INTRODUCTION TO GEOSTATISTICS
And
VARIOGRAM ANALYSIS

C&PE 940, 17 October 2005

Geoff Bohling
Assistant Scientist
Kansas Geological Survey
geoff@kgs.ku.edu
864-2093

Overheads and other resources available at:

http://people.ku.edu/~gbohling/cpe940

What is Geostatistics?

“Geostatistics: study of phenomena that vary in space and/or time” (Deutsch, 2002)

“Geostatistics can be regarded as a collection of numerical techniques that deal with the characterization of spatial attributes, employing primarily random models in a manner similar to the way in which time series analysis characterizes temporal data.” (Olea, 1999)

“Geostatistics offers a way of describing the spatial continuity of natural phenomena and provides adaptations of classical regression techniques to take advantage of this continuity.” (Isaaks and Srivastava, 1989)

Geostatistics deals with spatially *autocorrelated* data.

Autocorrelation: correlation between elements of a series and others from the same series separated from them by a given interval. (Oxford American Dictionary)

Some spatially autocorrelated parameters of interest to reservoir engineers: facies, reservoir thickness, porosity, permeability

Some Geostatistics Textbooks

C.V. Deutsch, 2002, *Geostatistical Reservoir Modeling*, Oxford University Press, 376 pages.

- Focuses specifically on modeling of facies, porosity, and permeability for reservoir simulation.

C.V. Deutsch and A.G. Journel, 1998, *GSLIB: Geostatistical Software Library and User's Guide, Second Edition*, Oxford University Press, 369 pages.

- Owner's manual for the GSLIB software library; serves as a standard reference for concepts and terminology.

P. Goovaerts, 1997, *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 483 pages.

- A nice introduction with examples focused on an environmental chemistry dataset; includes more advanced topics like factorial kriging.

E.H. Isaaks and R.M. Srivastava, 1989, *An Introduction to Applied Geostatistics*, Oxford University Press, 561 pages.

- Probably the best introductory geostatistics textbook; intuitive development of concepts from first principles with clear examples at every step.

P.K. Kitanidis, 1997, *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press, 249 pages.

- A somewhat different take, with a focus on generalized covariance functions; includes discussion of geostatistical inversion of (groundwater) flow models.

M. Kelkar and G. Perez, 2002, *Applied Geostatistics for Reservoir Characterization*, Society of Petroleum Engineers Inc., 264 pages.

- Covers much the same territory as Deutsch's 2002 book; jam-packed with figures illustrating concepts.

R.A. Olea, 1999, *Geostatistics for Engineers and Earth Scientists*, Kluwer Academic Publishers, 303 pages.

- Step by step mathematical development of key concepts, with clearly documented numerical examples.

Links to some software and online resources are available at

http://people.ku.edu/~gbohling/geostats

Basic Components of Geostatistics

(Semi)variogram analysis – characterization of spatial correlation

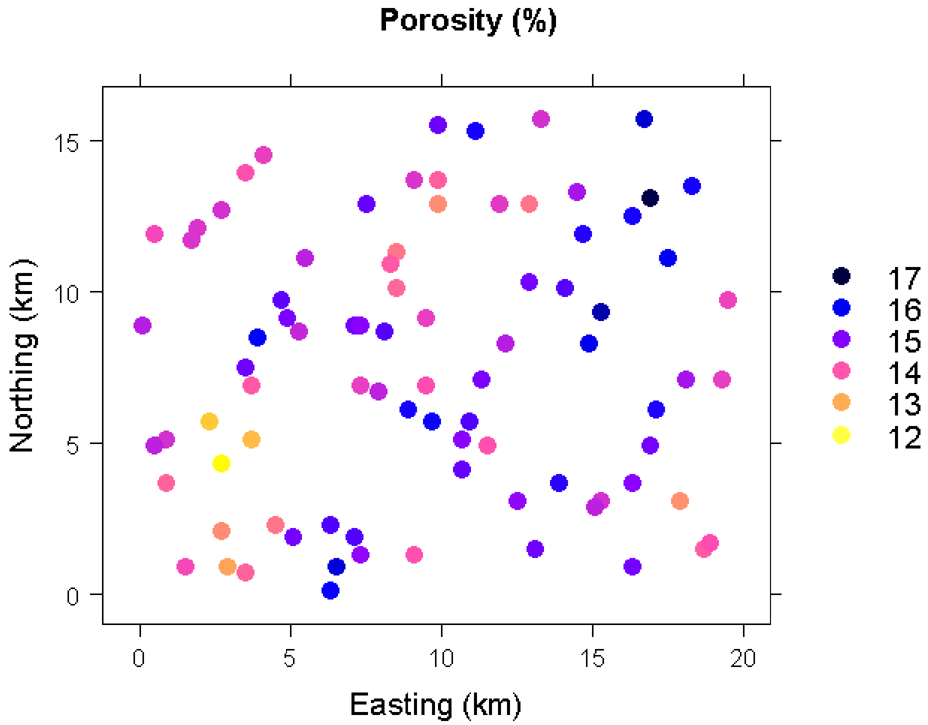
Kriging – optimal interpolation; generates best linear unbiased estimate at each location; employs semivariogram model

