

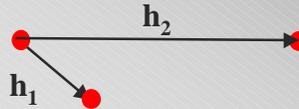
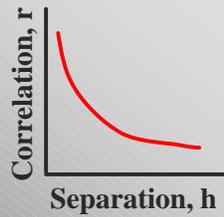
Section 3 Quantification of Spatial Continuity

- **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 geo-statistics apart from traditional statistics**
- **The traditional measurement of spatial correlation within geo-statistics is the semi-variogram, commonly called the variogram.**

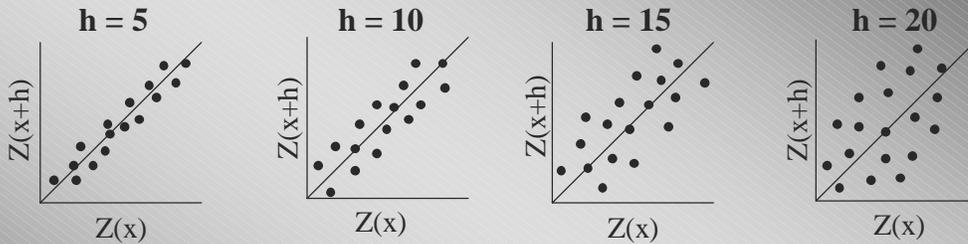
How does the correlation coefficient, r , behave as h increases?



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

Smart Sampling

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, things are very well behaved*
- *As the separations increase things are not as strongly correlated.*

Mound Accelerated Site Technology Deployment

3-4

A simpler concept than the variogram, the scatterplot gives a more intuitive view of spatial correlation, showing correlation dependency on lag spacing between any pair.

On a site with strong spatial correlation at $h = 100$, the points z are strongly correlated with those 100 feet away, $z(x+h)$.

As h increases, the correlation between values drops off. At very large separation, there is no correlation (if the concentration is 100 pCi/g at one point, on other side of the site it may be 0 or 1000 pCi/g).

- *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, there may not be good correlation between them (non-stationarity).*

Emphasis is on the relative difference in measurements, not the absolute magnitude.

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

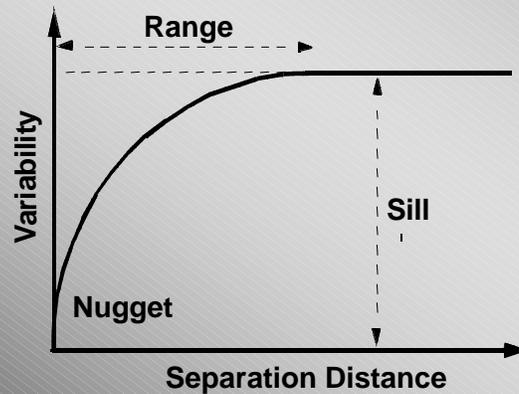
- *First order stationarity* refers to the mean value remaining constant in space
- *Second order stationarity* refers to the variance being constant in space

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 variance being constant in space

- *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.*



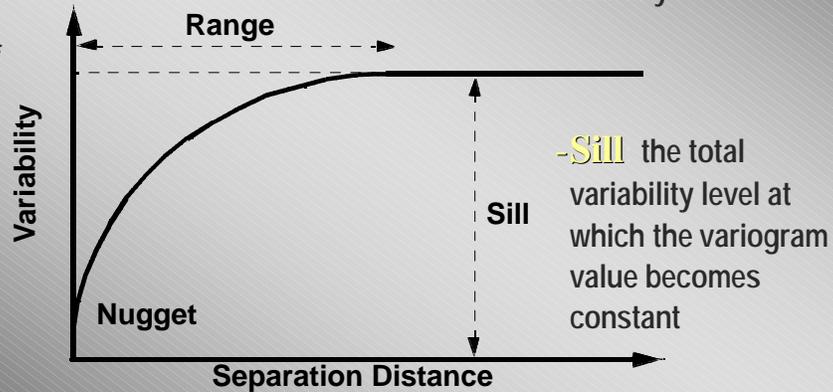
Smart Sampling

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

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

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

Where: g 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 .

