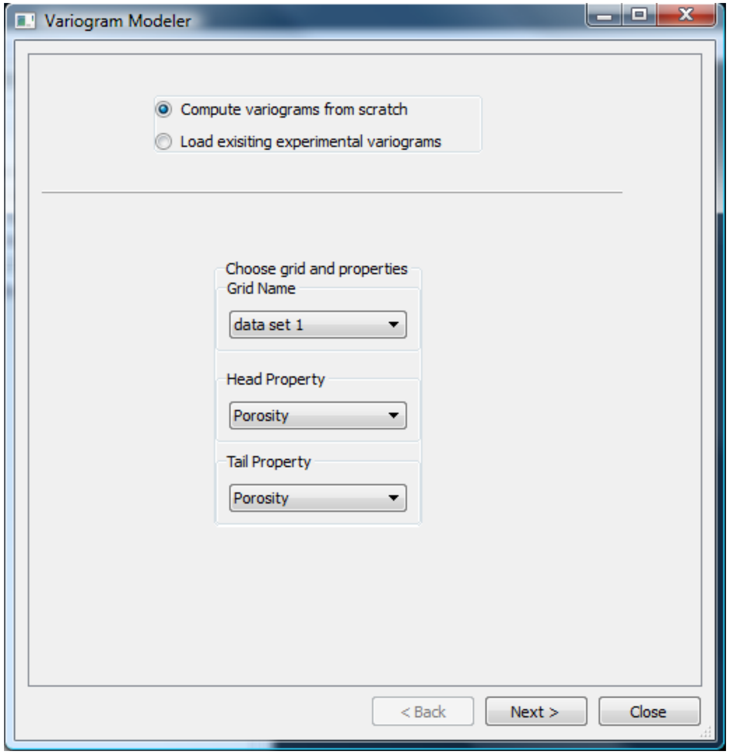
U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

IAEA International Atomic Energy Agency

SGeMS – Variogram Analysis

- Load the data sets that were used in the exploratory data analysis exercise.
- Start the variogram analysis by choosing the “Variogram” option under the “Data Analysis” pulldown menu.
- Choose the data set to analyze from the “Grid Name” pulldown menu and choose the parameter of interest from both the “Head Property” and “Tail Property” menus.



Variogram Modeler

● Compute variograms from scratch
○ Load exisiting experimental variograms

Choose grid and properties

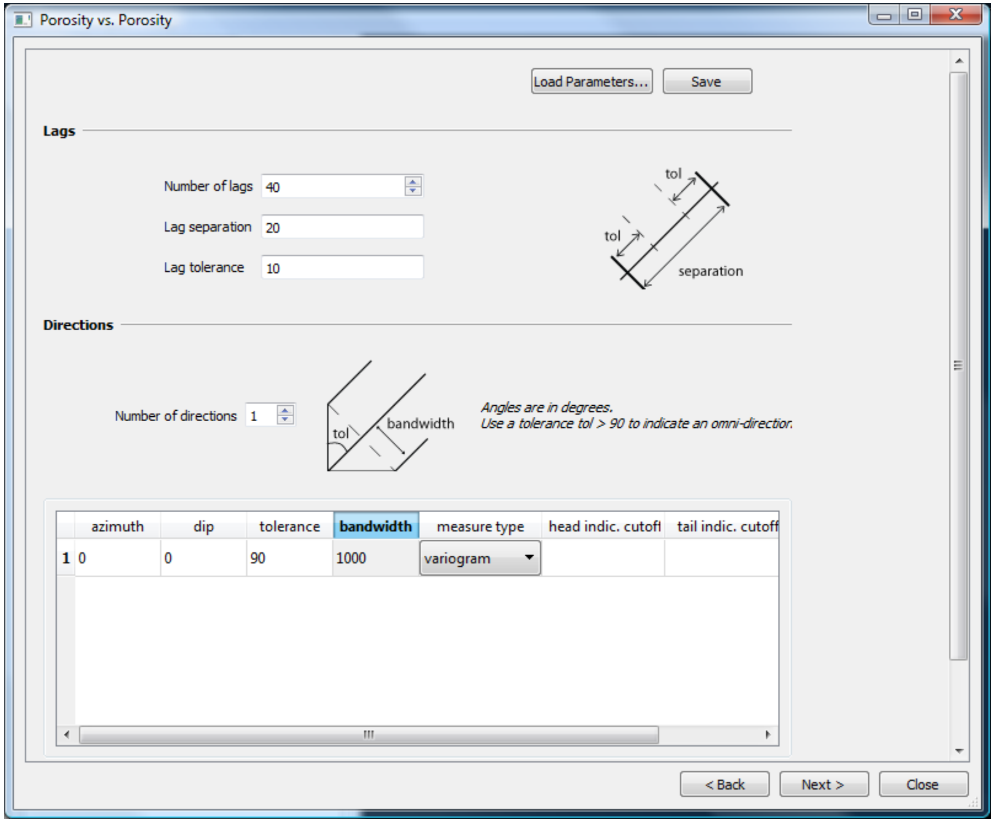
Grid Name: data set 1

Head Property: Porosity

Tail Property: Porosity

SGeMS – Variogram Analysis

- Enter the number of lags, lag separation, lag tolerance, azimuth, dip, directional tolerance, and bandwidth to control the construction of the experimental variogram.
- Note that multiple variogram plots can be generated simulataneously by increasing “Number of directions”.
- Search direction is defined by azimuth and dip, as explained in the figures to the right.



Lags

Number of lags: 40

Lag separation: 20

Lag tolerance: 10

Directions

Number of directions: 1

Angles are in degrees.
Use a tolerance tol > 90 to indicate an omni-direction

	azimuth	dip	tolerance	bandwidth	measure type	head indic. cutoff	tail indic. cutoff
1	0	0	90	1000	variogram		

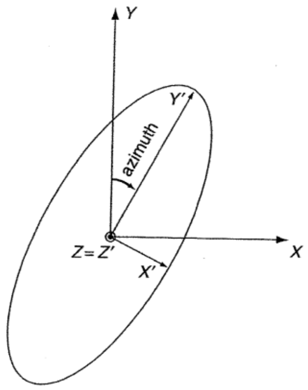


Figure 2.24 First rotation about Z'

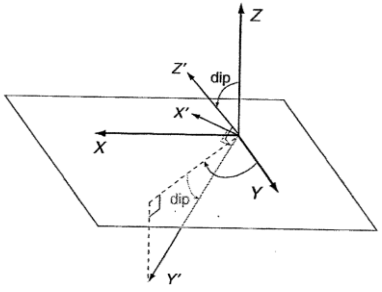
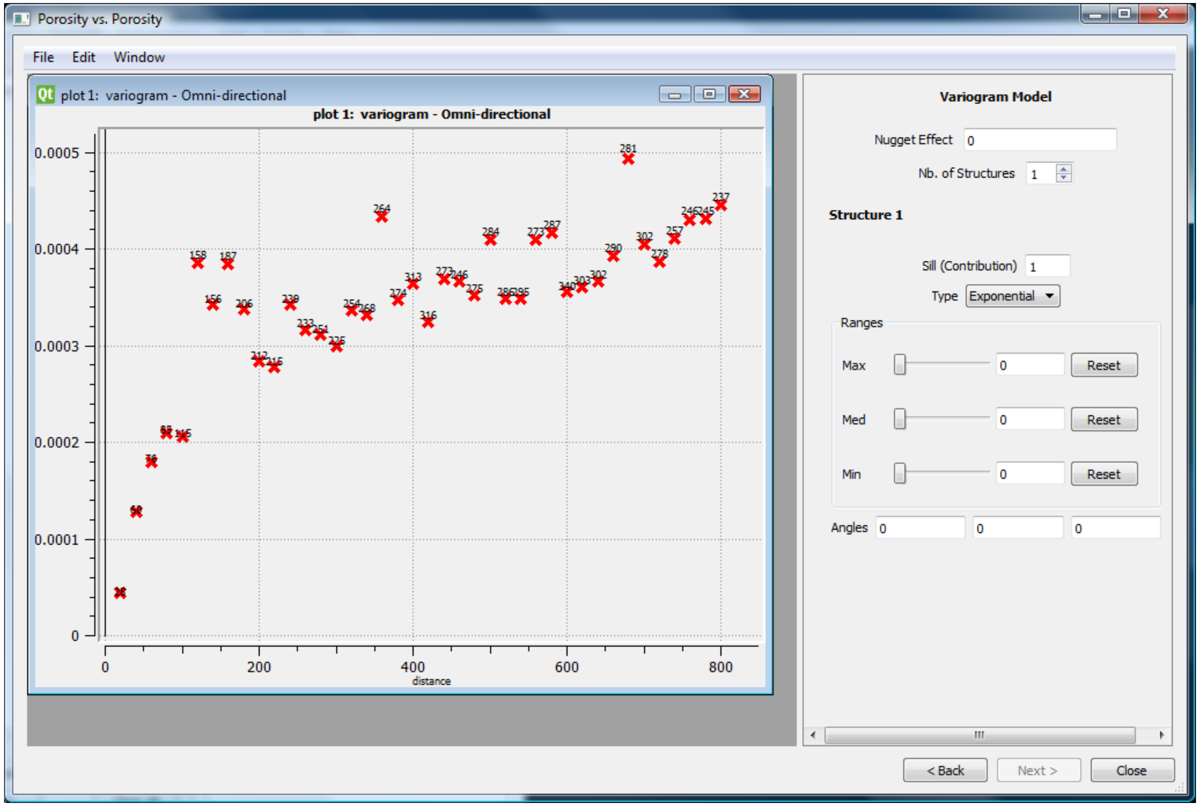


Figure 2.25 Second rotation about X'

SGeMS – Variogram Analysis

- **The experimental variogram is displayed in the next window.**
- **The number of data pairs that were used to calculate each plotted value are displayed by right clicking the mouse.**



Variogram Model

Nugget Effect 0

Nb. of Structures 1

Structure 1

Sill (Contribution) 1

Type Exponential

Ranges

Max 0

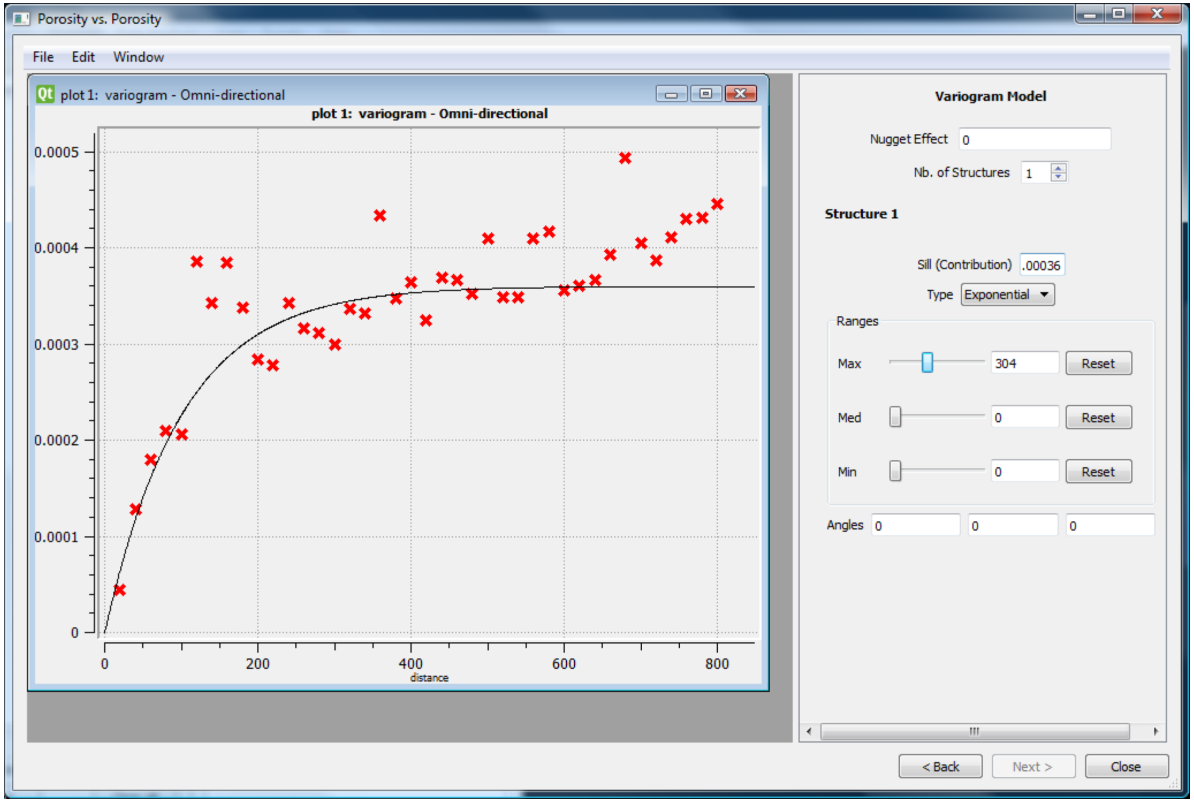
Med 0

Min 0

Angles 0 0 0

SGeMS – Variogram Analysis

- Fit the experimental variogram with a model created with the parameters entered in the panel to the right of the plot.
- Try different variogram types and increasing the number of structures to create variograms using linear combinations of models.



Variogram Model

Nugget Effect 0

Nb. of Structures 1

Structure 1

Sill (Contribution) .00036

Type Exponential

Ranges

Max 304

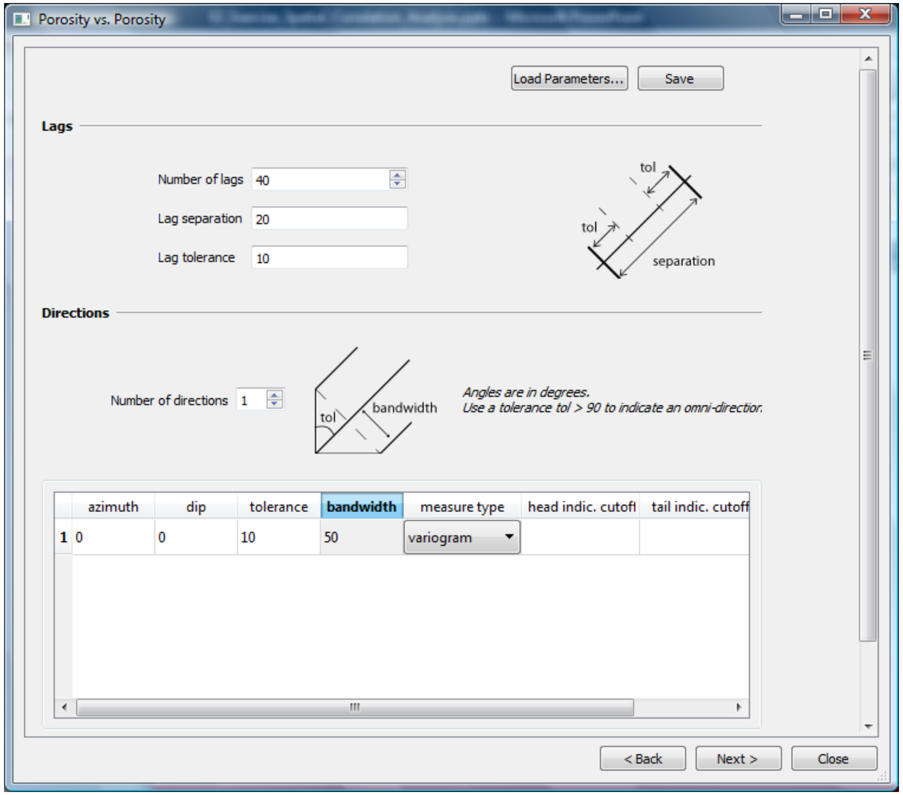
Med 0

Min 0

Angles 0 0 0

SGeMS – Variogram Analysis

- Save or write down the parameters for the variogram model that you have fit to the experimental variogram for later use in the estimation and simulation exercises.
- Load the other data sets (2-D data set 2, 2-D data set 2–exhaustive, and 3-D data set 3) and examine directional variograms to examine anisotropy in the horizontal and vertical directions.