Stochastic simulation – generation of multiple equiprobable images of the variable; also employs semivariogram model

Geostatistical routines are implemented in the major reservoir modeling packages like Petrel and Roxar’s Irap RMS; used in the generation of grids of facies, permeability, porosity, etc. for the reservoir.

Exploratory Analysis of Example Data

Our example data consist of vertically averaged porosity values, in percent, in Zone A of the Big Bean Field (fictitious, but based on data from a real field). Porosity values are available from 85 wells distributed throughout the field, which is approximately 20 km in east-west extent and 16 km north-south. The porosities range from 12% to 17%. Here are the data values posted at the well locations:

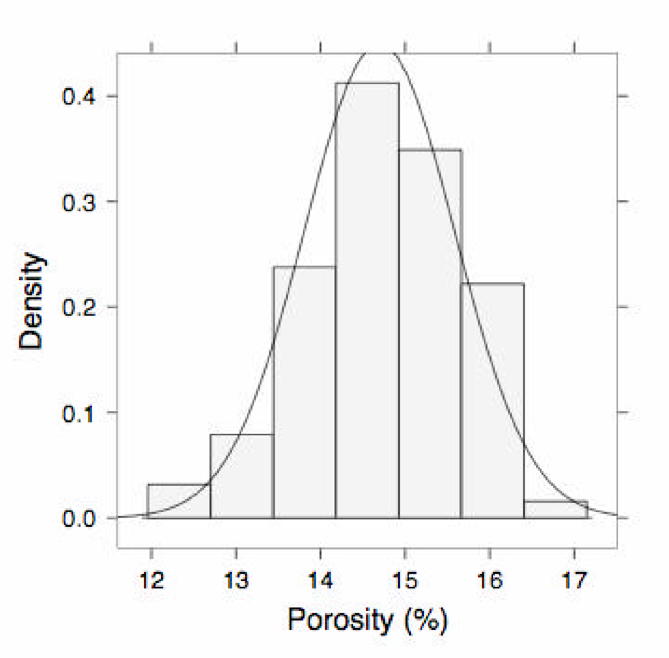


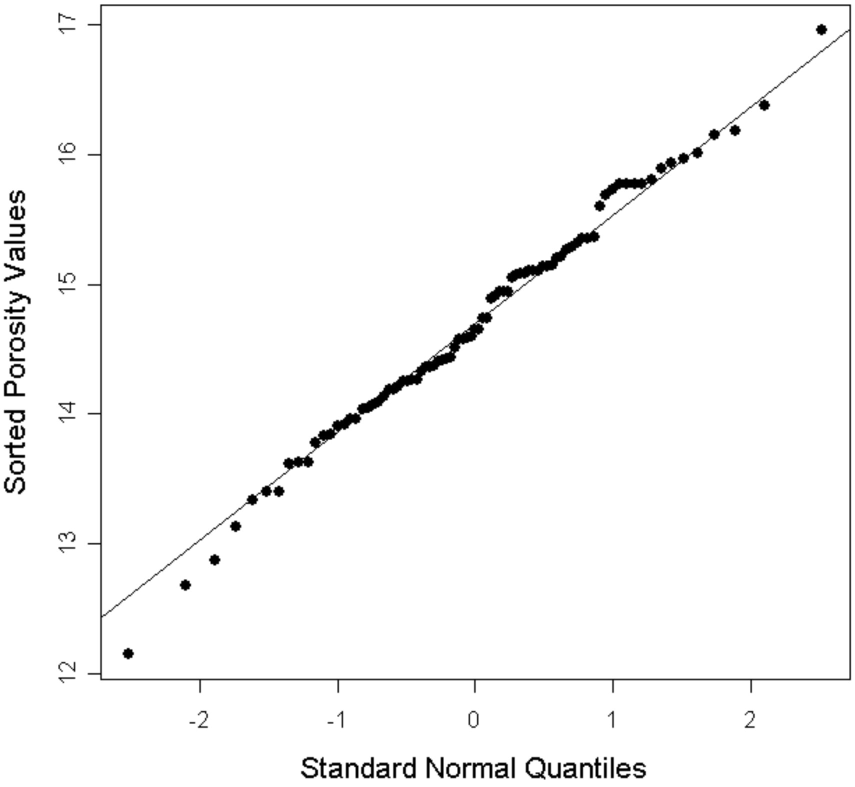
Geostatistical methods are optimal when data are