Mound Accelerated Site Technology Deployment

3-9

We look at a value at x and a value that is h away, take the difference and square it, then take $\frac{1}{2}$ the average. If you look at all pairs of x and $x+h$, h is defined, then sum up the squares of the differences and divide by $2n$.

There is some tolerance on the value of h .

A specified h , e.g. 10 ft, does not mean only the points that are exactly 10.0 ft apart. Instead it means all of the pairs that are between 5 and 15 ft away from each other.

$h = 10$ means 5 - 15
 $h = 20$ means 15-25
 $h = 30$ means 25-35

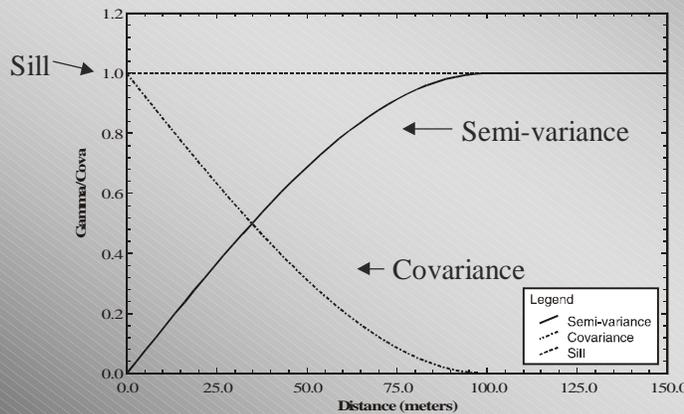
Note that everything is pair-wise.

The formula for the variogram looks like the formula for variance except here we are dealing with differences in values separated by a distance, h , not differences between each point and the mean of the distribution

If the values at x and $x+h$ are the same, the variability is zero. If some conditions are met (2nd order stationarity) then the equation applies.

At any one lag distance h there are n number of pairs, and you are integrating over the number of separation differences, N_h (the total number of h values considered). For a given value of h , find the difference between a pair of points ($z(x)-z(x+h)$) and square it to get rid of negatives (this also exaggerates the data, small values become smaller and large values become huge), divide by the number of pairs to get the average and halve that result. This gives a value for variability, a point on the variogram. Repeat this for each value of h .

$$C(h) = \text{Sill} - g(h)$$



Covariance is the inverse of the variogram

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

Mound Accelerated Site Technology Deployment

3-10

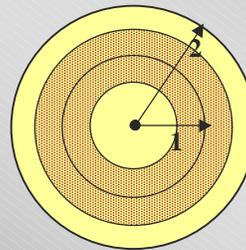
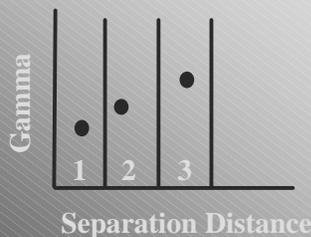
Generally, we deal with the variogram, but mathematically it is the covariance (correlation) used in solving the equations. Use the variogram when dealing with the samples, but realize that the software will use the covariance for estimation and simulation problems.

The covariance is equal to the sill value minus the variability. The covariance is the correlation. If we can make this assumption, then it means that variability is not a function of location, gamma is only a function of separation distance with no dependence on location.

Smart Sampling

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.



Mound Accelerated Site Technology Deployment

3-11

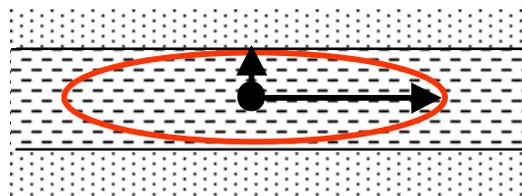
To determine how many samples are a given h away from a certain location, a search neighborhood is used.

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.

- Because of anisotropic deposition, rather than using a circular search neighborhood, you may want to use an ellipse oriented along the principle direction of correlation.

- For example in sedimentary layers there are changes vertically in types of rock and large sample variations, horizontally the beds are very similar even at large distances. Need to consider that things are fairly continuous along the bedding and consider that they are fairly variable across the bedding.