Porosity vs. Porosity

Lags

Number of lags: 40
Lag separation: 20
Lag tolerance: 10

Directions

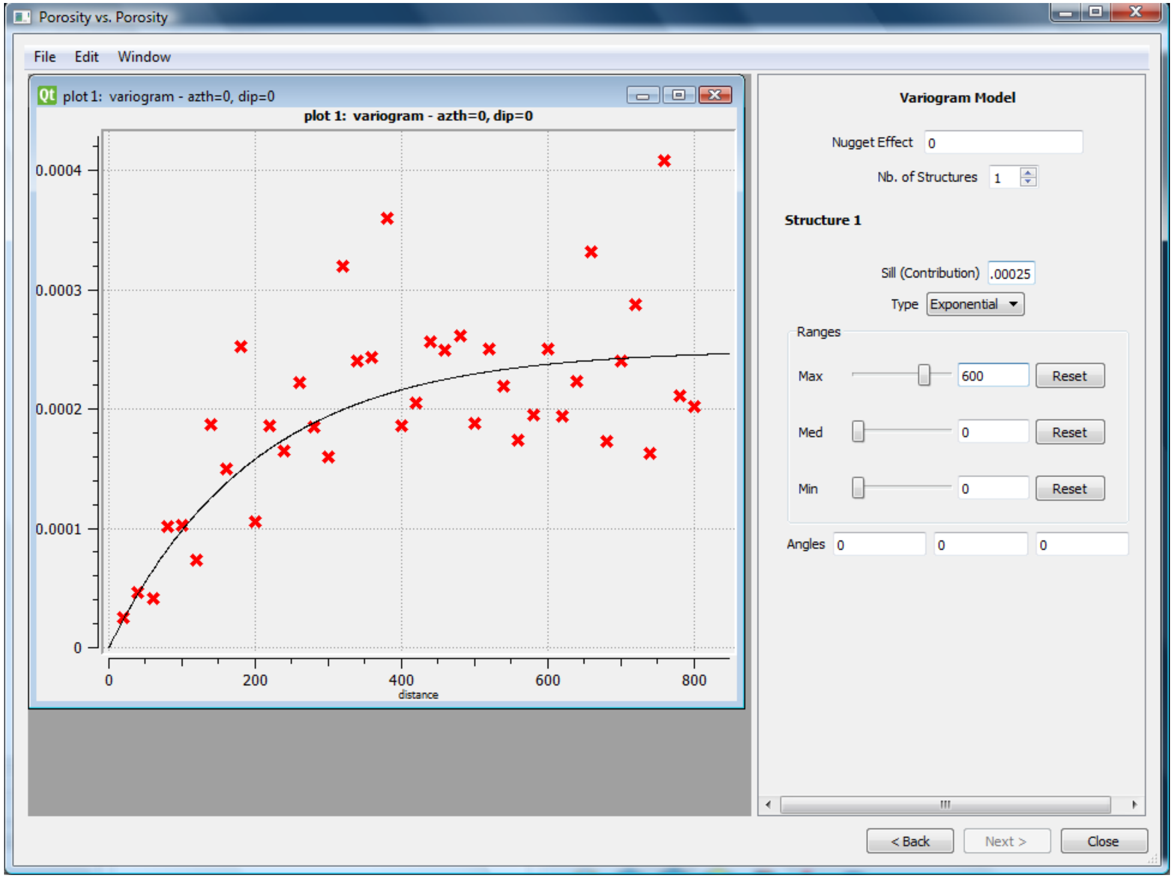
Number of directions: 1

Angles are in degrees.
Use a tolerance tol > 90 to indicate an omni-direction

	azimuth	dip	tolerance	bandwidth	measure type	head indic. cutoff	tail indic. cutoff
1	0	0	10	50	variogram		

SGeMS – Variogram Analysis

- **Fit a model to the experimental variograms created in different directions to evaluate the anisotropy in the spatial correlation.**
- **Save or write down the parameters for the variogram models that you have fit to the directional variograms for later use in the estimation and simulation exercises.**



Porosity vs. Porosity

plot 1: variogram - azth=0, dip=0

Variogram Model

Nugget Effect 0

Nb. of Structures 1

Structure 1

Sill (Contribution) .00025

Type Exponential

Ranges

Max 600

Med 0

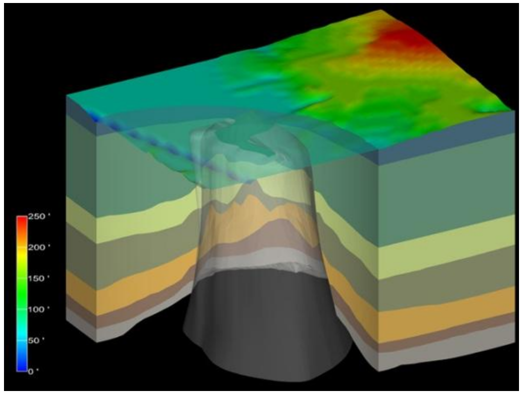
Min 0

Angles 0 0 0

SAND2013-xxxx

Exceptional service in the national interest

Sandia National Laboratories



Spatial Estimation and Simulation Methods

IAEA Training Course – July 1-5, 2013

Bill Arnold

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

IAEA International Atomic Energy Agency

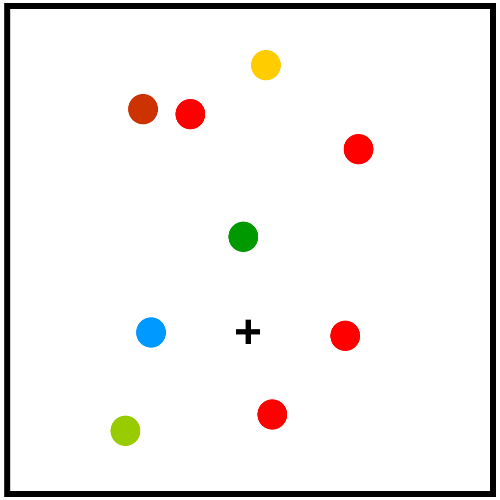
Outline

- Spatial estimation objectives
- Interpolation methods of estimation
- Evaluation of estimation methods
- Kriging
- Spatial simulation versus estimation
- Sequential simulation methods

Point Estimation

Point estimation is used to estimate the value of a property (porosity, permeability, fracture frequency, etc.) at some point in the ground or in a space based on linear combinations of the surrounding data.

Kriging is a form of estimation (interpolation algorithm).
The kriging algorithm is also the basis of the simulation techniques used in geostatistics.



Estimation Techniques

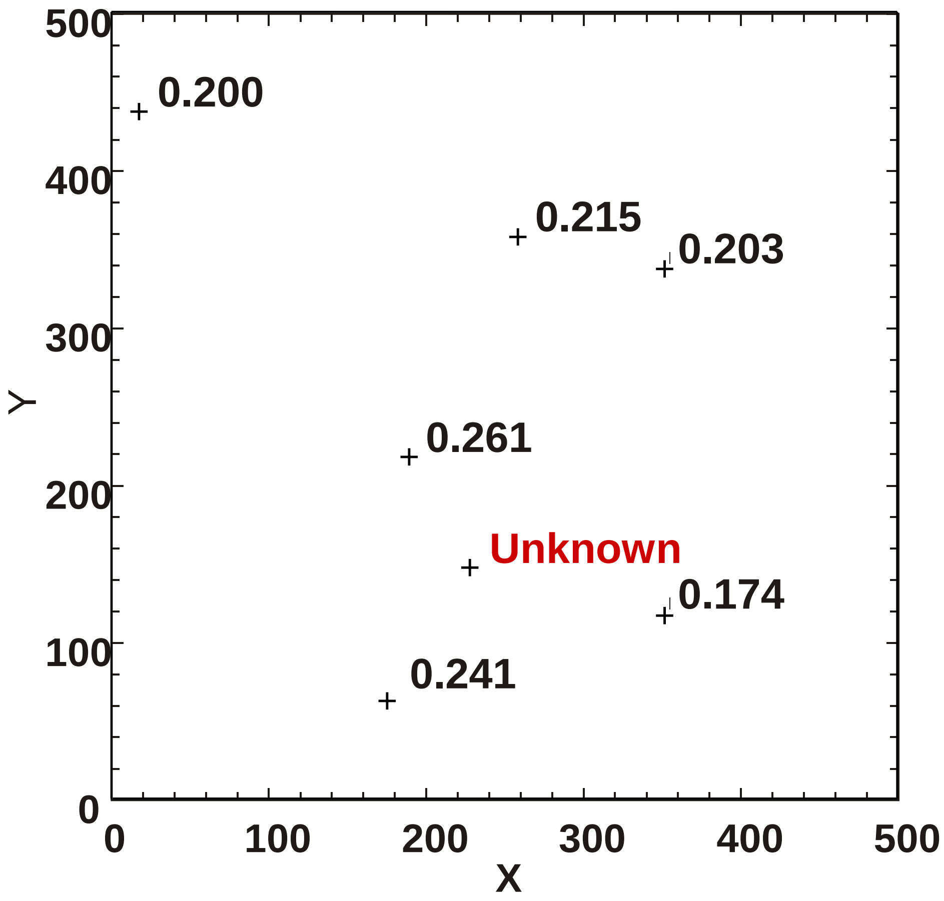
Example data-driven point estimation techniques (require some data to exist already) that interpolate to the surrounding locations:

- **Nearest Neighbor Polygons**

  (aka Theissen or Voronoi polygons)
- **Local Mean** **(using surrounding data)**
- **Inverse Distance Squared**

We will look at the three techniques above, as well as kriging. There are other techniques such as: trend surfaces (polynomial fit) and splines.

Estimation Example



- Porosity measured at 6 points.
- Make estimate of porosity at unknown point, x_0.

Nearest Neighbor Polygons

- Construct polygons around the samples that divide the space into regions
- Everywhere inside of the polygon is closer to the sample point enclosed by that polygon than to any other sample point

Advantages:

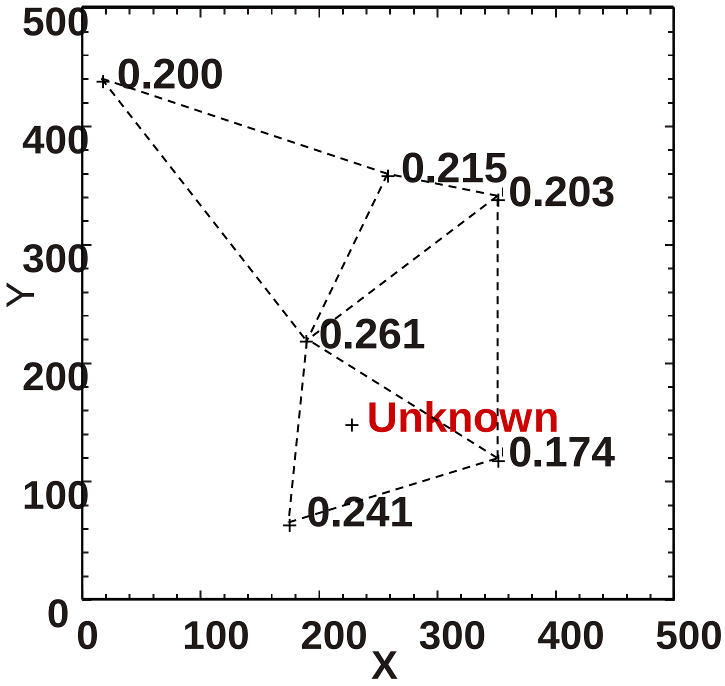
- Simple, fast, exact interpolator (at a point where the value is known, it returns that exact value)

Disadvantages:

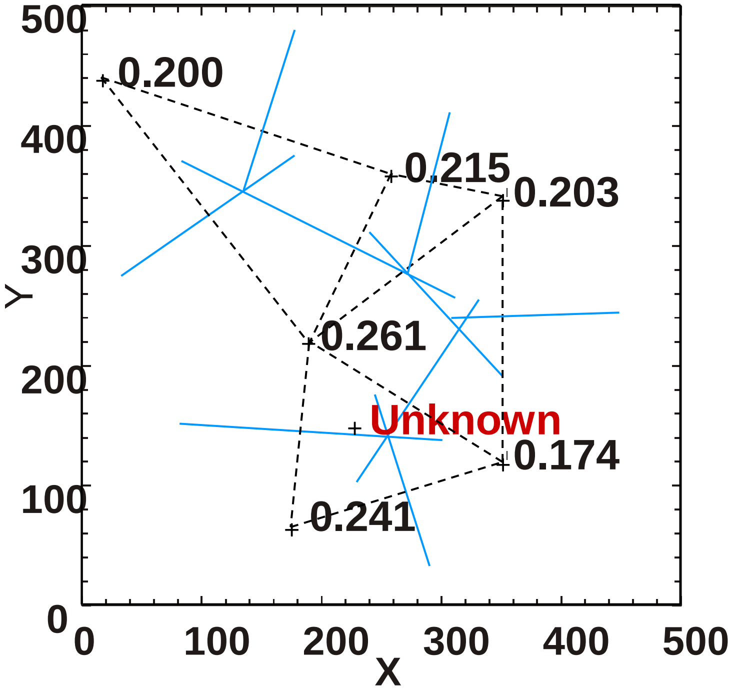
- Discontinuities at polygon boundaries
- If data are sparse and somewhat unevenly spaced, the global estimation is dominated by the sparsely located points

Nearest Neighbor Polygons

- Connect each sample point to the neighboring sample points to create a series of triangles



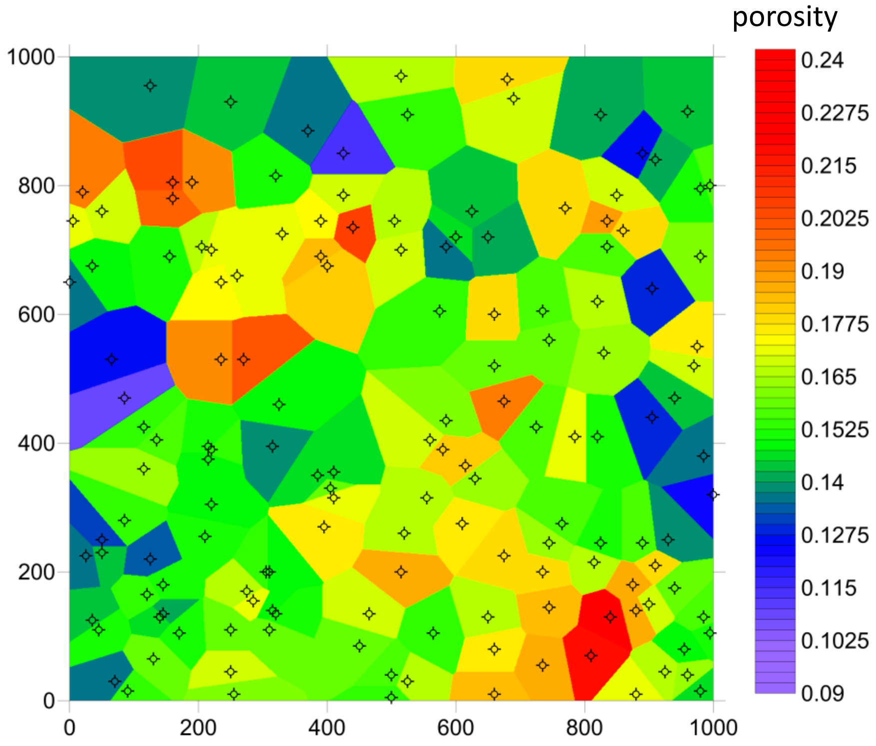
- Draw a perpendicular bisector through each line.



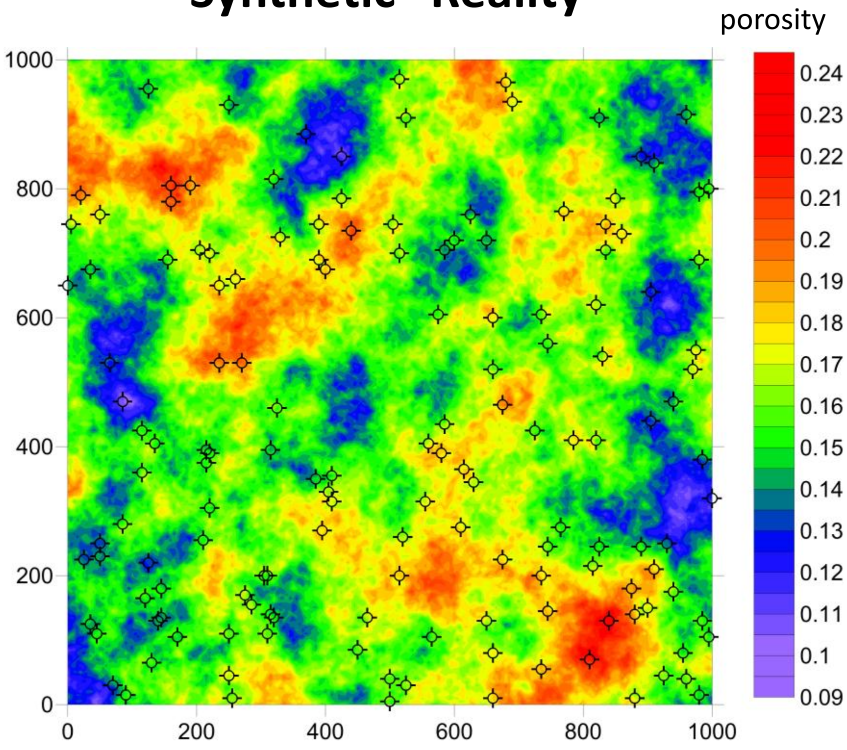
Estimated value at Unknown = .261

Nearest Neighbor Polygons

Example from workshop exercise data set



Synthetic “Reality”



Local Mean Estimation

- Use the mean of surrounding data as an estimate of the value at target location

Advantages:

- Simple, fast, few large errors (near the edges of the domain)