- normally distributed and
- stationary (mean and variance do not vary significantly in space)

Significant deviations from normality and stationarity can cause problems, so it is always best to begin by looking at a histogram or similar plot to check for normality and a posting of the data values in space to check for significant trends. The posting above shows some hint of a SW-NE trend, which we will check later.

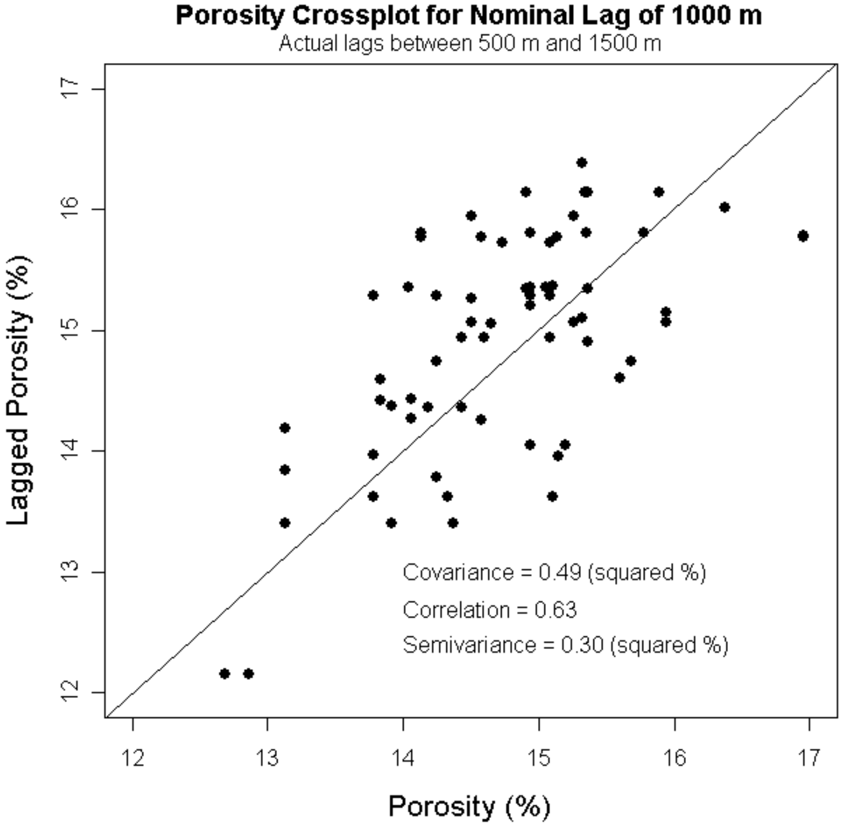
Looking at the histogram (with a normal density superimposed) and a normal quantile-quantile plot shows that the porosity distribution does not deviate too severely from normality:





Spatial Covariance, Correlation and Semivariance

You have already learned that covariance and correlation are measures of the similarity between two different variables. To extend these to measures of spatial similarity, consider a scatterplot where the data pairs represent measurements of the same variable made some distance apart from each other. The separation distance is usually referred to as “lag”, as used in time series analysis. We’ll refer to the values plotted on the vertical axis as the lagged variable, although the decision as to which axis represents the lagged values is somewhat arbitrary. Here is a scatterplot of porosity values at wells separated by a nominal lag of 1000 m:



Because of the irregular distribution of wells, we cannot expect to find many pairs of data values separated by exactly 1000 m, if we find any at all. Here we have introduced a “lag tolerance” of 500 m, pooling the data pairs with separation distances between 500 and 1500 m in order to get a reasonable number of pairs for computing statistics. The actual lags for the data pairs shown in the crossplot range from 566 m to 1456 m, with a mean lag of 1129 m.