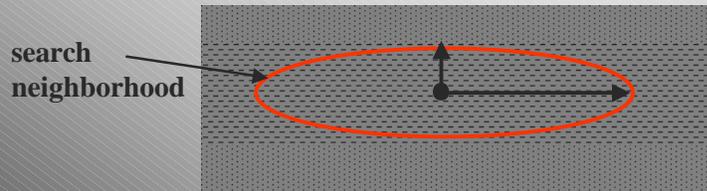


Smart Sampling

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*

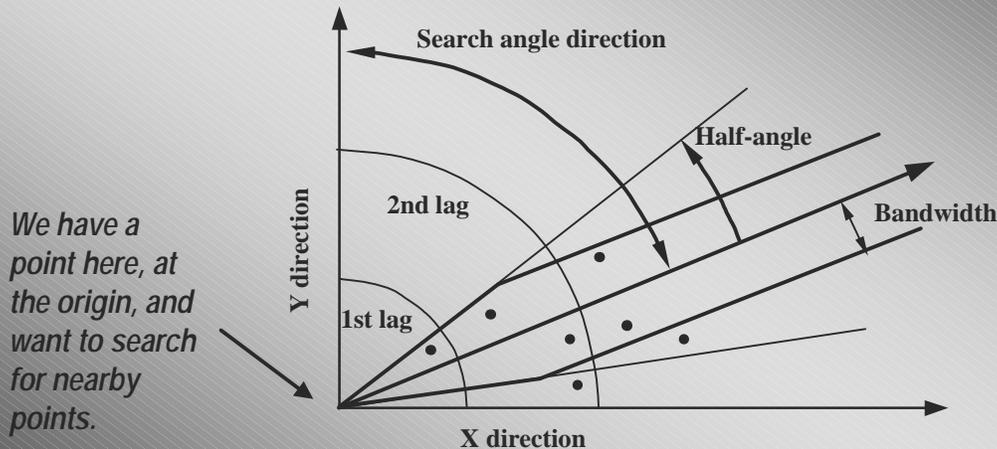


Mound Accelerated Site Technology Deployment

3-12

When dealing with sedimentary layers, you need to consider that things are fairly continuous along the bedding and fairly variable across the bedding.

Use knowledge of data deposition to customize search along a preferred orientation. Orient search along this direction, the *search direction*.



Mound Accelerated Site Technology Deployment

3-13

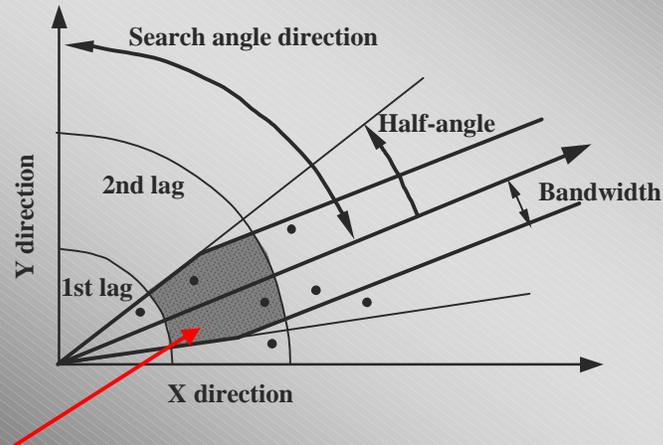
The search neighborhood diagram is a template which can be moved to different points and different directions. Use your knowledge of data deposition to customize search along a preferred orientation, the *search direction*.

Use a *bandwidth* to constrain the search and prevent searching outside the boundaries of the band of relevant data around the preferential direction. The *bandwidth* should not be bigger than the *lag*, because the search area would become too broad without a well defined search direction.

Only count the points that fall in the template within the bandwidth. The tighter and more focused the bandwidth, the fewer pairs found in that direction. You need to have a wide enough bandwidth to get a statistically valuable number of pairs. Rule of thumb is 30 pairs in each lag distance.