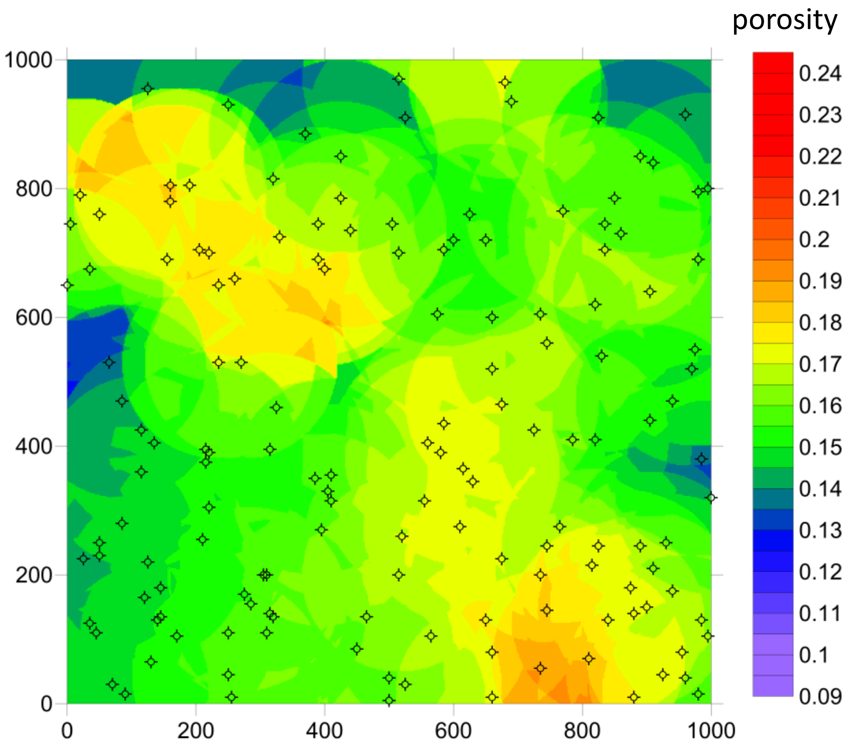
Disadvantages:

- Not an exact interpolator (the average of the few surrounding data points won’t necessarily return the exact value for a known point)
- Definition of "surrounding data“ may be difficult?
- It has a smoothing effect on the data values. Any extreme values, high or low, will get smoothed out as they are averaged in with the surrounding values

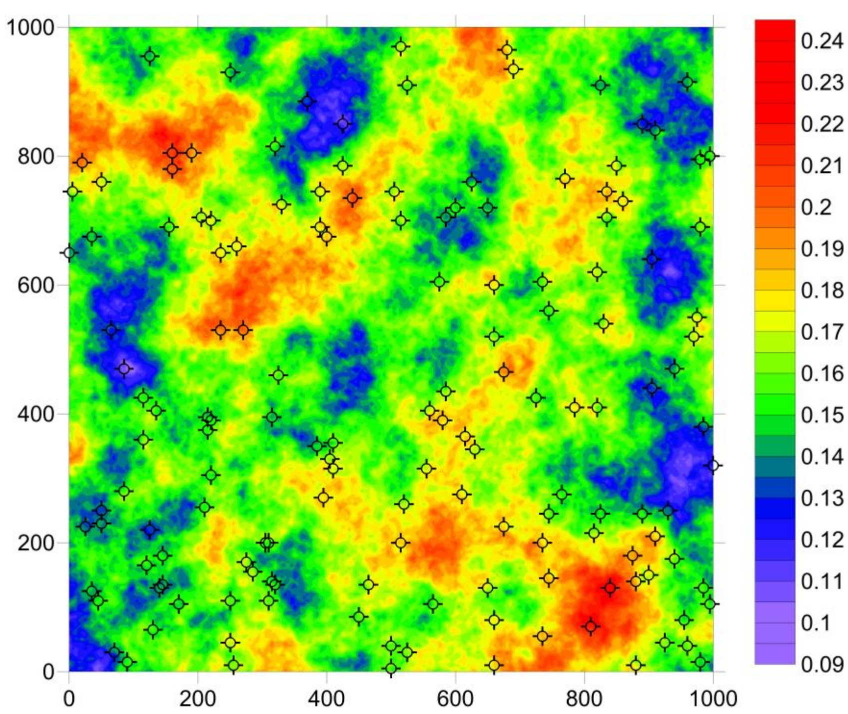
For the example shown: mean = 1/n*SUM(data) = 0.216

Local Mean Estimation

Example from workshop exercise data set

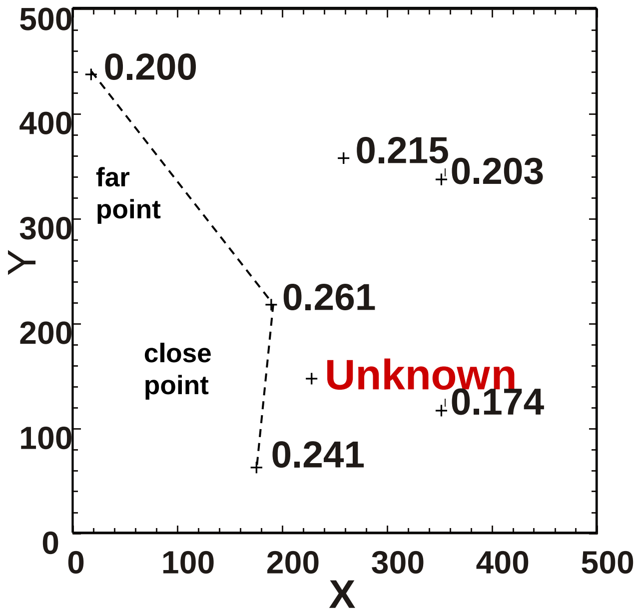


Synthetic "Reality"



Local Mean Estimation

- Should each surrounding point be weighted evenly?
- Do closer points have a greater influence on the estimate than data that are farther away?
- Should the more distant points be included in the average? If so, should they be given less weight?



Next we will examine a technique that weights the data relative to their distance away from the point being estimated.

Inverse Distance Estimation

- Create weights for the data values that are inversely proportional to the distance from the unknown location,
- The weighting function is the inverse distance raised to a power, ω

$$est = \frac{\sum_{i=1}^{n}\left[\left(\frac{1}{d_i^{\omega}}\right)z_i\right]}{\sum_{i=1}^{n}\frac{1}{d_i^{\omega}}}$$

Advantages:

- Simple, fast and includes distance in calculation of weights

Disadvantages:

- Not an exact interpolator
- As d goes to 0, the estimator “blows up”.

Where:

d_i is the distance

z_i is the sample value

Inverse Distance Estimation

Inverse Distance Squared ($\omega = 2$)

Estimation of Value at (235.0, 155.0)

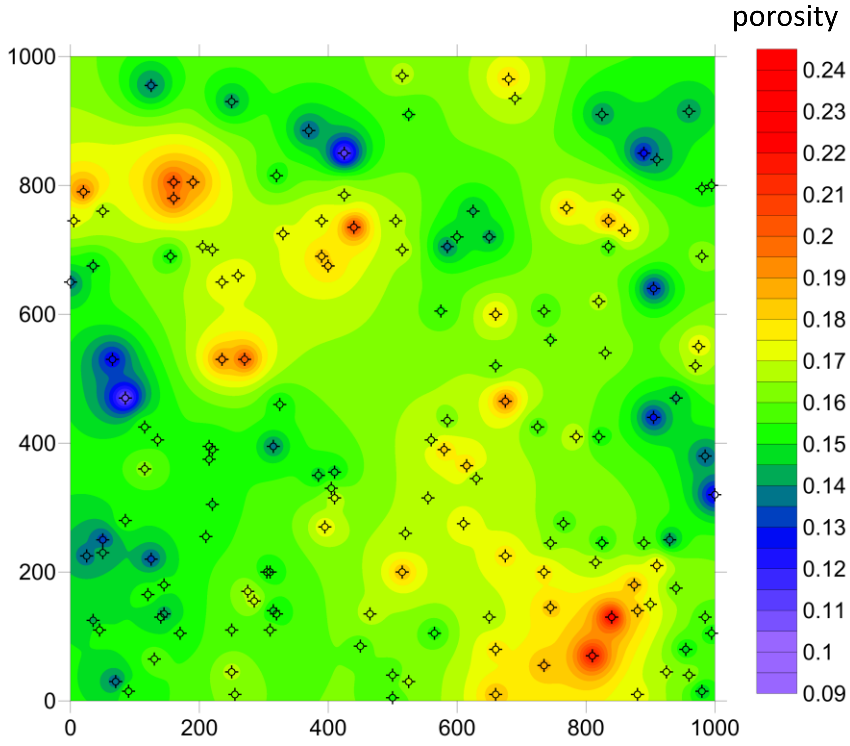
Sample #	X	Y	Value	Distance from X_0, Y_0	Normalized Weight	Weighted Value
1	195	225	0.261	80.62	0.403	0.105
2	355	225	0.174	123.69	0.172	0.03
3	355	345	0.203	224.72	0.052	0.01
4	265	365	0.215	212.13	0.058	0.013
5	185	75	0.241	94.34	0.295	0.071
6	25	445	0.2	358.05	0.02	0.004

The final estimate is the sum of weighted values.

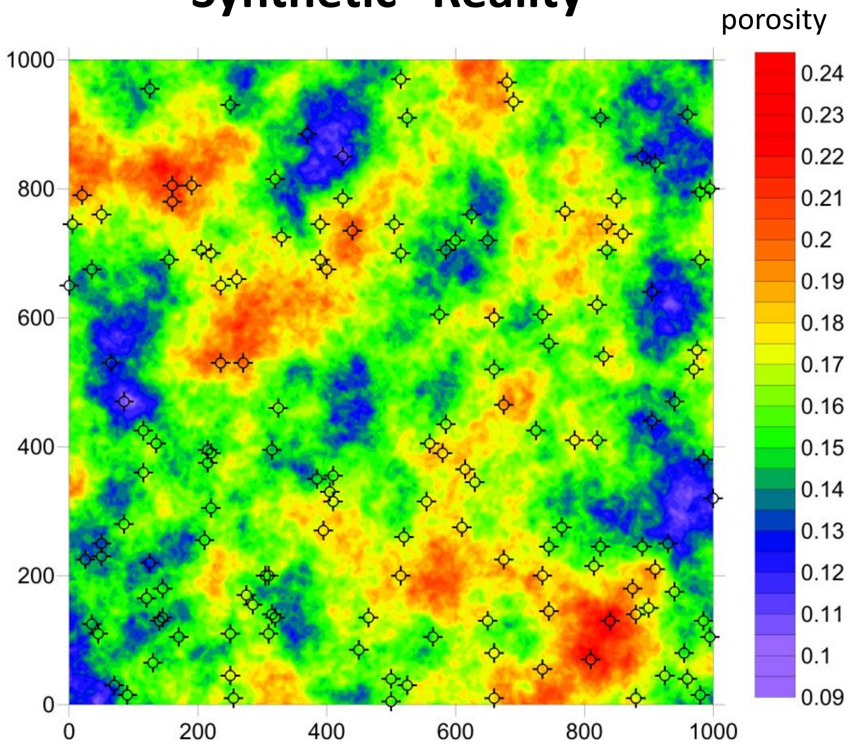
Estimate of value at Unknown = 0.233

Inverse Distance Estimation

Example from workshop exercise data set



Synthetic “Reality”



Evaluating Estimation Methods

What attributes/statistics could be used to determine whether or not a technique is worthwhile?

- **Estimate a large number of points (100's-1000's) and then take a sample at each location. Look at how well the estimates and actual values compare.**

  (Practically possible with a subset of a large data set)

- **Look at the mean error as a measure of bias across all of those locations. Want as many over-estimates as under-estimates so that the mean = 0 in terms of error.**

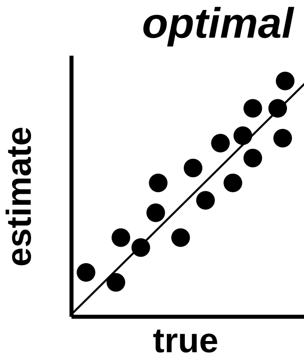
- **Look at the spread, or variance, of the errors. Want it to be minimal.**

Evaluating Estimation Methods

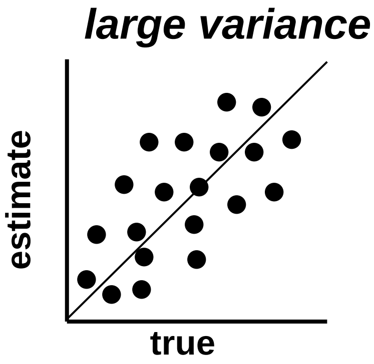
Cross-validation:	Pull each datum out of the model individually and use the surrounding data to re-estimate the removed datum. Then compare the estimate to the actual value.
Jackknifing:	Hold back some of original data and use the remaining data to estimate those locations. Then compare the real values with the estimates.
Both techniques:	Examine a scatterplot of the actual values vs. estimates. Map the residuals to make sure they are not always over estimated in one region and underestimated in another.

Evaluating Estimation Methods

- Build a model of the concentration at each point, the estimate.
- Take a sample and see how well the estimate and true values correlate.



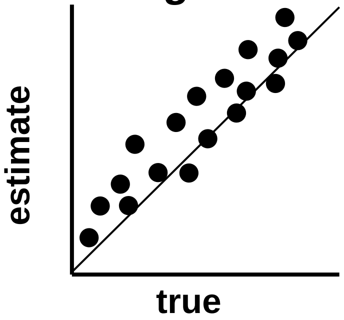
Optimal: The distribution is centered on the 45 degree line (accurate, unbiased) with a small spread.



Large Variance: The distribution is accurate and unbiased, but the estimates are more variable causing a wider spread in the distribution (imprecise).

Evaluating Estimation Methods

high bias

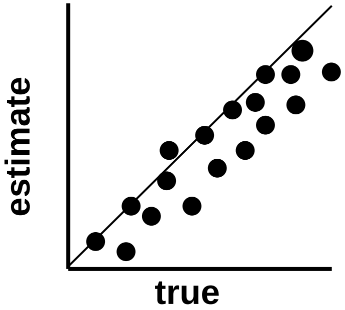


The distribution is precise but biased.

High bias: overestimates the true value

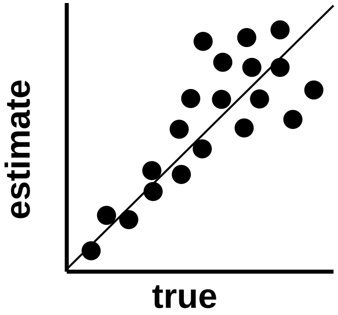
Low bias: underestimates the true value

low bias



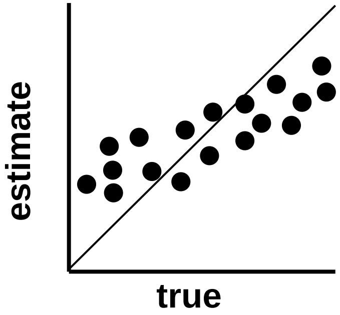
Heteroscedastic: The variance changes as a function of the value. So as the values increase, quality of the fit about the 45 degree line deteriorates.

heteroscedastic

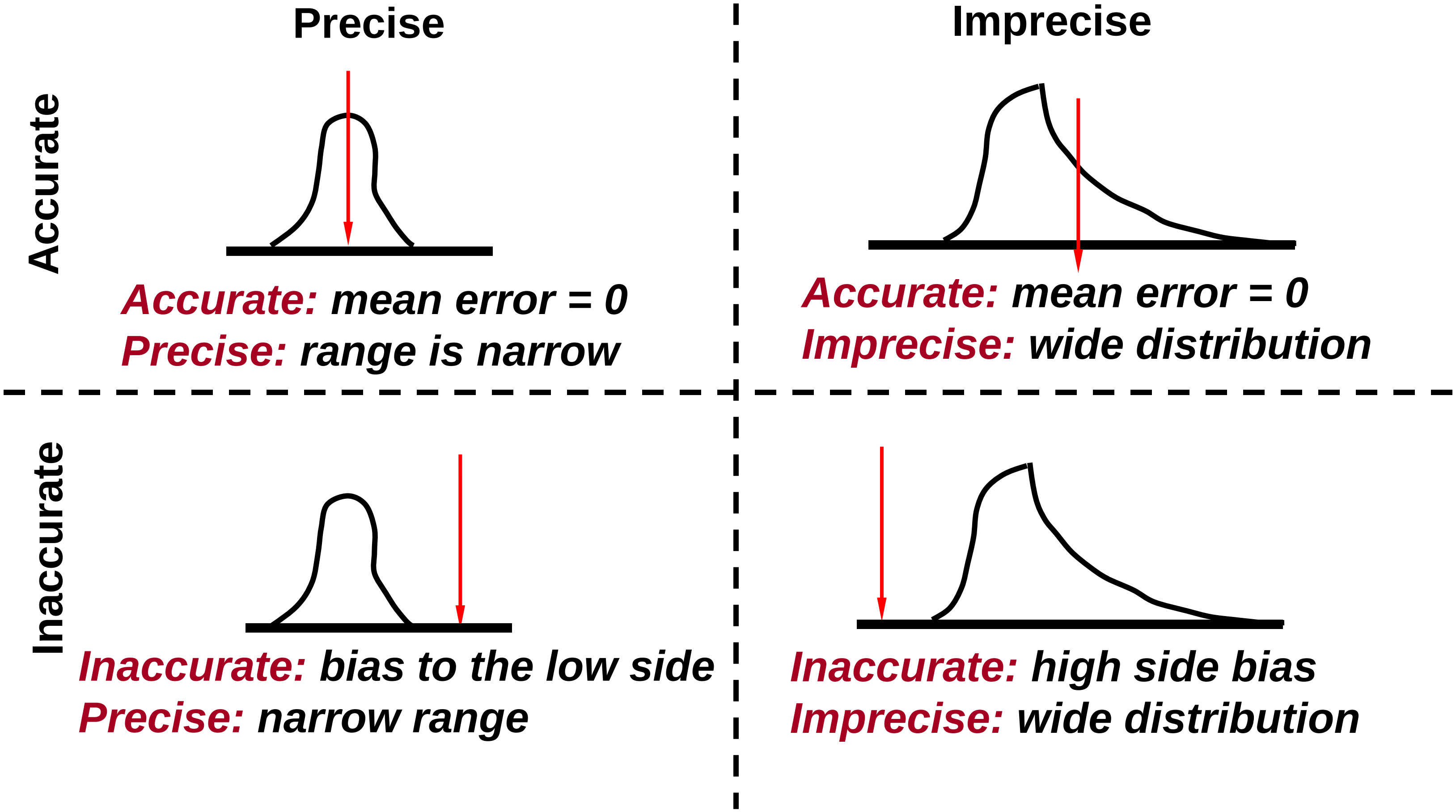


Conditional: A small subsection appears to be optimal, but the low values tend to be overestimated and high values tend to be underestimated.

conditional bias



Precision and Accuracy



Precise

Imprecise

Accurate

Inaccurate

***Accurate:* mean error = 0**
***Precise:* range is narrow**

***Accurate:* mean error = 0**
***Imprecise:* wide distribution**

***Inaccurate:* bias to the low side**
***Precise:* narrow range**

***Inaccurate:* high side bias**
***Imprecise:* wide distribution**

Geostatistical Estimation: Kriging

Kriging is an estimator that uses a weighted linear combination of surrounding data to produce unbiased, minimum variance estimates.