The three statistics shown on the crossplot are the covariance, correlation, and semivariance between the porosity values on the horizontal axis and the lagged porosity values on the vertical axis. To formalize the definition of these statistics, we need to introduce some notation. Following standard geostatistical practice, we’ll use:

$\mathbf{u}$: vector of spatial coordinates (with components x, y or “easting” and “northing” for our 2D example)

$z(\mathbf{u})$: variable under consideration as a function of spatial location (porosity in this example)

$\mathbf{h}$: lag vector representing separation between two spatial locations

$z(\mathbf{u}+\mathbf{h})$: lagged version of variable under consideration

Sometimes $z(\mathbf{u})$ will be referred to as the “tail” variable and $z(\mathbf{u}+\mathbf{h})$ will be referred to as the “head” variable, since we can think of them as being located at the tail and head of the lag vector, $\mathbf{h}$. The scatterplot of tail versus head values for a certain lag, $\mathbf{h}$, is often called an **h**-scattergram.

Now, with $N(\mathbf{h})$ representing the number of pairs separated by lag $\mathbf{h}$ (plus or minus the lag tolerance), we can compute the statistics for lag $\mathbf{h}$ as

Covariance: $$C(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{a=1}^{N(\mathbf{h})} z(\mathbf{u}_a) \cdot z(\mathbf{u}_a + \mathbf{h}) - m_{\mathbf{0}} \cdot m_{+\mathbf{h}}$$

Correlation: $$r(\mathbf{h}) = \frac{C(\mathbf{h})}{\sqrt{s_{\mathbf{0}} \cdot s_{+\mathbf{h}}}}$$

Semivariance: $$g(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{a=1}^{N(\mathbf{h})} \left[z(\mathbf{u}_a + \mathbf{h}) - z(\mathbf{u}_a) \right]^2$$

where $m_{\mathbf{0}}$ and $m_{+\mathbf{h}}$ are the means of the tail and head values:

$$m_{\mathbf{0}} = \frac{1}{N(\mathbf{h})} \sum_{a=1}^{N(\mathbf{h})} z(\mathbf{u}_a) \qquad m_{+\mathbf{h}} = \frac{1}{N(\mathbf{h})} \sum_{a=1}^{N(\mathbf{h})} z(\mathbf{u}_a + \mathbf{h})$$

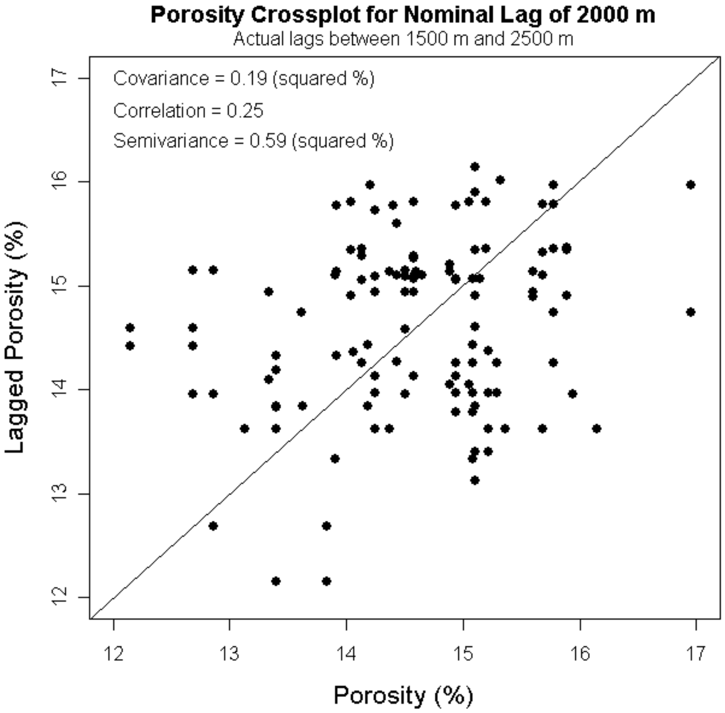
and $s_{\mathbf{0}}$ and $s_{+\mathbf{h}}$ are the corresponding standard deviations:

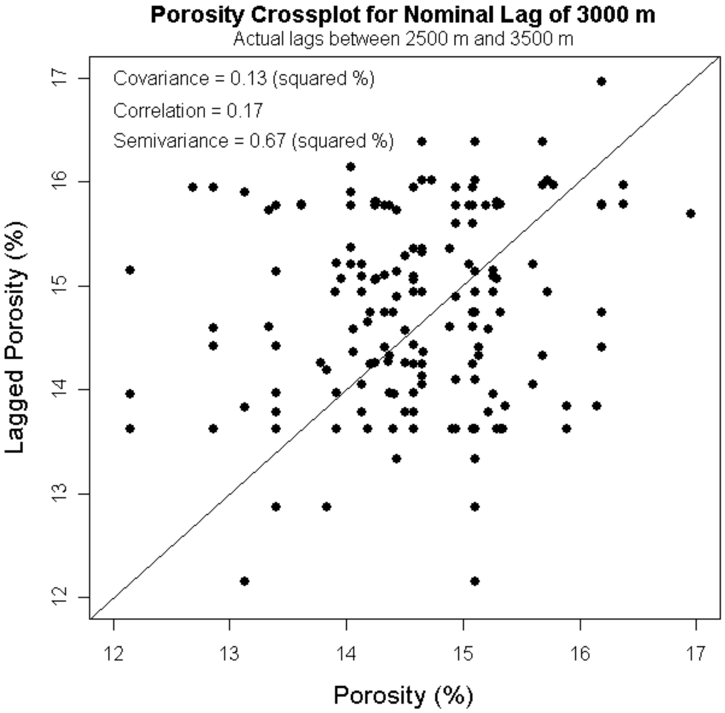
$$s_{\mathbf{0}} = \frac{1}{N(\mathbf{h})} \sum_{a=1}^{N(\mathbf{h})} \left[z(\mathbf{u}_a) - m_{\mathbf{0}} \right]^2 \qquad s_{+\mathbf{h}} = \frac{1}{N(\mathbf{h})} \sum_{a=1}^{N(\mathbf{h})} \left[z(\mathbf{u}_a + \mathbf{h}) - m_{+\mathbf{h}} \right]^2 .$$

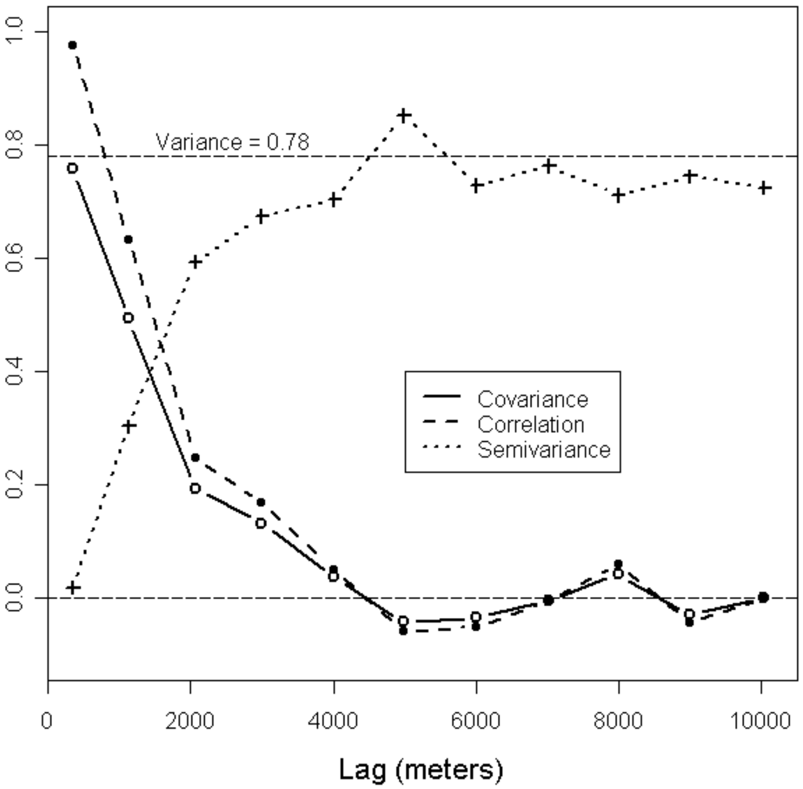
Note that these definitions for $C(\mathbf{h})$ and $g(\mathbf{h})$ use $N(\mathbf{h})$ in the denominator, not $N(\mathbf{h}) - 1$.

The semivariance is the moment of inertia or spread of the **h**-scattergram about the 45° (1 to 1) line shown on the plot. Covariance and correlation are both measures of the similarity of the head and tail values. Semivariance is a measure of the dissimilarity.

Here are the **h**-scatterplots for nominal lags of 2000 m and 3000 m. Note that the covariance and correlation decrease and the semivariance increases with increasing separation distance.