Control the rate at which the search neighborhood reaches the *bandwidth* with the *half angle*. The *half-angle* is unrelated to the *search direction*. There is a separate variogram for each direction; within the angle and bandwidth, one point on the variogram will be generated for each lag.

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

Mound Accelerated Site Technology Deployment

3-14

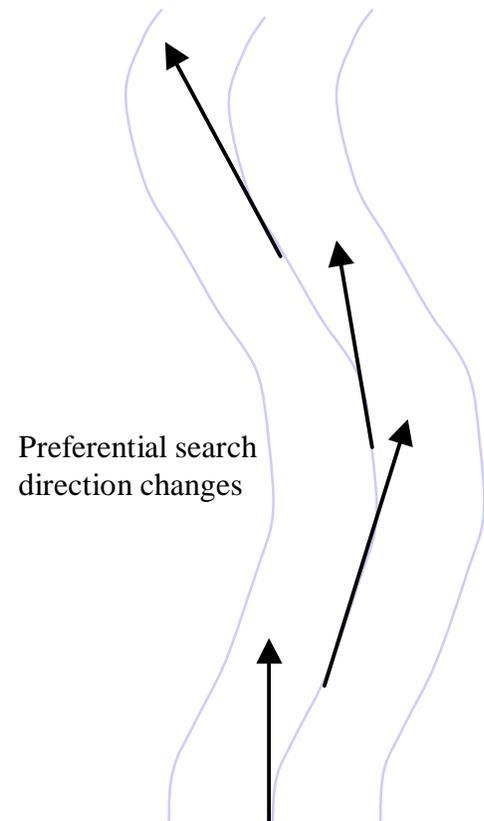
Only count the points that fall in the template (within the bandwidth).

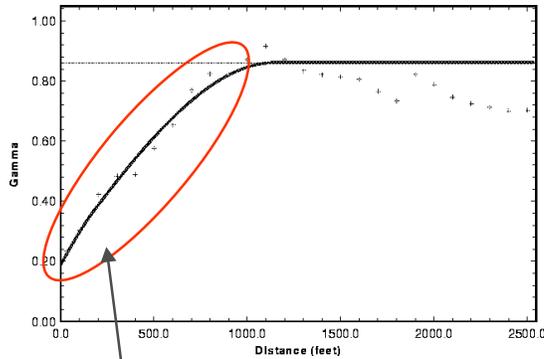
The tighter and more focused the bandwidth, the fewer pairs found in that direction.

In geostatistical practice, the search angle is constant throughout the domain, which may or may not be desirable.

For example: how to construct a variogram on grain sizes in a river channel deposit when the channel is not straight.

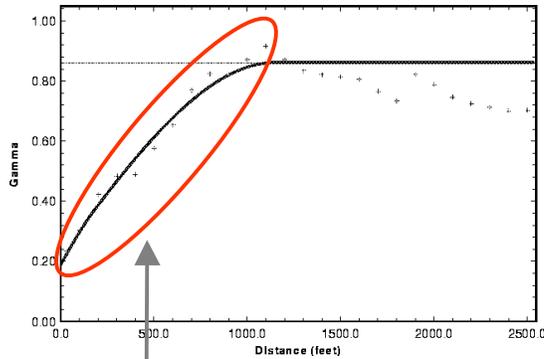
Cannot adapt the search angle to follow the curves of the channel because h is a vector, specified over entire domain.





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

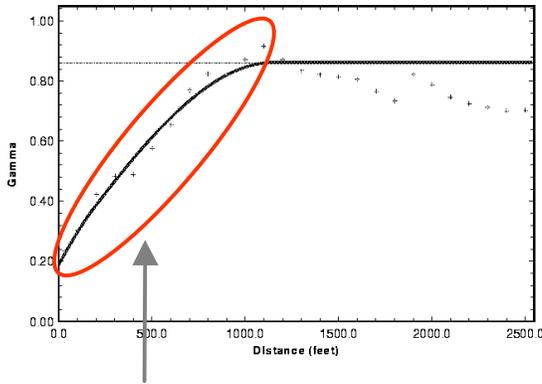
Typically the experimental variogram starts at small values, increases, then follows, on average, the sill. In modeling, the emphasis is placed on fitting the experimental variogram prior to the sill.