Kriging weights are not based on Euclidean distance, but use the geometry defined by the variograms.

Ordinary Kriging (OK): Allows for local re-estimation of the global mean. The estimate is the sum of the product of the weights and the z values.

$$\mathrm{est} = \sum_{i=1}^{n} \omega_i z_i$$

Simple Kriging (SK): Enforces the global mean on to each estimate. Sums the weighted residuals from the surrounding data points and adds that sum onto the global mean.

$$\mathrm{est} = \mathrm{mean} + \sum_{i=1}^{n} \omega_i \left[z_i - \mathrm{mean} \right]$$

Kriging

Best Linear Unbiased Estimator (B.L.U.E.)

Does Ordinary Kriging (OK) fit the requirements of a B.L.U.E.?

Best:	Minimizes the variance of the residuals (precise estimate)
Linear:	Employs a weighted linear combination of the surrounding data
Unbiased:	Attempts to make the mean residual equal to zero

Kriging – Matrix Formulation

Calculation of Kriging Weights

Local covariance matrix that describes covariance between all samples in the local search neighborhood

$$\underbrace{\begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} & 1 \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} & 1 \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} & 1 \\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46} & 1 \\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56} & 1 \\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66} & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}}_{\mathbf{C}} \cdot \underbrace{\begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \\ \omega_5 \\ \omega_6 \\ \mu \end{bmatrix}}_{\vec{\omega}} = \underbrace{\begin{bmatrix} C_{10} \\ C_{20} \\ C_{30} \\ C_{40} \\ C_{50} \\ C_{60} \\ 1 \end{bmatrix}}_{\vec{\mathbf{D}}}$$

Vector of covariances between each point in the search neighborhood and the location being estimated

To solve for vector of weights use matrix algebra:

$$\omega = \mathbf{C}^{-1} * \mathbf{D}$$

Kriging – Matrix Formulation

C_{11} is the covariance at zero separation, the sill value

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} & 1 \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} & 1 \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} & 1 \\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46} & 1 \\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56} & 1 \\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66} & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \\ \omega_5 \\ \omega_6 \\ \mu \end{bmatrix} = \begin{bmatrix} C_{10} \\ C_{20} \\ C_{30} \\ C_{40} \\ C_{50} \\ C_{60} \\ 1 \end{bmatrix}$$

C_{61} is the covariance between points 6 and 1

Add an extra row and column to assure unbiasedness

Lagrange parameter, from adding a row for unbiasedness

Kriging – Covariance Matrix

- There is no guarantee of a unique solution to the matrix system. To ensure that there is only one unique solution, the system must be positive definite
- For estimates that are weighted linear combinations of other values, the variance about those estimates must be greater than or equal to zero
- Positive definite condition can be achieved by modeling variograms with positive definite functions, as long as one of the standard models is used (which are positive definite functions) the covariances will create a positive definite set of matrices

Kriging – Unbiasedness

We want the mean, or expectation, of the errors to equal zero.

- **Define error as: the estimated value – actual value**

$$E(x_0) = \sum_{i=1}^{n} \omega_i \overset{\downarrow}{Z}(x_i) - \overset{\downarrow}{Z}(x_0)$$

- **Realize that the mean is stationary, so both the estimate and the actual value have the same mean.**
- **If the average error is set to zero, then:**

$$\sum_{i=1}^{n} \omega_i = 1.0$$

Kriging – Lagrange Parameter

- **Lagrange parameter, m, solves the problem of n+1 equations and only n unknowns created by the unbiasedness constraint**
- **Lagrange parameter is essentially another unknown**

$$\sum_{i=1}^{n} \omega_i = 1 \xrightarrow{\quad} \sum_{i=1}^{n} \omega_i - 1 = 0 \xrightarrow{\quad} 2\mu\left(\sum_{i=1}^{n} \omega_i - 1\right) = 0$$

Kriging – Summary

Covariances in $\vec{D}$ act like inverse distance weights (two points close together have a high covariance, as the distance increases the covariance approaches zero).

- However, the weight as a function of distance is not limited to simple powers, it can fit with more complex variogram models.

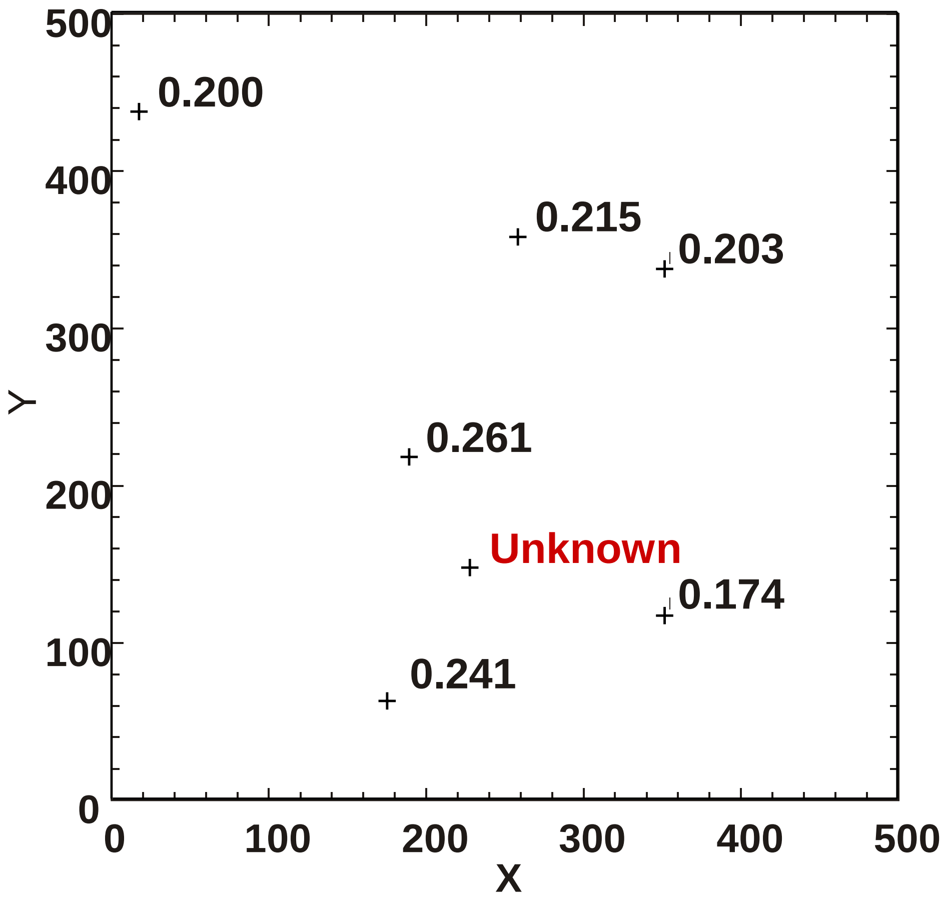
Distances are not in Euclidian space, but are relative to variogram range.

- Can provide anisotropy and weights of zero, if a point is beyond the range of the variogram it will be assigned a zero weight

Covariances in C act to decluster the data

- C^{-1} is adjusting the weights in $\vec{D}$ for data redundancy

Kriging – Results



- Porosity measured at 6 points.
- Make estimate of porosity at unknown point, x_0.

Kriging – Results

For the example problem, the normal-score variogram model is spherical, with a range of 125.0 (N-S) and 100.0 (E-W) and a nugget of 0.0

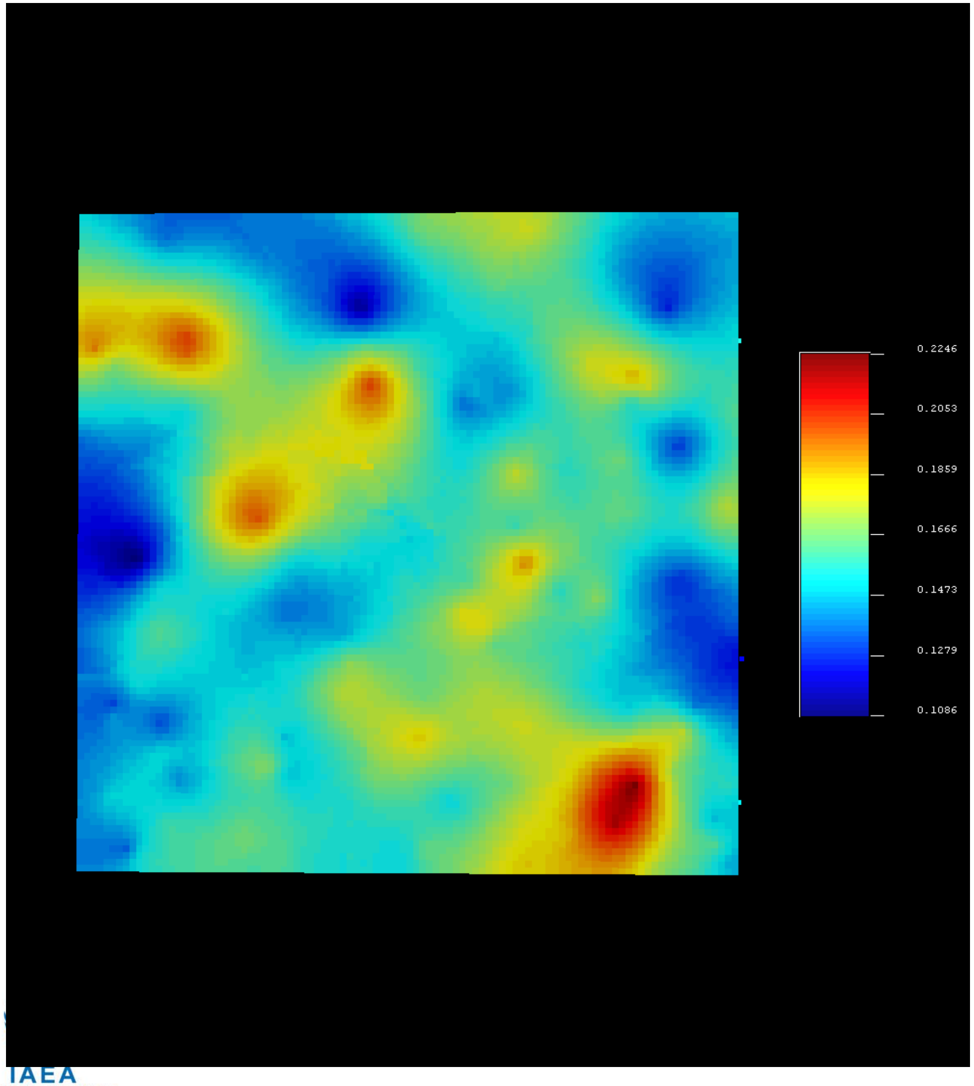
Location	Distance from X_0	Kriging Weight (radii = range)
1	80.6	0.541
2	123.7	0
3	224.7	0
4	212.1	0
5	94.3	0.459
6	358.1	0

Note that only points 1 and 5 fall within the search ellipse

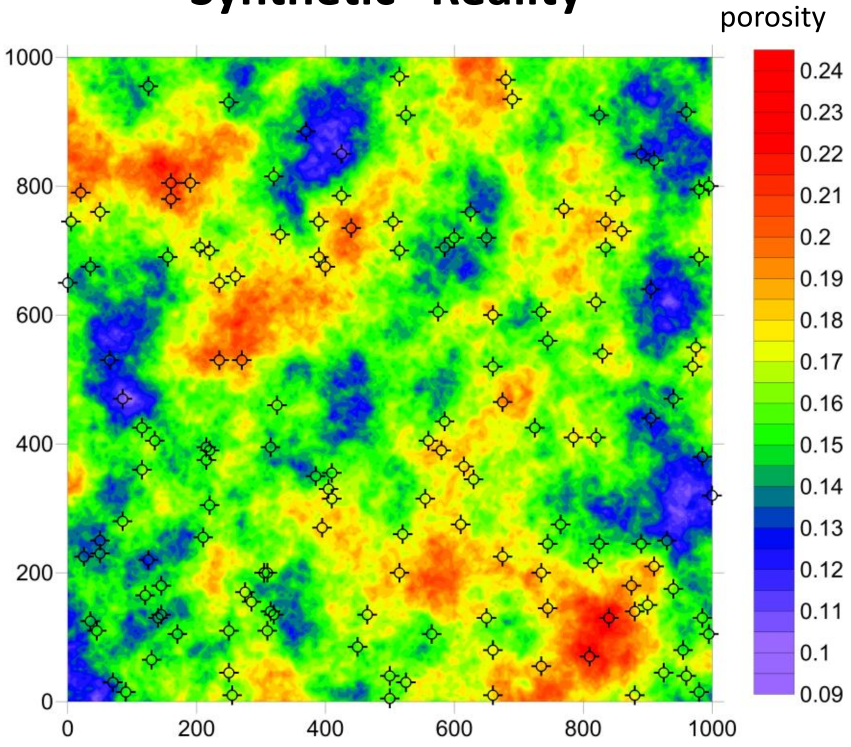
Estimate of value at x_0 = 0.250

Kriging – Results

Example from workshop exercise data set



Synthetic “Reality”



Kriging – Error Variance

- Kriging is unique among spatial estimation techniques in attempting to minimize errors, making the distribution tight.
- Concisely, it is the variance of the errors that is minimized

$$\sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2$$

Where each error, e_i = the estimate - the true value
and $\bar{e}$ is the mean error across the domain

Kriging – Error Variance

Long Version: Derive a model of the error variance and minimize the modeled variance by setting partial derivative for each weighted covariance (w.r.t each weight) between datum and estimation location to zero.