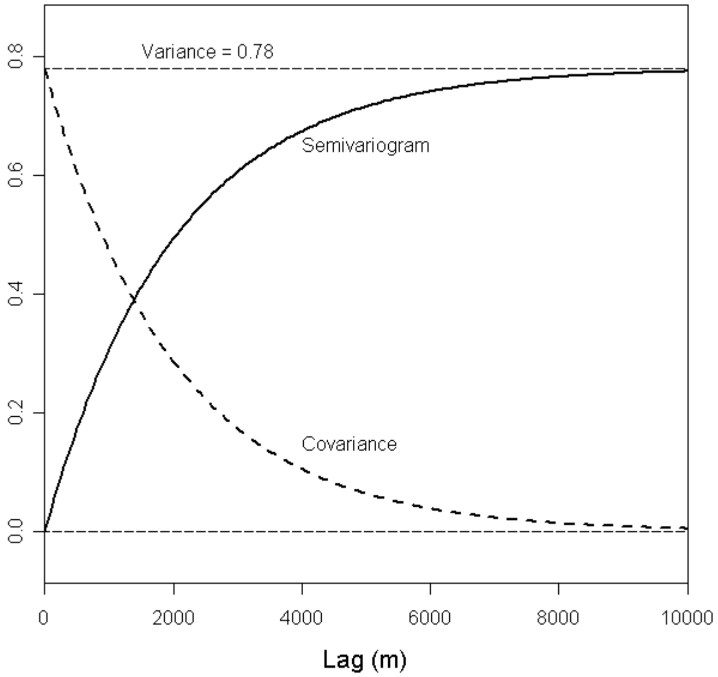
The plot above shows all three statistics versus actual mean lag for the contributing data pairs at each lag. The shortest lag shown (the nominally “zero” lag) includes six data pairs with a mean lag of 351 m. The correlation versus lag is referred to as the *correlogram* and the semivariance versus lag is the *semivariogram.* The covariance versus lag is generally just referred to as the covariance function.

The empirical functions that we have plotted – computed from the sample data – are of course just estimators of the theoretical functions $C(\mathbf{h})$, $r(\mathbf{h})$, and $g(\mathbf{h})$, which can be thought of as population parameters. Estimating these functions based on irregularly distributed data (the usual case) can be very tricky due to the need to pool data pairs into lag bins. Larger lag spacings and tolerances allow more data pairs for estimation but reduce the amount of detail in the semivariogram (or covariance or correlogram). The problem is particularly difficult for the shorter lags, which tend to have very few pairs (six in this example). This is unfortunate, since the behavior near the origin is the most important to characterize.

Under the condition of *second-order stationarity* (spatially constant mean and variance), the covariance function, correlogram, and semivariogram obey the following relationships:

$$C(\mathbf{0}) = \mathrm{Cov}(Z(\mathbf{u}), Z(\mathbf{u})) = \mathrm{Var}(Z(\mathbf{u}))$$
$$r(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0})$$
$$g(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

In words, the lag-zero covariance should be equal to the global variance of the variable under consideration, the correlogram should look like the covariance function scaled by the variance, and the semivariogram should look like the covariance function turned upside down:



In practice, the estimated versions of the functions will violate these relationships to a greater or lesser extent due to sampling limitations and deviations from second-order stationarity.

Unlike time series analysts, who prefer to work with either the covariance function or the correlogram, geostatisticians typically work with the semivariogram. This is primarily because the semivariogram, which averages squared differences of the variable, tends to filter the influence of a spatially varying mean. Also, the semivariogram can be applied whenever the first differences of the variable, $Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h})$, are second-order stationary. This form of stationarity, referred to as the *intrinsic hypothesis*, is a weaker requirement than second-order stationarity of the variable itself, meaning that the semivariogram can be defined in some cases where the covariance function cannot be defined. In particular, the semivariance may keep increasing with increasing lag, rather than leveling off, corresponding to an infinite global variance. In this case the covariance function is undefined.