Plot the gamma value for each lag and fit a model to the points.

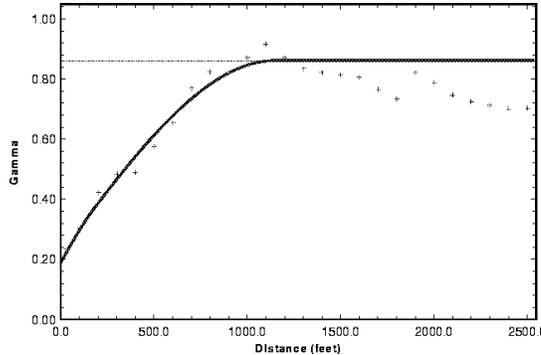
Algorithms require an analytical expression to fit the experimental variogram

*The points are the experimental variogram.
The variogram model is shown by the line.*



Model must yield a positive definite determinant.

'For an estimate that is a weighted linear combination of other values, the variance about that estimate must be ≥ 0 '
(After Isaaks and Srivastava, 1989)



Generally, stick with 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.

The experimental variogram (shown by red crosses) shows values for each lag (x-axis) and average gamma value.

Fit a model to the points (the points are the empirical estimate, the theoretical estimate is the line)... The theoretical model is required for simulation so that you can calculate values from the formula for the covariance. Algorithms require an analytical expression to fit the experimental variogram.

Model must yield a *positive definite* determinate.

Need to emphasize the variogram at smaller h, before it hits the sill, because that's where you're working; it's the part of the variogram that the model needs to fit best.

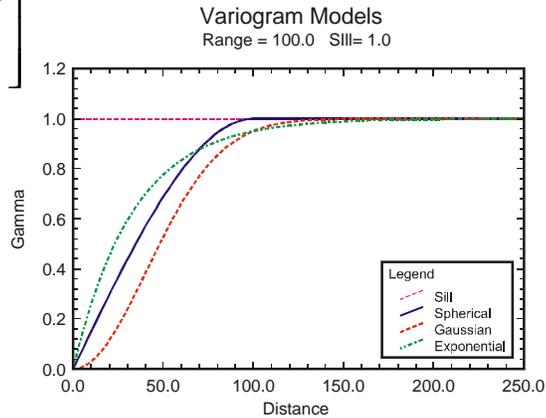
At this point in the process, you're trying to figure out if the data is strongly correlated or not. The fewer data you have, the more you have to take by faith.

Least squares efforts are not productive. Trying to look at correlation.

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

$$h \geq a: \gamma(h) = C$$

Where **C** = sill value
a = range
h = lag distance



The spherical model tends to be linear at the origin

Where **C** = sill value
a = range
h = lag distance

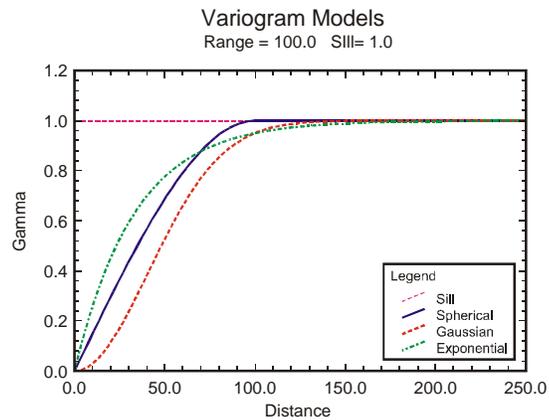
Spherical: Tends to be linear at origin

Gaussian: Allows modeling when variability increases very slowly with increasing h

Exponential: Constantly increasing level of variability and an asymptotic approach to the sill

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

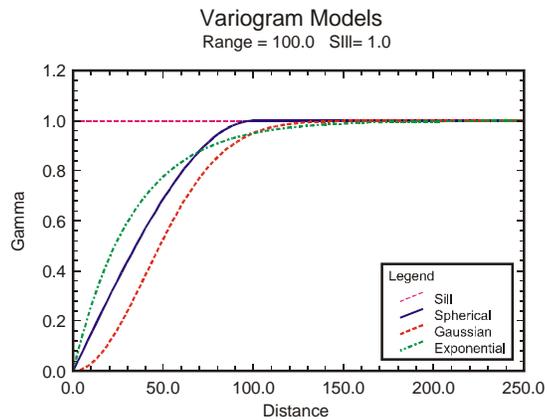
Where **C** = sill value
a = range
h = lag distance



The Gaussian model allows modeling when variability increases very slowly with increasing h .