Short Version: The estimate of error variance is the total variance minus the weighted sum of covariances in $\vec{D}$ plus the Lagrange parameter

$$\hat{\sigma}_e^2 = \sigma_{Data}^2 - \sum_{i=1}^{n} \left(\omega_i C_{i0}\right) + \mu$$

in matrix form:

$$\hat{\sigma}_e^2 = \sigma_{Data}^2 - \left(\vec{\omega} \cdot \vec{D}\right)$$

where $\hat{\sigma}_e$ is the local estimate of the error variance

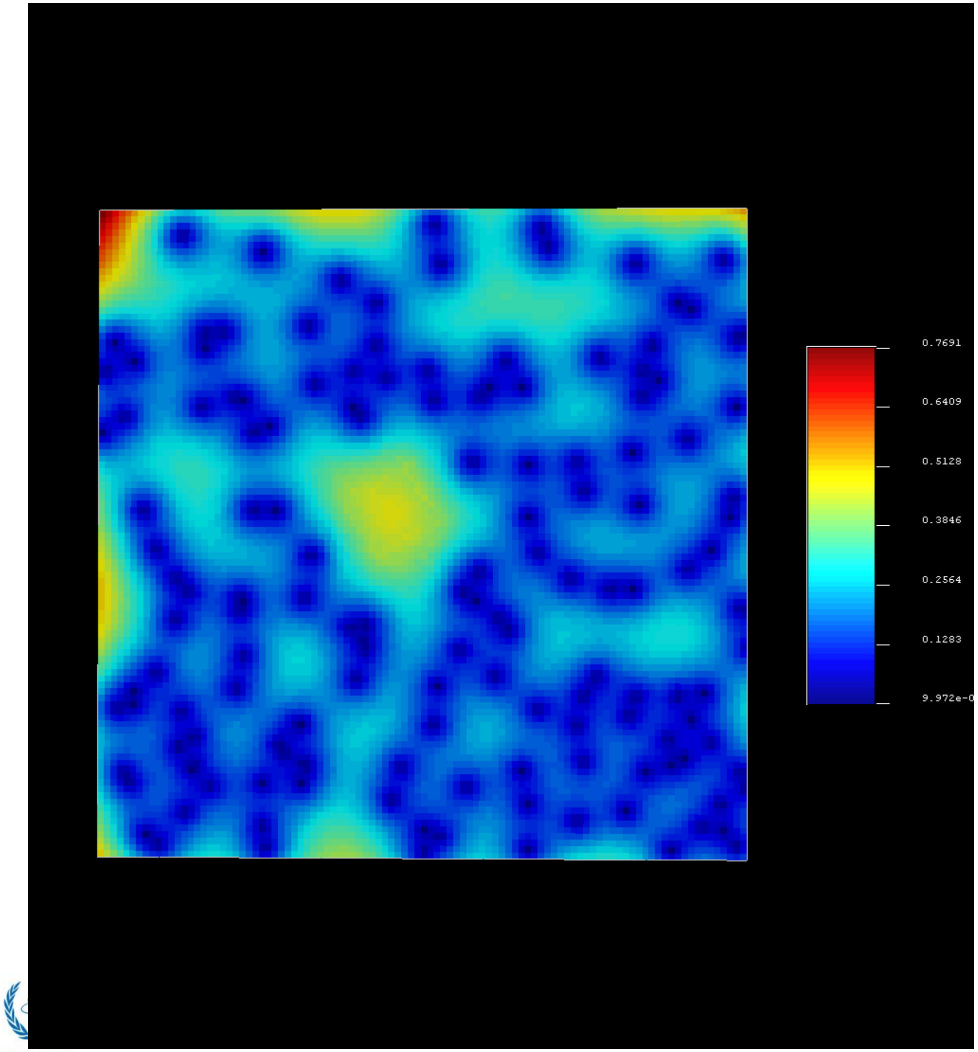
Kriging – Error Variance

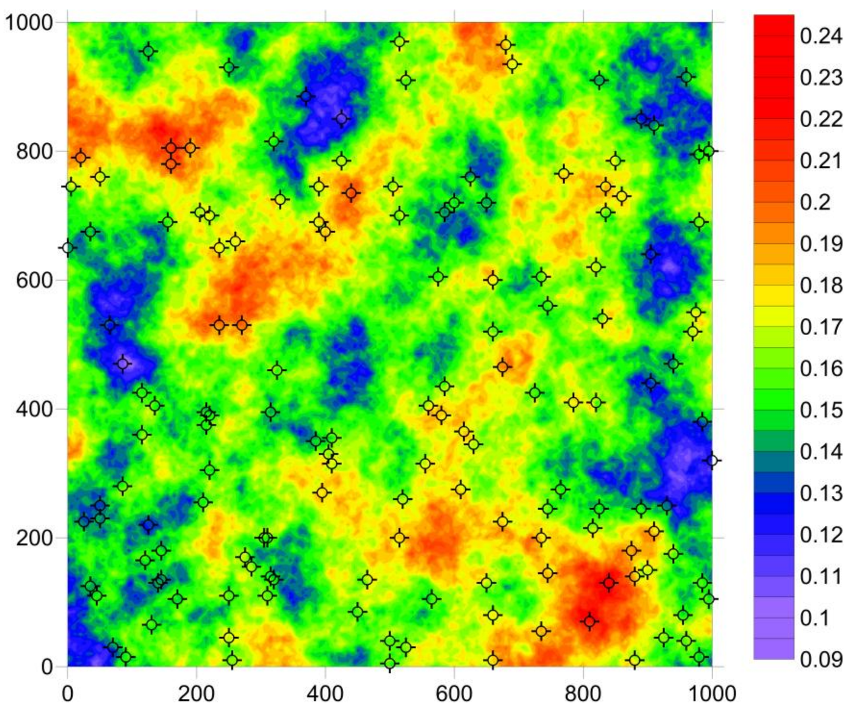
- **Error variance is also called kriging variance or estimation variance**
- **Error variance is <u>not</u> a function of data values but only of sample configuration**
- **Error Variance is equal to zero at data location**
- **Like the entire kriging system, distribution of errors is <u>non-parametric</u>**

(although a gaussian distribution is often assumed for errors)

Kriging – Error Variance

Example from workshop exercise data set





Synthetic "Reality"

Spatial Estimation Summary

- Review of three traditional point estimation techniques on a simple example problem
- Introduction of techniques that can be used to assess the accuracy and precision of estimators
- Introduction to kriging including the matrix formulation of the kriging system
- Assessment of kriging system as a "BLUE"
- Introduction to kriging variance

Spatial Simulation

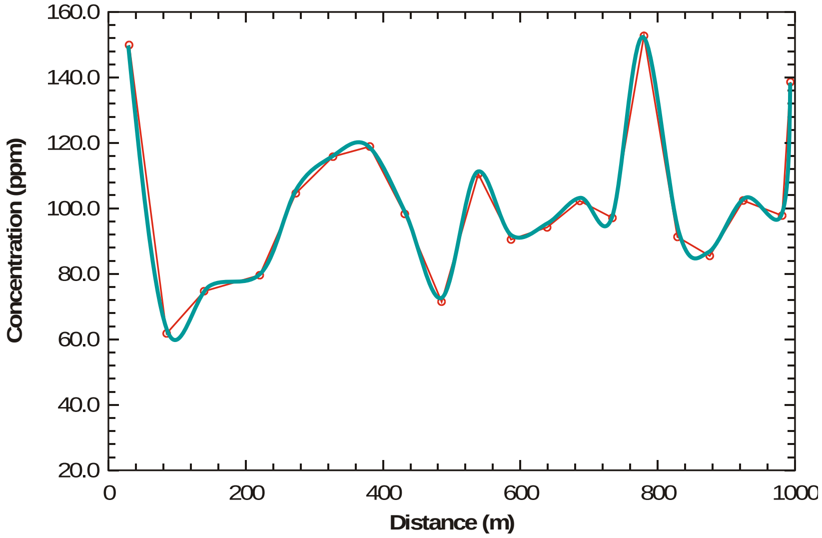
- Probability-based techniques (Monte Carlo process) on spatially correlated distributions
- Sacrifices the local best estimate for the reproduction of global statistics and features
- Simulation process can create any number of equally probable realizations (maps), all of which honor the available information
- Simulation allows for evaluation of joint uncertainty (accuracy) at multiple locations
- A large suite of geostatistical simulation methods are available

Spatial Simulation

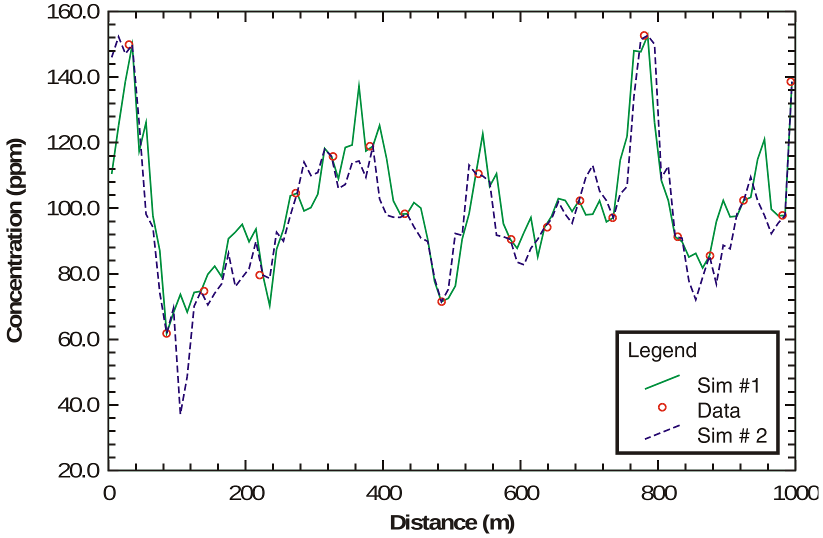
- Simulation provides a more realistic picture of natural complexity and heterogeneity relative to kriging
- Simulation can provide an idea of "best," "most likely" and "worst" cases for a given problem
  - bounding cases
- Simulation is a basis for Monte Carlo risk analysis where a full distribution of results is necessary
  - full distribution
- Simulation reproduces the observed level of variability or heterogeneity at a site
- Effectively extrapolates parameter values, whereas kriging only interpolates parameter values

Estimation versus Simulation

Estimation (Kriging)



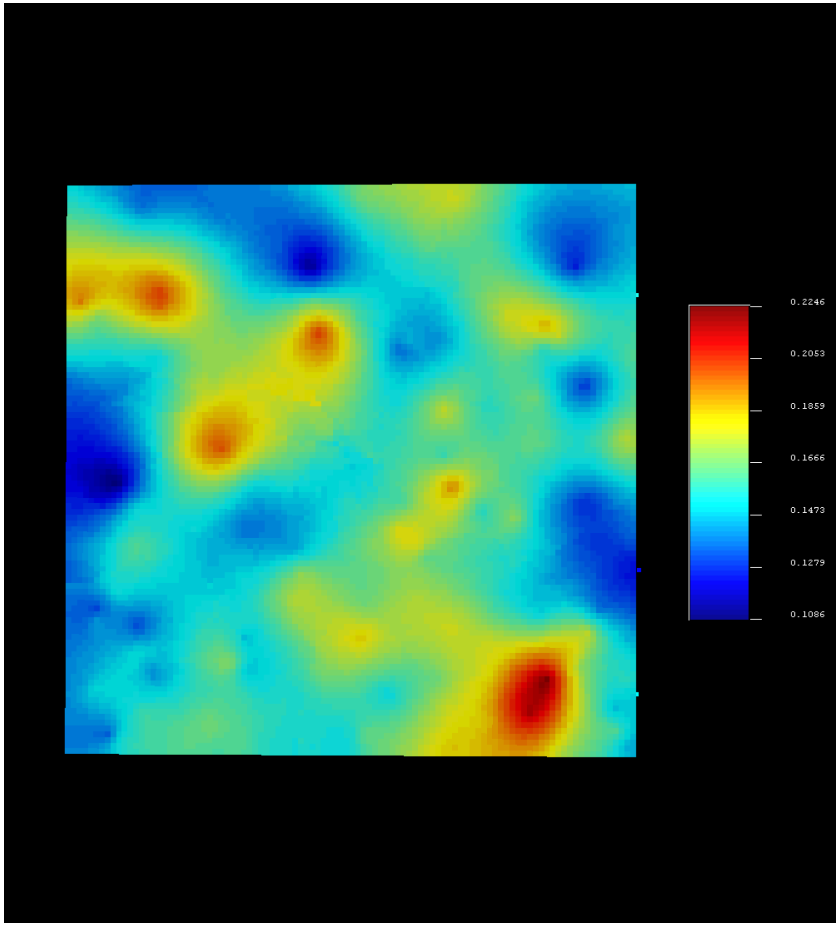
Simulation



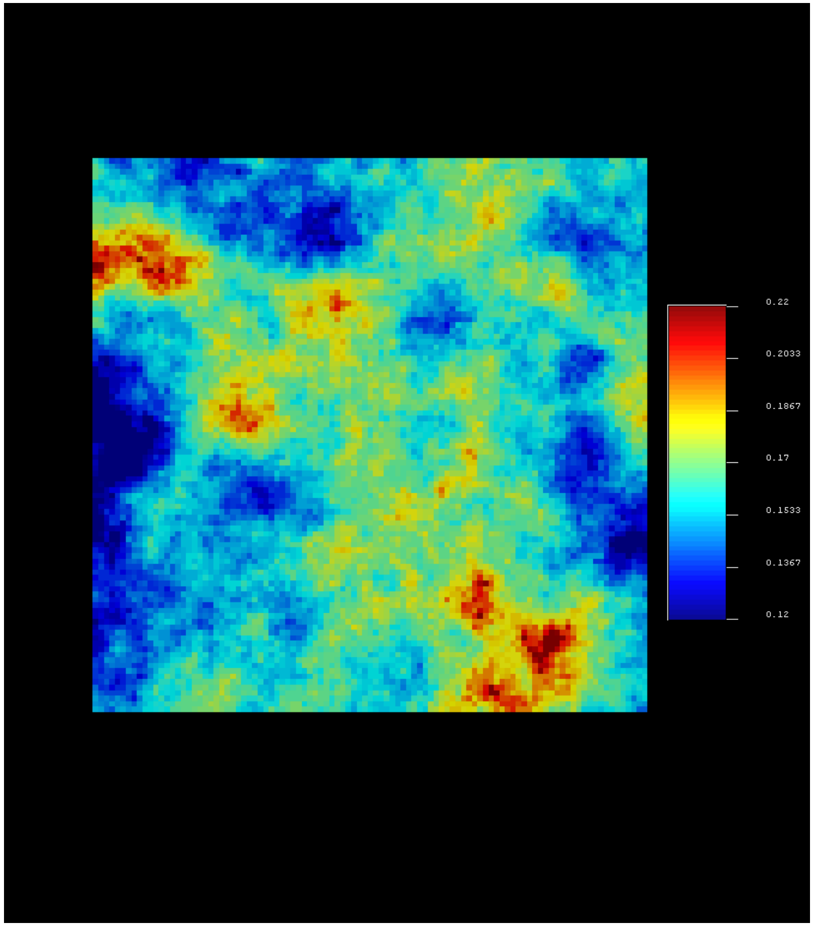
Estimation versus Simulation

Example from workshop exercise data set

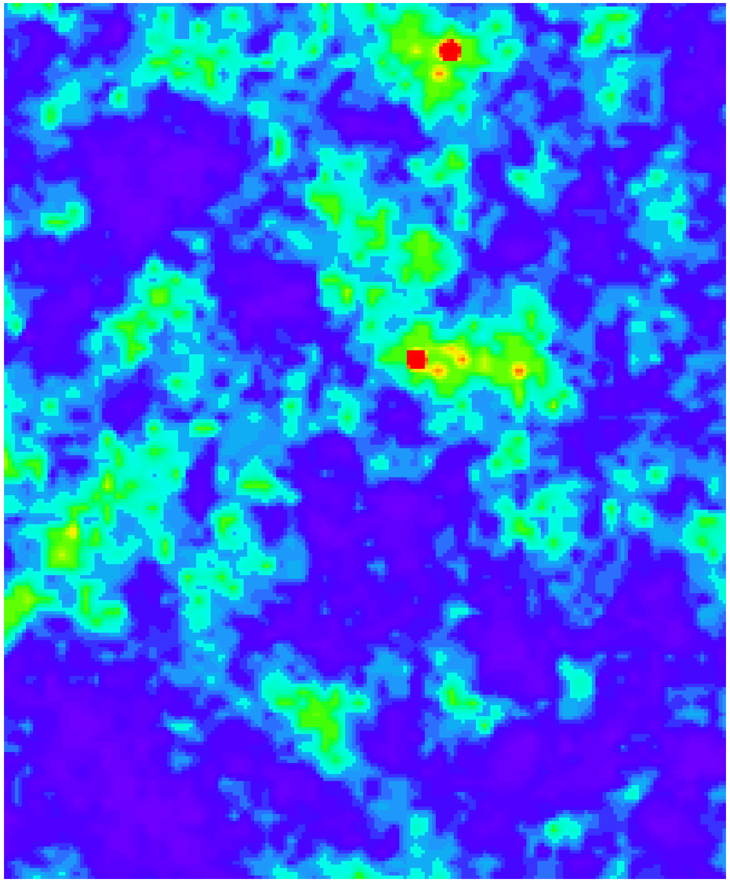
Ordinary Kriging (Estimation)



Sequential Gaussian Simulation



Simulation Example



The Monte Carlo process can create multiple images of the activity at the site.