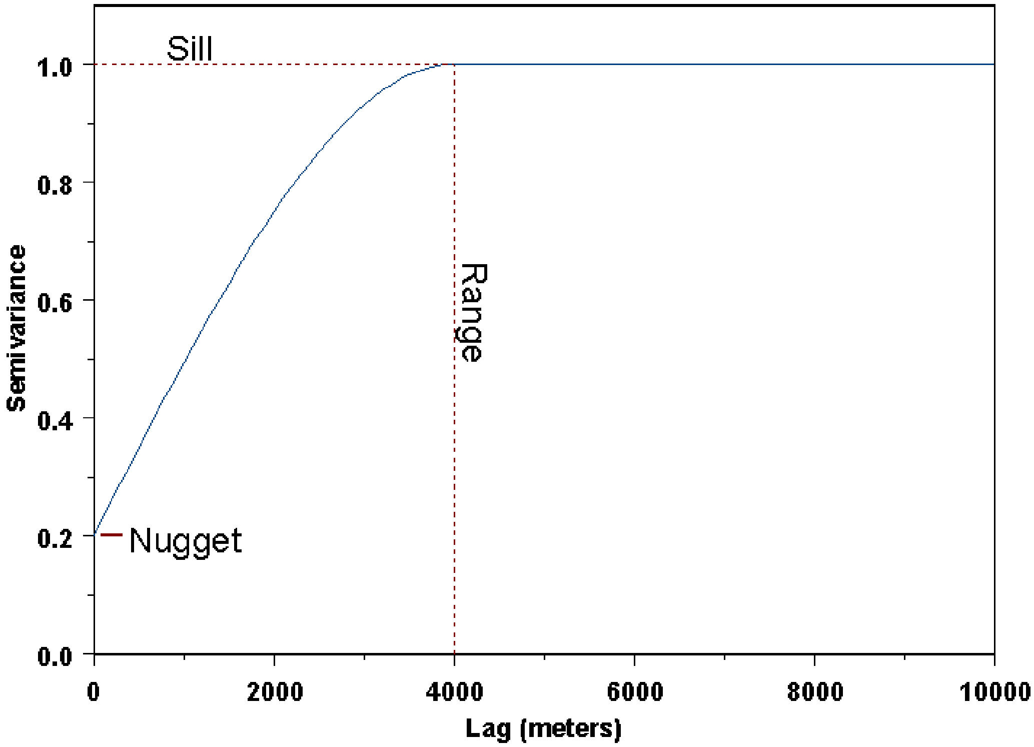
Trend

If the empirical semivariogram continues climbing steadily beyond the global variance value, this is often indicative of a significant spatial trend in the variable, resulting in a negative correlation between variable values separated by large lags. Three options for dealing with lag include:

1) Fit a trend surface and work with residuals from the trend
2) Try to find a “trend-free” direction and use the variogram in that direction as the variogram for the “random” component of the variable (see the section on anisotropy, below)
3) Ignore the problem and use a linear or power variogram

The semivariogram for the porosity data does not seem to indicate a significant trend.

Characteristics of the Semivariogram



Sill: The semivariance value at which the variogram levels off. Also used to refer to the “amplitude” of a certain component of the semivariogram. For the plot above, “sill” could refer to the overall sill (1.0) *or* to the difference (0.8) between the overall sill and the nugget (0.2). Meaning depends on context.

Range: The lag distance at which the semivariogram (or semivariogram component) reaches the sill value. Presumably, autocorrelation is essentially zero beyond the range.

Nugget: In theory the semivariogram value at the origin (0 lag) should be zero. If it is significantly different from zero for lags very close to zero, then this semivariogram value is referred to as the nugget. The nugget represents variability at distances smaller than the typical sample spacing, including measurement error.

Modeling the Semivariogram

For the sake of kriging (or stochastic simulation), we need to replace the empirical semivariogram with an acceptable semivariogram model. Part of the reason for this is that the kriging algorithm will need access to semivariogram values for lag distances other than those used in the empirical semivariogram. More importantly, the semivariogram models used in the kriging process need to obey certain numerical properties in order for the kriging equations to be solvable. (Technically, the semivariogram model needs to be *non-negative definite*, in order the system of kriging equations to be *non-singular*.) Therefore, geostatisticians choose from a palette of acceptable or *licit* semivariogram models.

Using *h* to represent lag distance, *a* to represent (practical) range, and *c* to represent sill, the five most frequently used models are:

Nugget: $$g(h) = \begin{cases} 0 & \text{if } h = 0 \\ c & \text{otherwise} \end{cases}$$

Spherical: $$g(h) = \begin{cases} c \cdot \left(1.5\left(\frac{h}{a}\right) - 0.5\left(\frac{h}{a}\right)^3 \right) & \text{if } h \le a \\ c & \text{otherwise} \end{cases}$$