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

Where **C** = sill value
a = range
h = lag distance

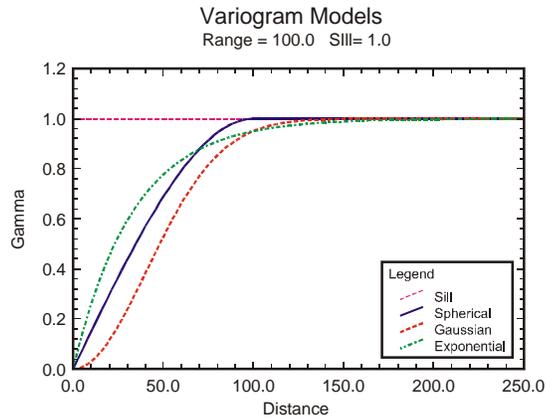


The Exponential model displays a constantly increasing level of variability and an asymptotic approach to the sill

May not be able to tell which model is the exact fit. Typically experimental variograms show variability, so exact choice of model may be critical.

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.

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



- *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*

$$C(h)_{\text{total}} = C_1(h) + C_2(h) + \dots + C_n(h)$$

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

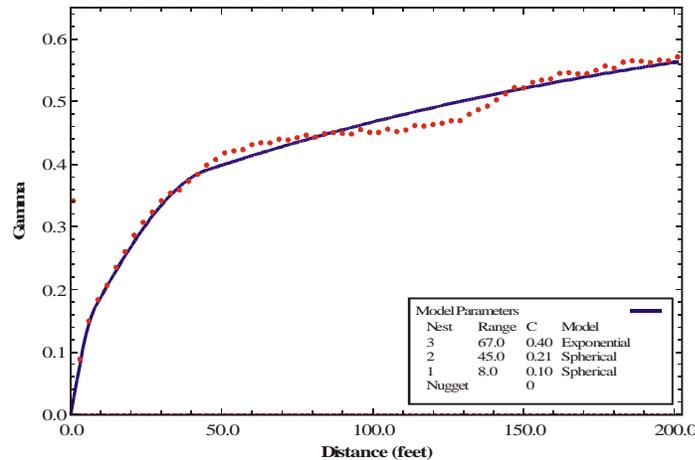
- *Combinations of different spherical, Gaussian, or exponential models give a legitimate model with a positive definite covariance matrix.*

Combinations of different spherical, Gaussian, or exponential models give a legitimate model with a positive definite covariance matrix.

Models can be added together to capture a particular curve that you may want to interpret

Example shows three nested models:

**two spherical
and
one exponential.**



Mound Accelerated Site Technology Deployment

3-24

Below sample spacing, fitting the nugget is interpretation. Is it best to have large nugget or to add another structure with a short range?

(need to mark on graph where the breaks are between models) - Maya
take out dot on vertical axis, it's meaningless (at gamma = .35) - Maya

- *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: *Constant Sill, Range changes with direction*

Software almost always requires anisotropy to be geometric.

Zonal: *Constant Range, Sill changes with direction*

The level of variability is different in different directions.

Zonal anisotropy is more complex than geometric anisotropy, and requires a few “tricks” to implement

Geometric Anisotropy is typically dealt with by simply transforming the coordinates.

Zonal Anisotropy usually means the data is too sparse, more sampling would eliminate the anisotropy. It is more complex and requires a few “tricks” to implement

Software almost always requires anisotropy to be geometric. Each model within a nested model may have it’s own anisotropy, so can get very complicated. Really don’t need to worry about subtleties.