Every image honors:

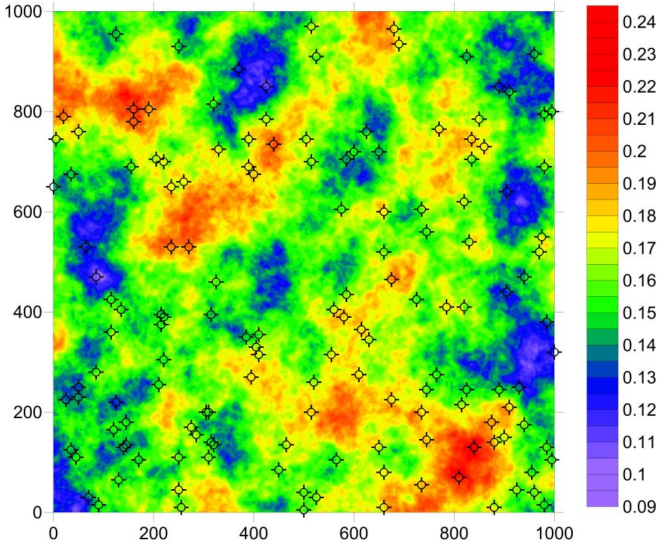
- Sample data
- Histogram
- Variogram

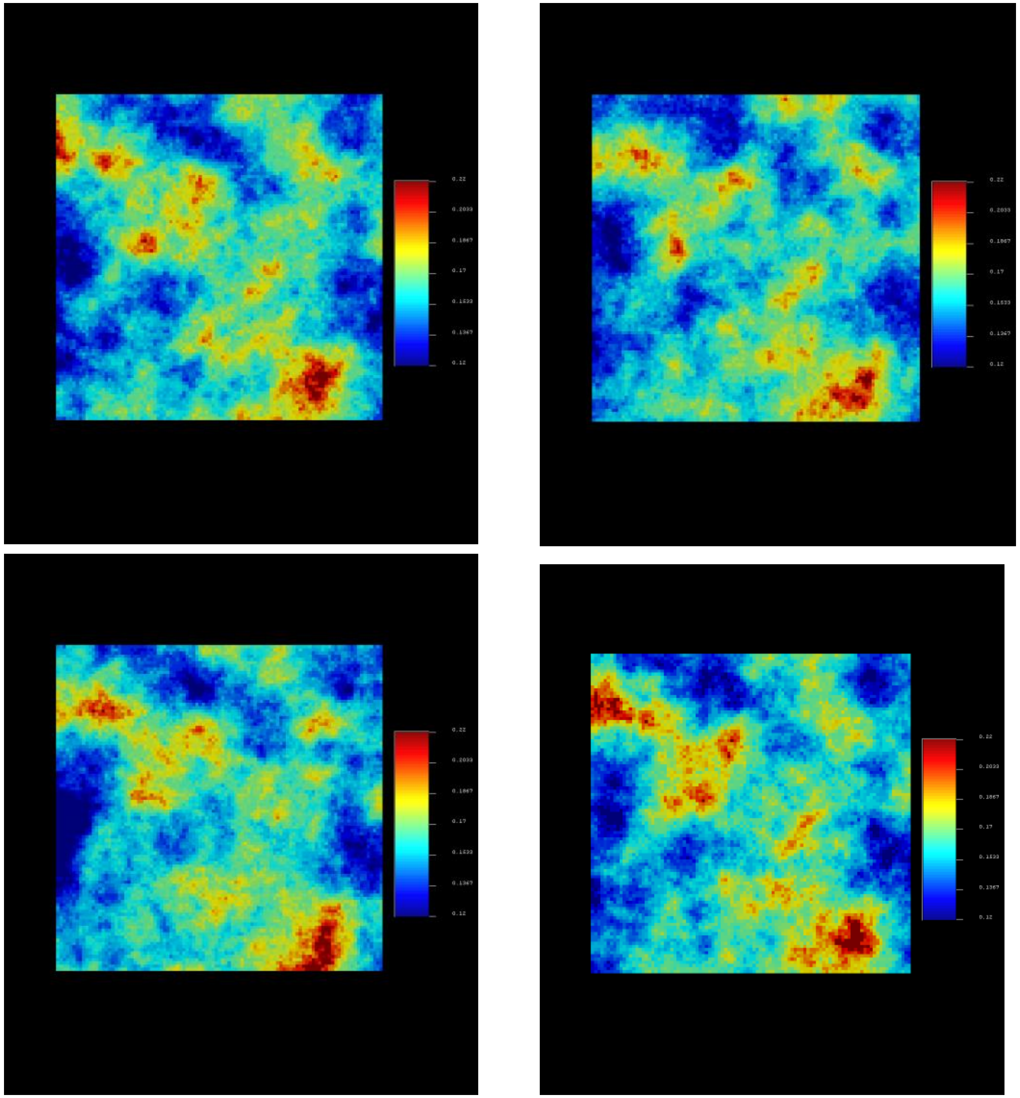
Given our knowledge of the site, *every image is a plausible depiction of the real parameter distribution.*

Simulation Example

Example from workshop exercise data set

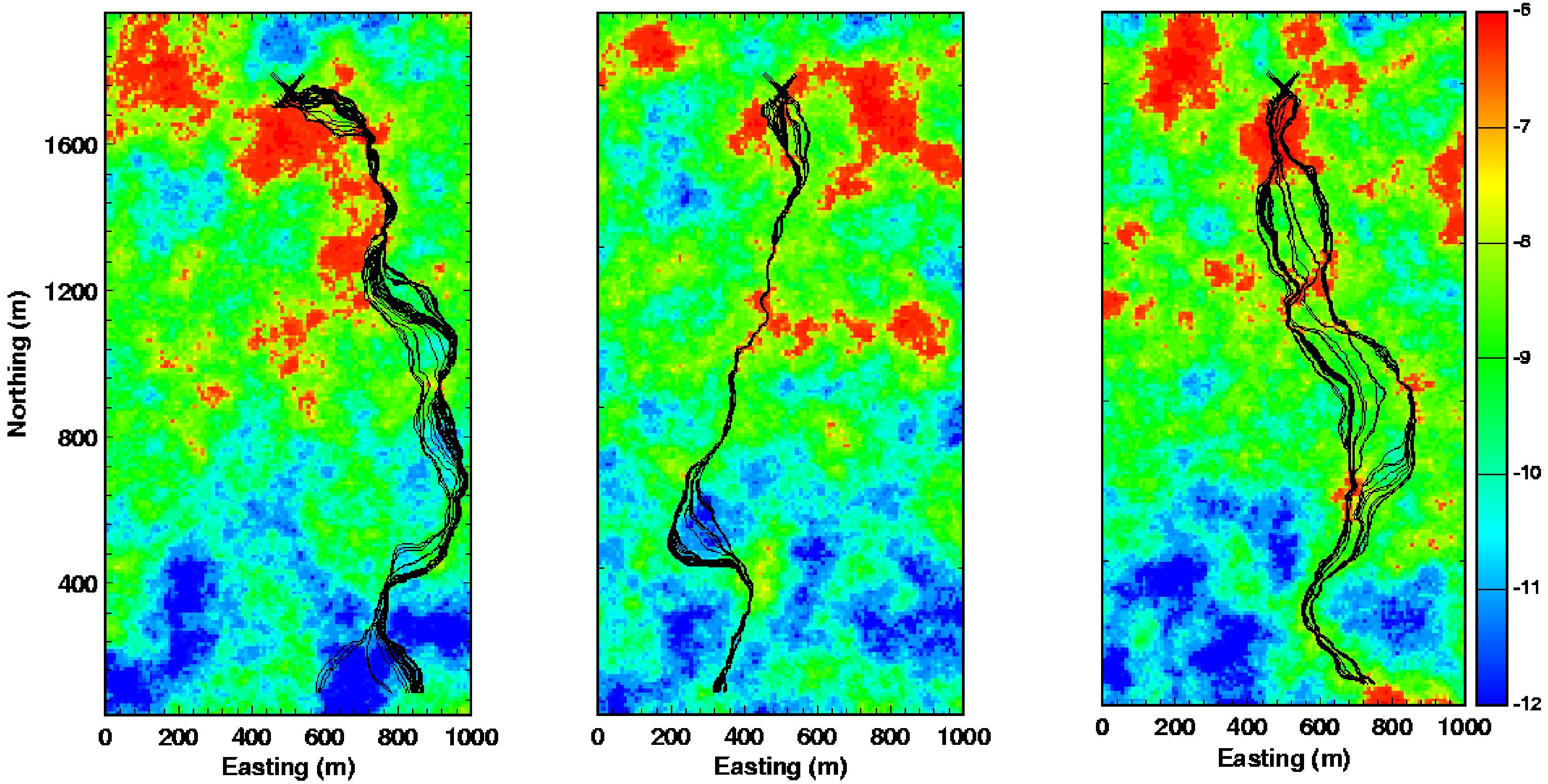
Synthetic "Reality"



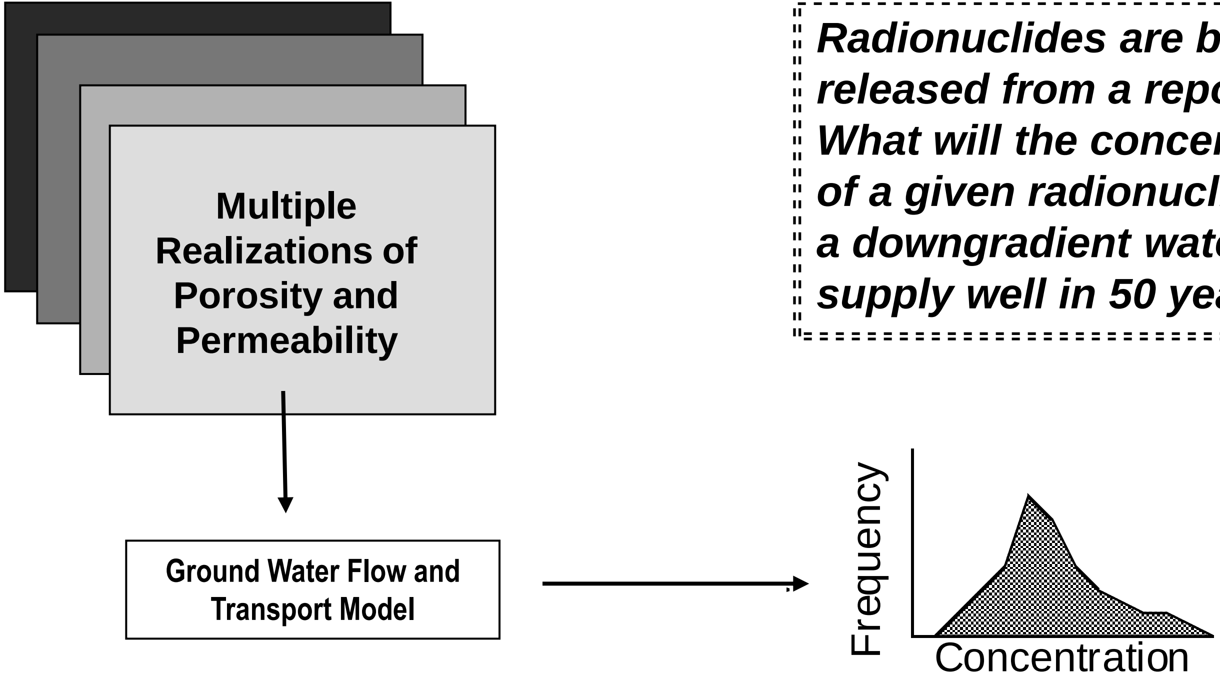


Transport Example

Uncertainty in spatial distribution of hydraulic properties leads to uncertainty in transport results



Transfer Function



Multiple Realizations of Porosity and Permeability

Ground Water Flow and Transport Model

Radionuclides are being released from a repository. What will the concentration of a given radionuclide be at a downgradient water supply well in 50 years?

General Types of Simulation

Parametric: requires transform of the data to a parametric space, simulation in that space and then back-transform to raw data space.

Example: Gaussian simulation using the normal-score transform
Advantage: only requires one-variogram model
Disadvantage: does not reproduce variogram at extremes of the distribution

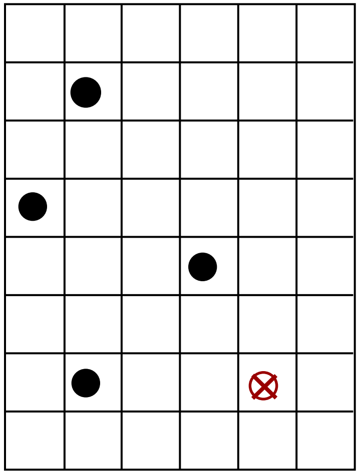
Non-Parametric: requires discretization of data into classes and a variogram model at each threshold or class.

Example: indicator simulation of geologic facies (sand, silt, clay)
Advantage: Reproduces each variogram at each class/threshold
Disadvantage: requires variogram modeling for each class/threshold

Sequential Simulation

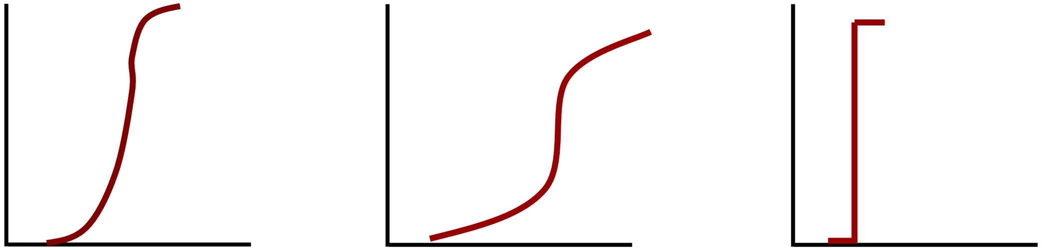
- Map the conditioning data onto a grid
- Randomly visit all other grid nodes
- Use kriging system to create a local cdf based on surrounding data for each node
- Draw a random value from the cdf to get the simulated value at that location
- Consider each simulated point as a conditioning value for future cdf construction
- Continue until all nodes have a simulated value
- Reinitialize random number generator and begin next realization

Sequential Simulation



- Use the kriging system to create a local cdf based on the surrounding data points of the first node.
- This cdf can be parametric or indicator
- Draw a random number between 0 and 1, and assign the value for that probability to the node.
- For the remainder of this realization, the newly defined node is treated as a sample point.

Local cdfs reflect proximity to data locations



Well defined
Close to data

Poorly defined
Far from data

On top of / right next to value
On top of data point

CDF Construction

Gaussian: normal-score transform allows the kriging estimate and kriging variance to define the local cdf.

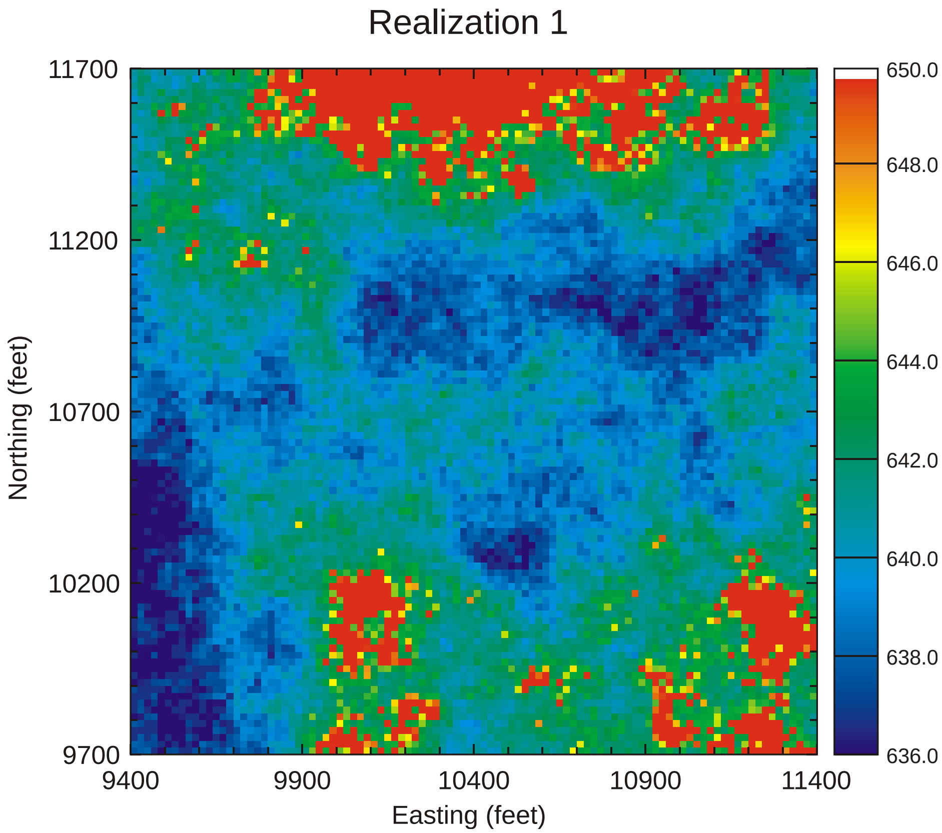
$$\text{mean} = \text{est} = \sum_{i=1}^{n} \omega_i z_i \qquad \text{variance} = \hat{\sigma}_e^2 = \sigma_{\text{Data}}^2 - \sum_{i=1}^{n} \left(\omega_i C_{i0}\right) + \mu$$

Indicator: construct the cdf through indicator kriging at each threshold z_k.

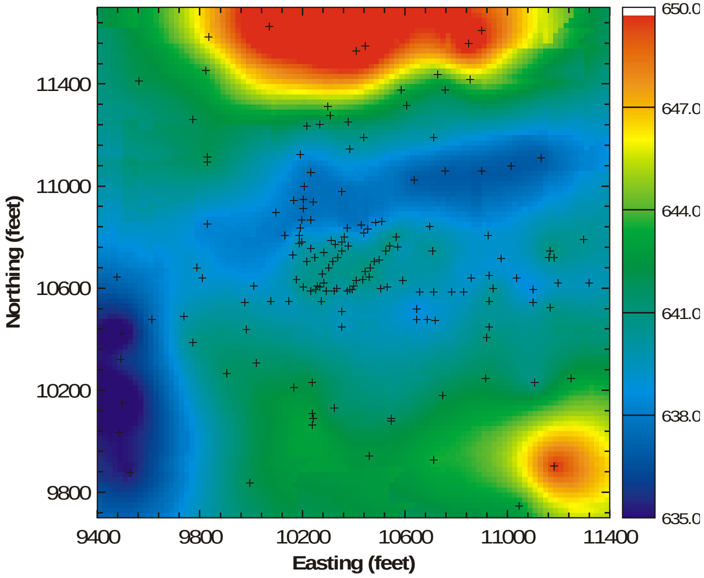
The expected value of the cdf at any threshold is estimated by the weighted linear combination of surrounding indicator data.