- **Look for anisotropy with a variogram map.**

Anisotropy in the variable requires fine tuning of the search neighborhood.

Geometric anisotropy

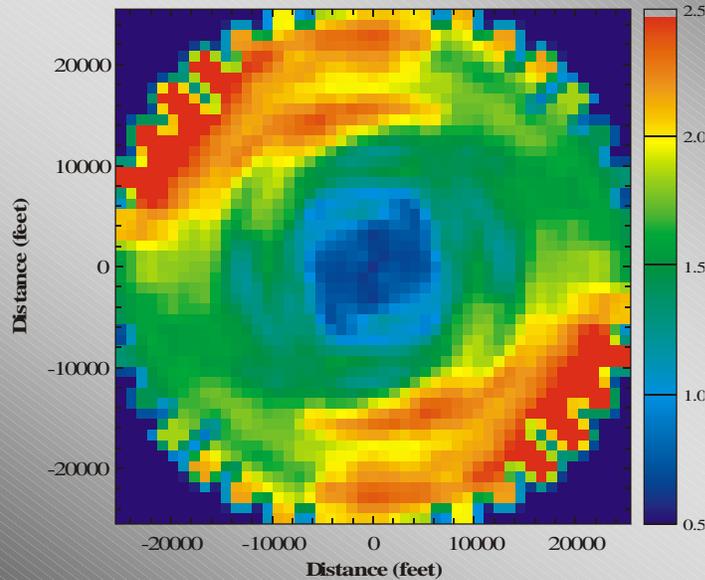
is typically dealt with by transforming the coordinates

Zonal anisotropy

usually means data are too sparse, more sampling would eliminate the anisotropy

Each model within a nested model may have it's own anisotropy.

Can get very complicated, but can often ignore subtleties



Mound Accelerated Site Technology Deployment

3-27

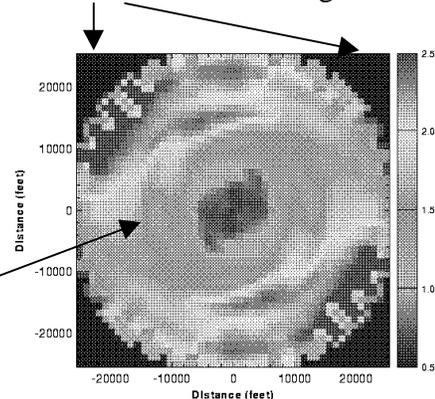
One way to look for anisotropy is with a variogram map

While keeping a small bandwidth and changing the search direction, calculate the variogram around a circle. Plot the variogram values (color scale) in all directions to get a variogram map. The map shows the general direction of strongest correlation. A donut shape means the variogram is isotropic.

In this map, the smallest ranges (least correlation) occur in the SE-NW direction. The greatest amount of correlation is in the NE-SW direction.

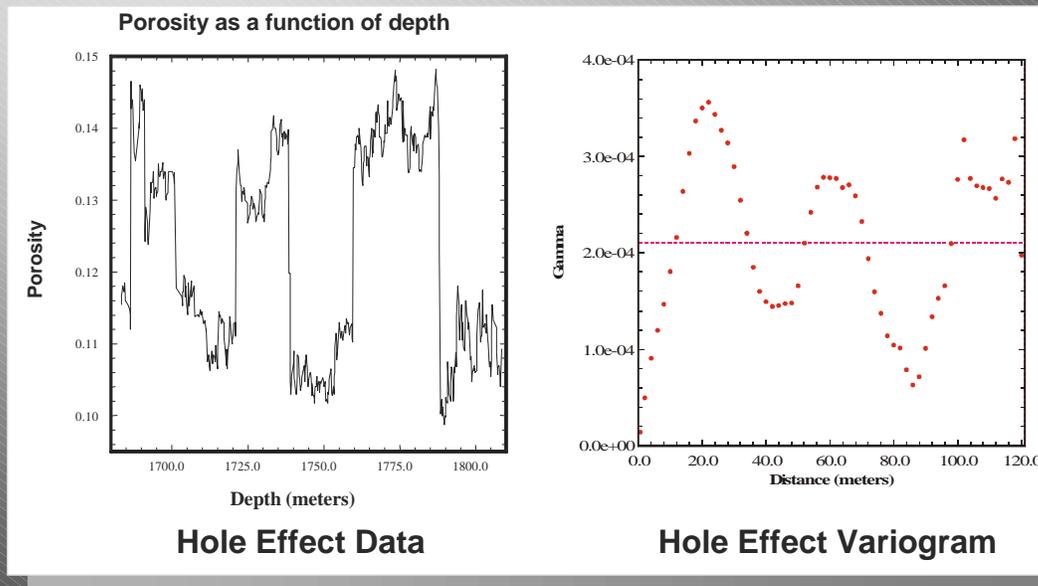
When modeling, only the major and minor directions of correlation are used to define the variogram space. You must have enough data to look at every possible direction and distance combination. Often this cannot be calculated.

The corners are out of range of the data



Should be mirror images along each axis because looking east from one sample to another is the same as looking west from the other sample.

Smoothing of the data may cause what appears to be areas where the variogram value rises and then drops.



Mound Accelerated Site Technology Deployment

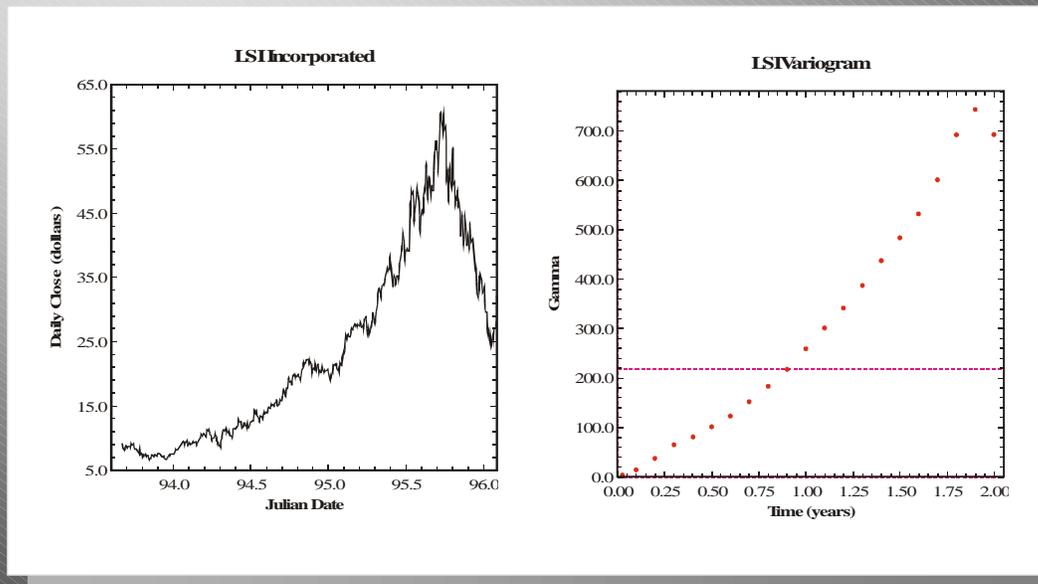
3-28

The data plot above shows porosity as a function of depth, drilling through alternating layers of sandstone and shale.

If you compute a variogram across those separations without respect to whether you are keeping it within a certain rock type, you can get a variogram with periodicity.

With this sort of data set, you start at small separation (h) typically staying within the rock type and have small variability. As the separation (h) increases, you are comparing two different rock types and get a very high variability. When the h increases enough that you are comparing the same rock type in two different layers, the variability drops down, etc.

Typically easiest way to get rid of the hole effect is by working only in same rock types. Otherwise, you can fit complicated models to it.



Mound Accelerated Site Technology Deployment

3-29

This data shows a constantly increasing stock price that eventually crashes.

The variogram is typical of data that has a trend: data are non-stationary, the mean changes depending on location. A trend in the data creates a parabolic-type increase and never reaches a sill. You cannot fit a model to it.

To get rid of trend effect, fit a curve to the data and then calculate and model the variograms using the residuals.