Gaussian Simulation Example

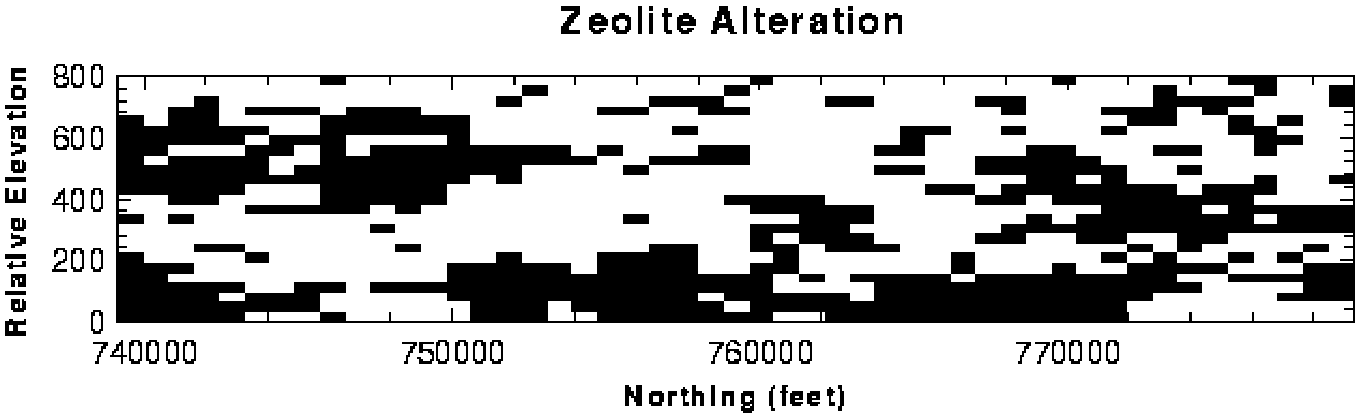
Aquifer bottom elevation, Portsmouth, Ohio



Simulation shows greater variability in elevation values relative to kriged map



Indicator Simulation Example



Kriging versus Simulation

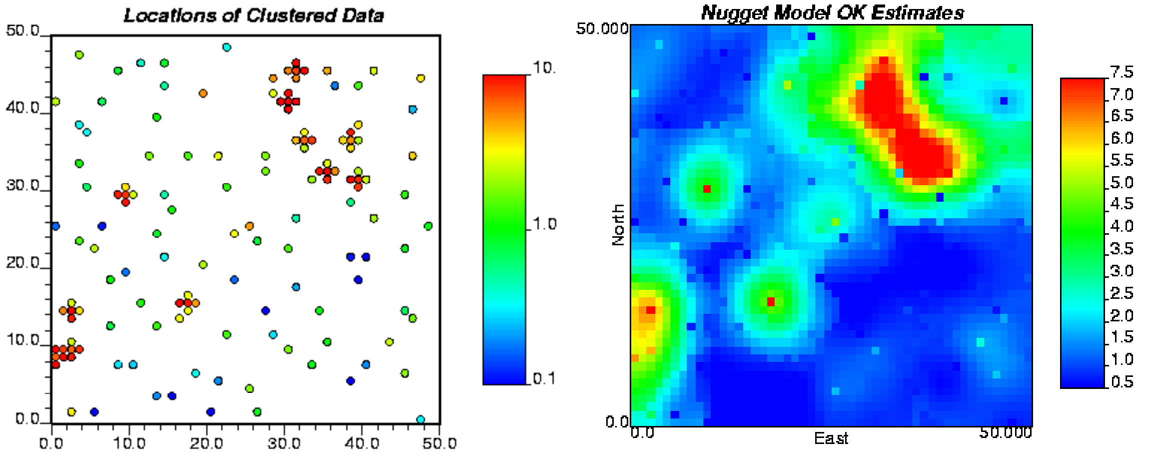
Kriging: smoothing effect of interpolation will produce:

1) A longer range variogram in the output than the input model
2) Less variability in the output field than the input data (distribution gets squeezed)

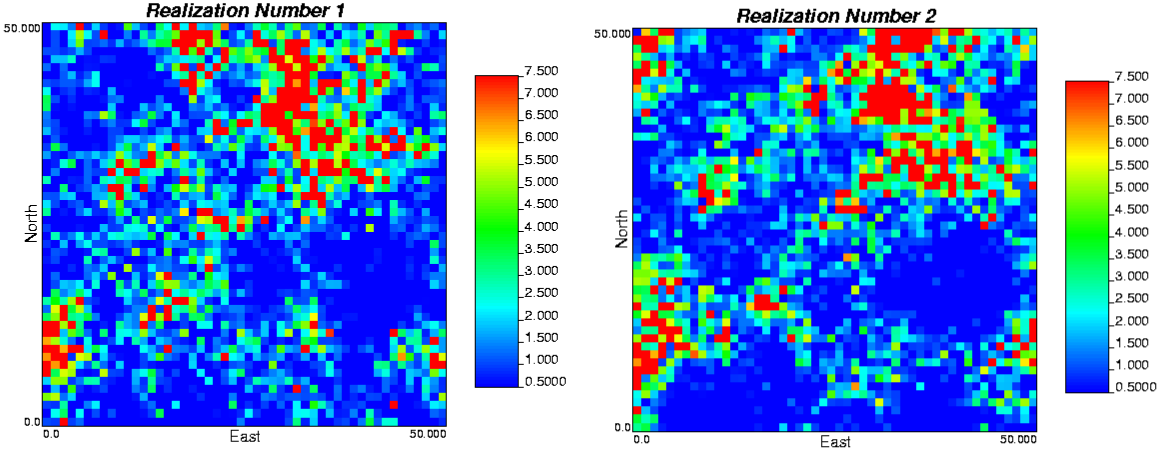
Simulation: attempts to reproduce the input histogram and variogram (the input univariate and bivariate data distributions, respectively) within the limits of "ergodic fluctuations".

Kriging versus Simulation

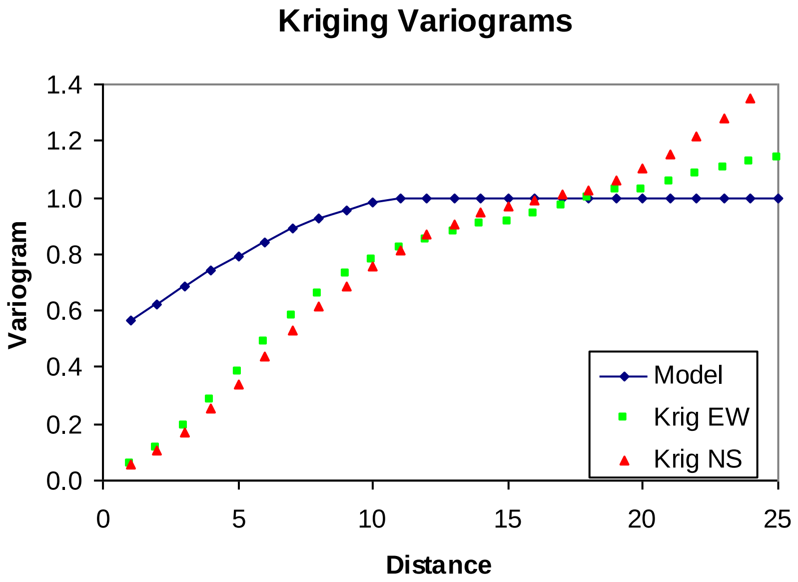
Sample Data and Kriging Estimate

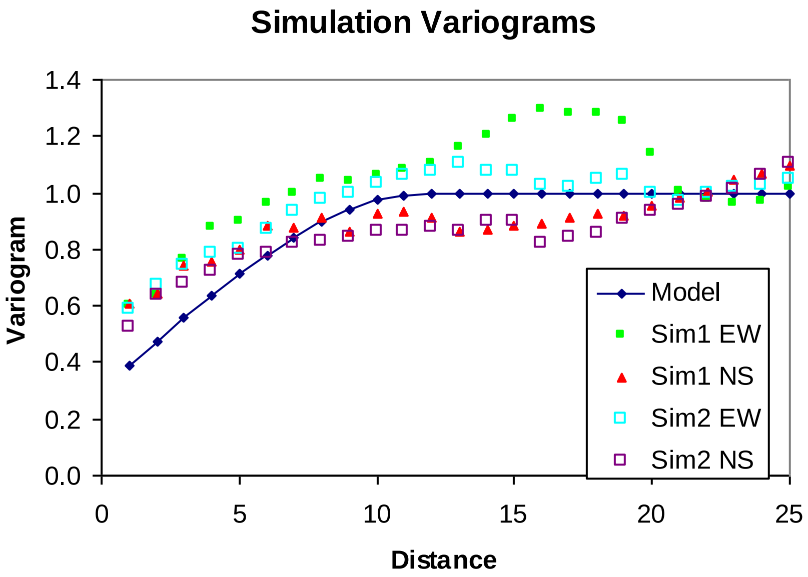


Two example realizations from simulation



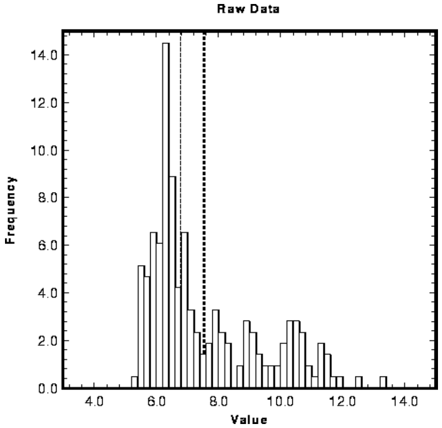
Kriging versus Simulation

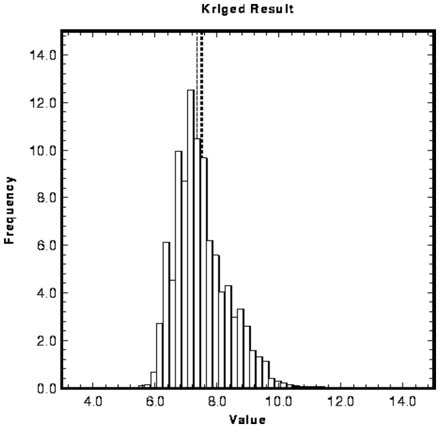




Kriging versus Simulation

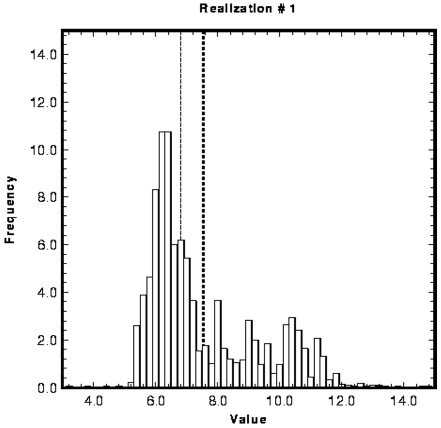
Histogram of raw data

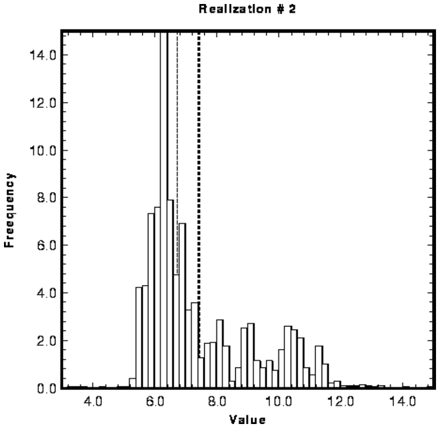




Histogram of kriged data

Histogram of Realization 1 (simulation)





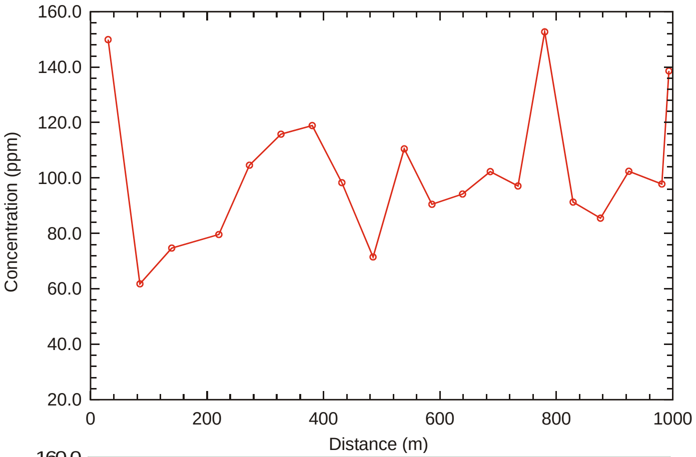
Histogram of Realization 2 (simulation)

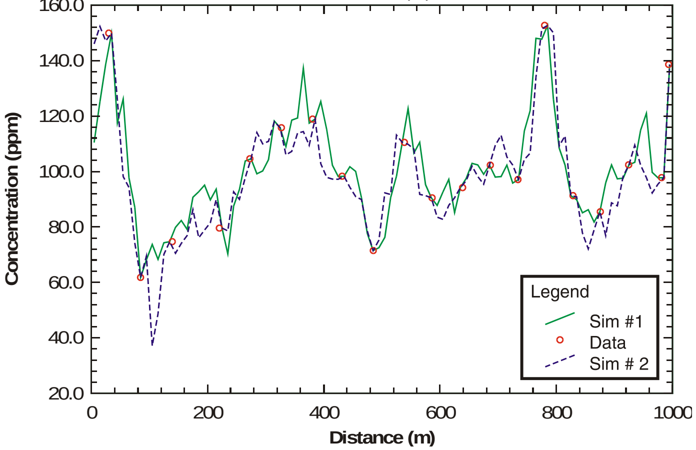
Kriging versus Simulation

Parameter	Raw Data (n=214)	Kriged Map (n=5329)	Realization 1 (n=5329)	Realization 2 (n=5329)
Mean	7.53	7.49	7.41	7.52
Median	6.79	7.34	6.70	6.81
Standard Deviation	1.81	0.86	1.75	1.80
Minimum	5.20	5.46	3.18	3.10
Maximum	13.24	13.13	14.93	14.88
10th Percentile	5.80	6.46	5.80	5.80
90th Percentile	10.47	8.75	10.39	10.47

Kriging versus Simulation

- **Kriging** reduces variance but retains the mean of the input data
- **Kriging,** as an interpolator, does not produce values outside minimum and maximum of sample data
- **Simulation** can produce values above and below the maximum and minimum sample data because it draws from a fully defined cdf [0,1] at each location.

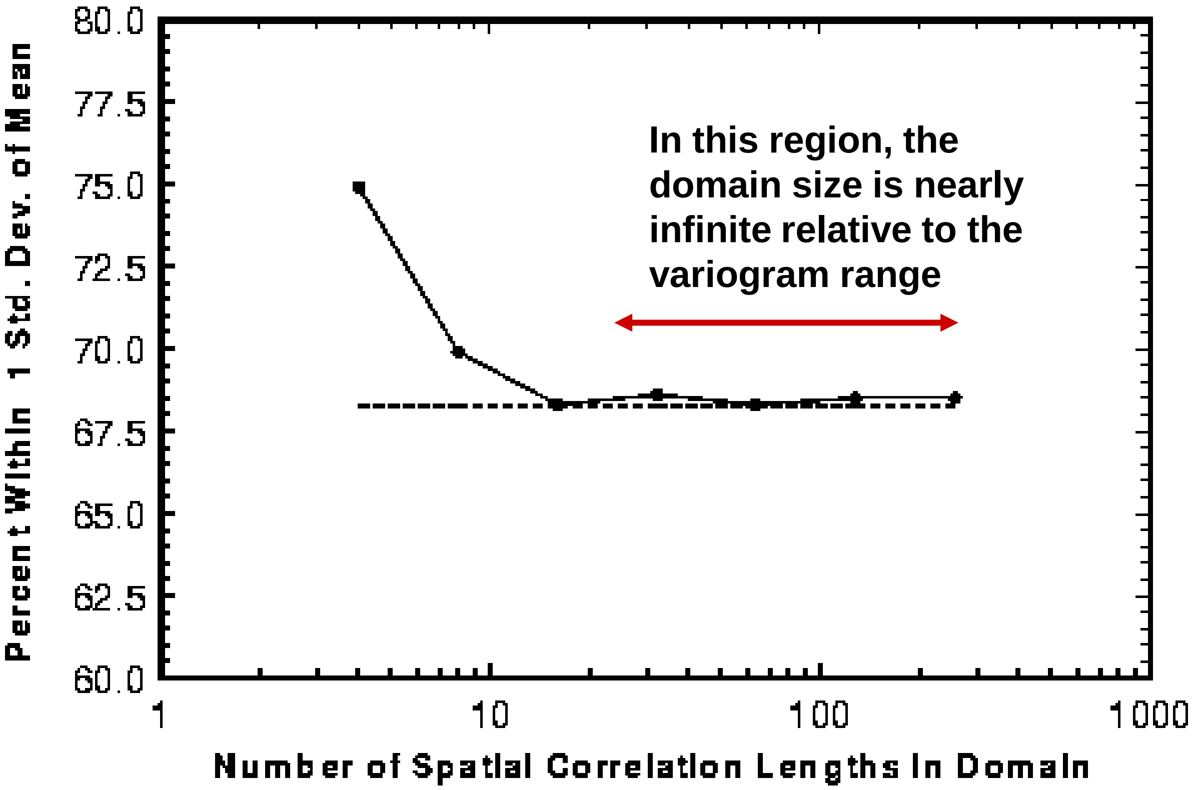




Ergodic Fluctuations

- **Ergodic fluctuation** is defined as the difference between the input model and the statistics of a realization.
- Input models are generally based on data from a limited sample size
- The underlying model is said to be ergodic in the paramenter α if the realization statistics tend toward α as the size of the field increases

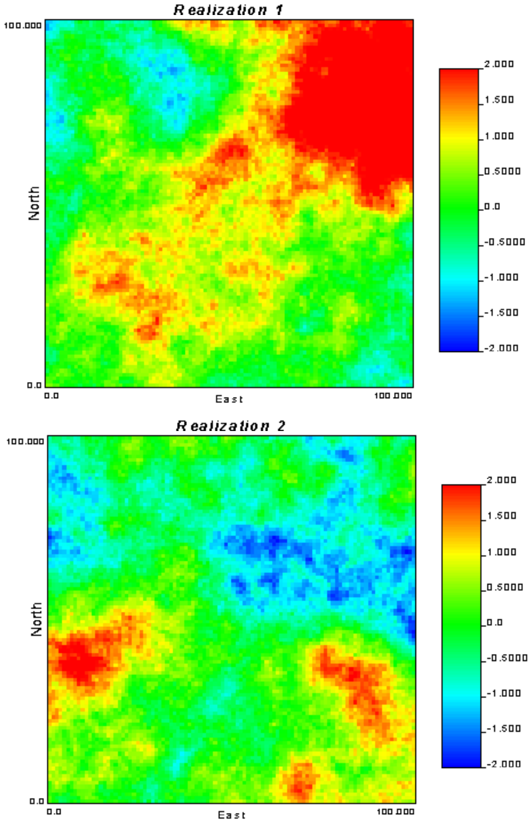
Ergodic Fluctuations



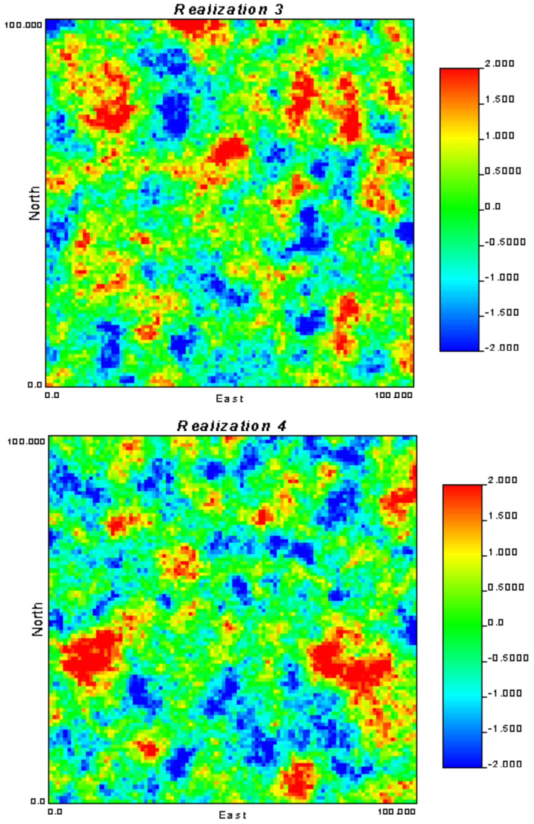
In this region, the domain size is nearly infinite relative to the variogram range

Ergodic Fluctuations

Range = 80.0



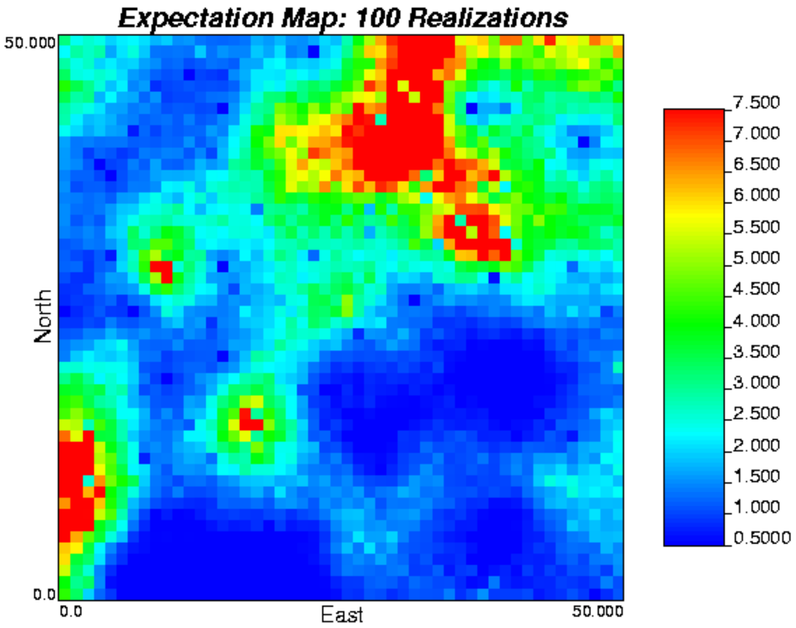
Range = 10.0

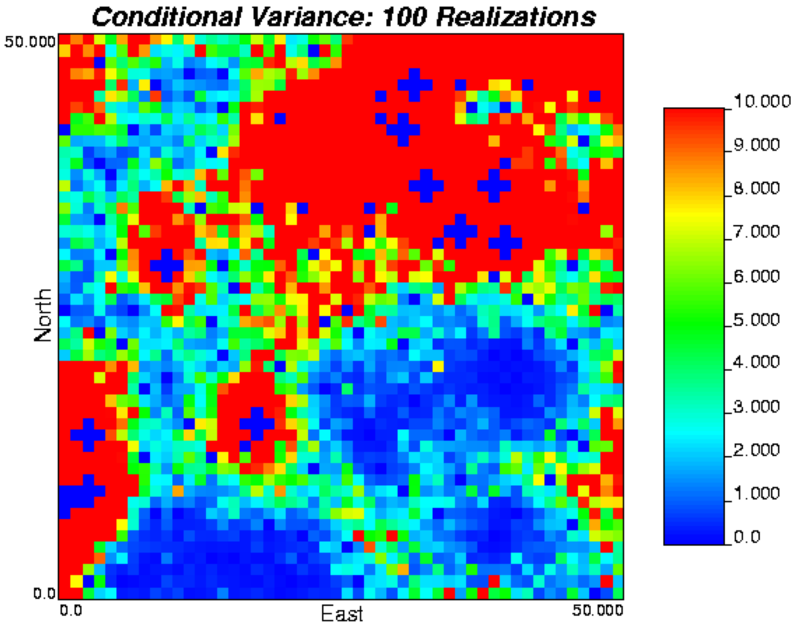


Simulation Post Processing

Expectation, or Average, Map

Conditional Variance Map

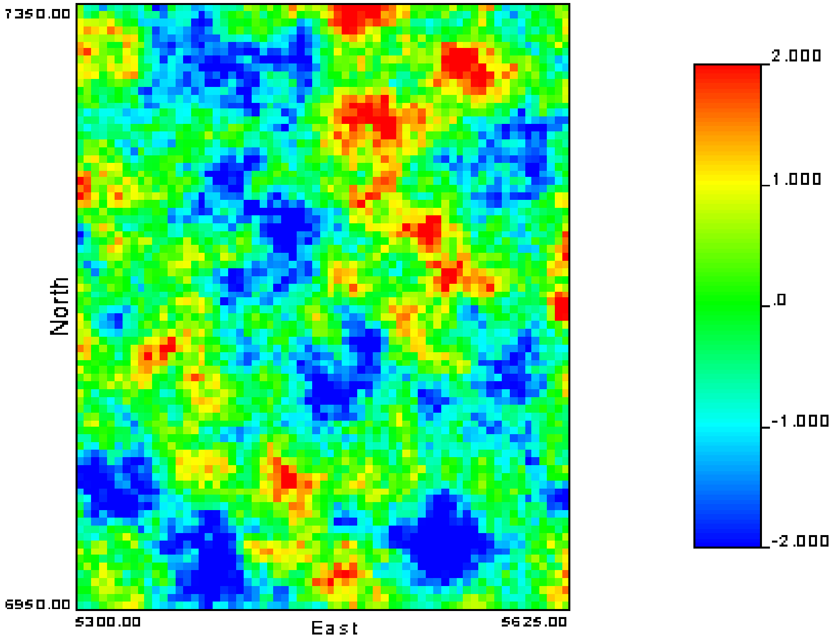




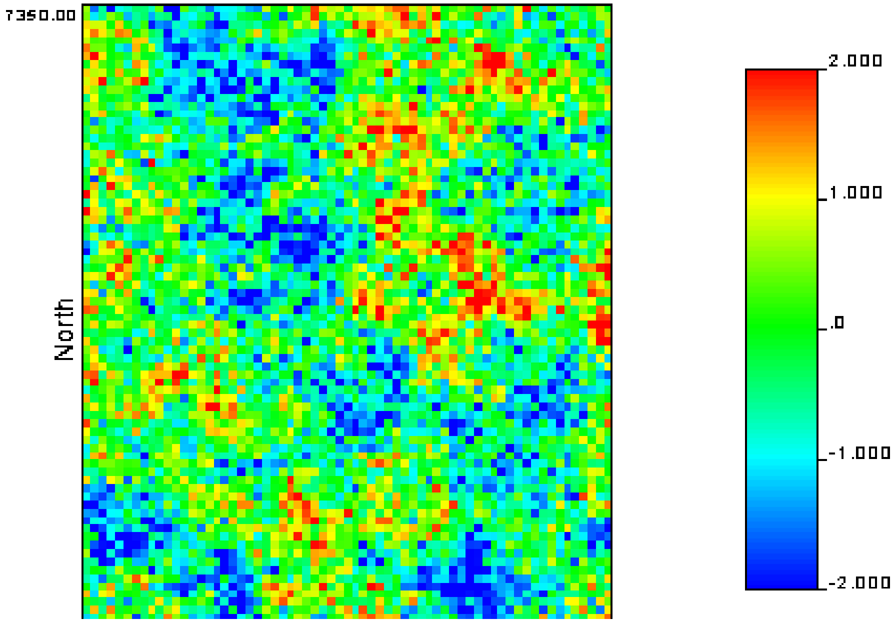
In the limit of an infinite number of realizations, these two maps will be identical to the kriging map and the kriging variance map

Simulation – Nugget Effect

Nugget = 0.00

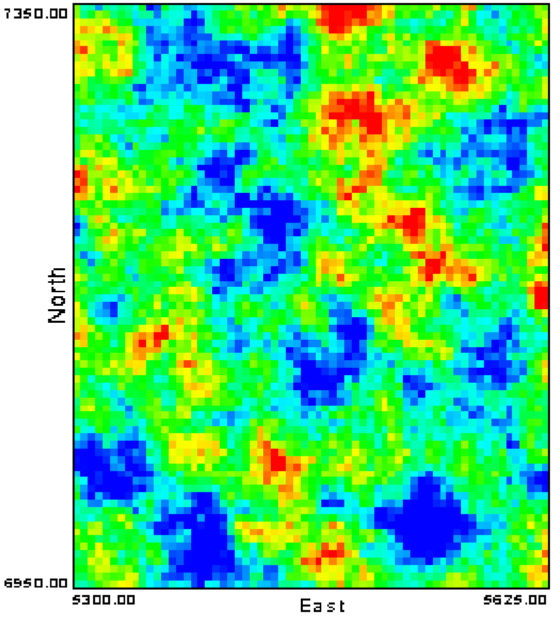


Nugget = 0.40

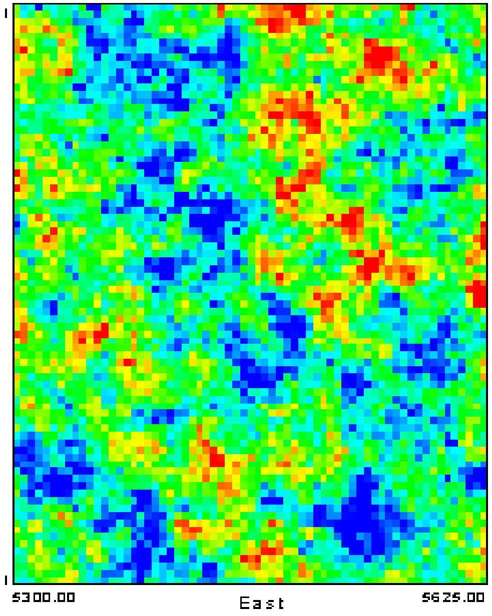


Simulation – Variogram Models

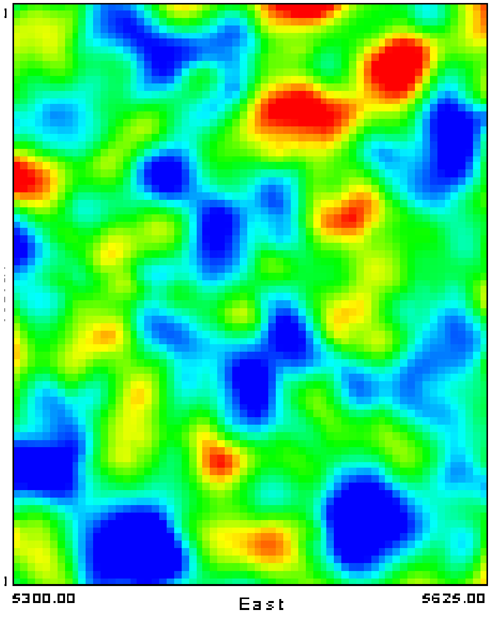
Spherical



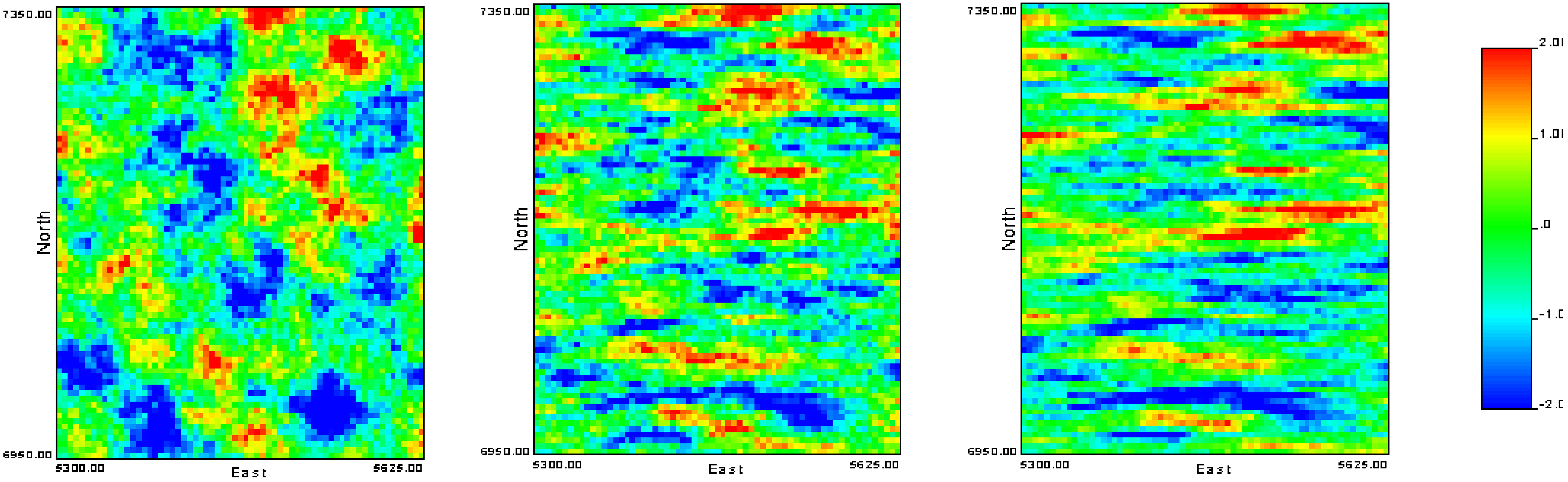
Exponential



Gaussian



Simulation – Anisotropy



Isotropic

Anisotropy = 4.0

Anisotropy = 12.0

Simulation Summary

- Introduction to geostatistical simulation and different simulation algorithms
- Mechanics of sequential simulation
- Differences of parametric and non-parametric simulation
- Comparison of simulation to kriging
- The ergodic assumption
- Postprocessing of simulations for probability maps