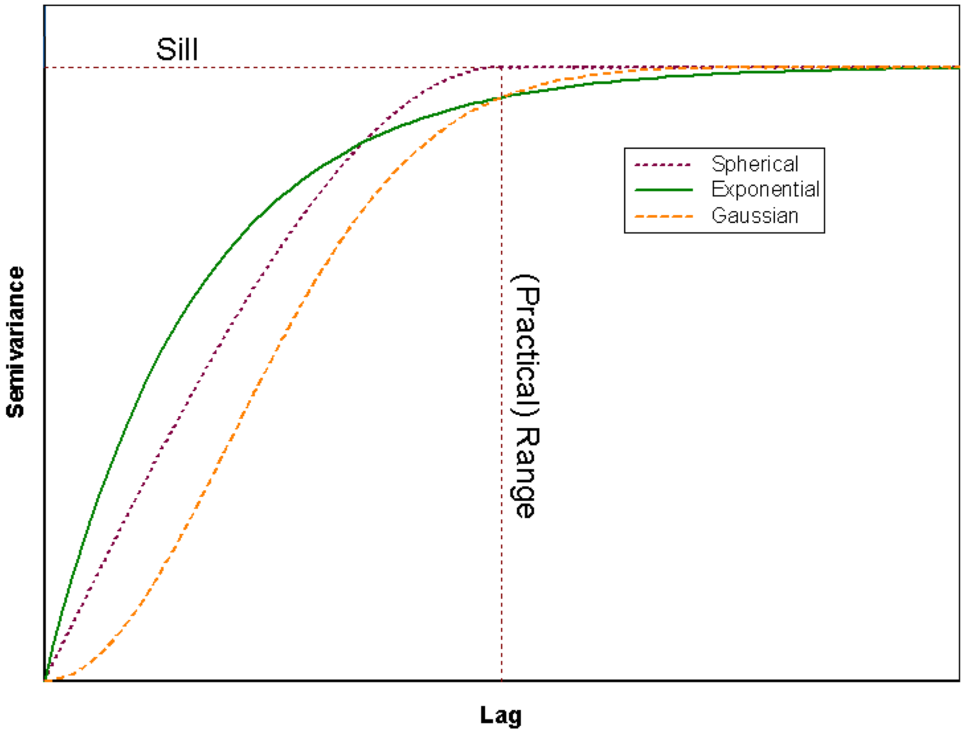
Exponential: $$g(h) = c \cdot \left(1 - \exp\left(\frac{-3h}{a} \right) \right)$$

Gaussian: $$g(h) = c \cdot \left(1 - \exp\left(\frac{-3h^2}{a^2} \right) \right)$$

Power: $$g(h) = c \cdot h^w \quad \textit{with } 0 < w < 2$$

The nugget model represents the discontinuity at the origin due to small-scale variation. On its own it would represent a purely random variable, with no spatial correlation.

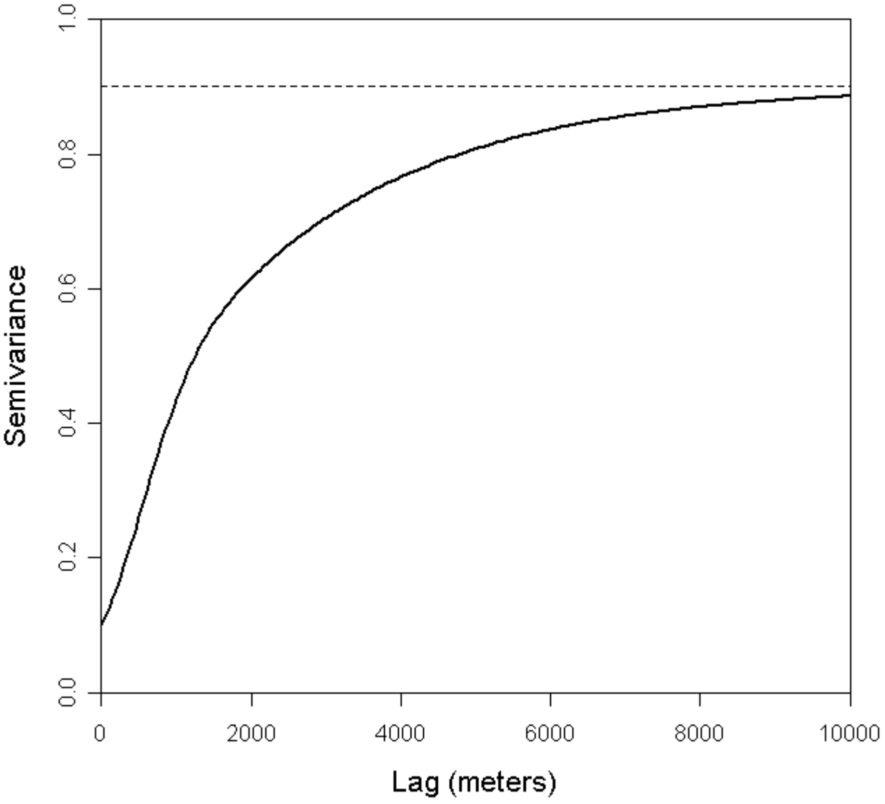
The spherical model actually reaches the specified sill value, *c*, at the specified range, *a*. The exponential and Gaussian approach the sill asymptotically, with *a* representing the practical range, the distance at which the semivariance reaches 95% of the sill value. These three models are shown below:



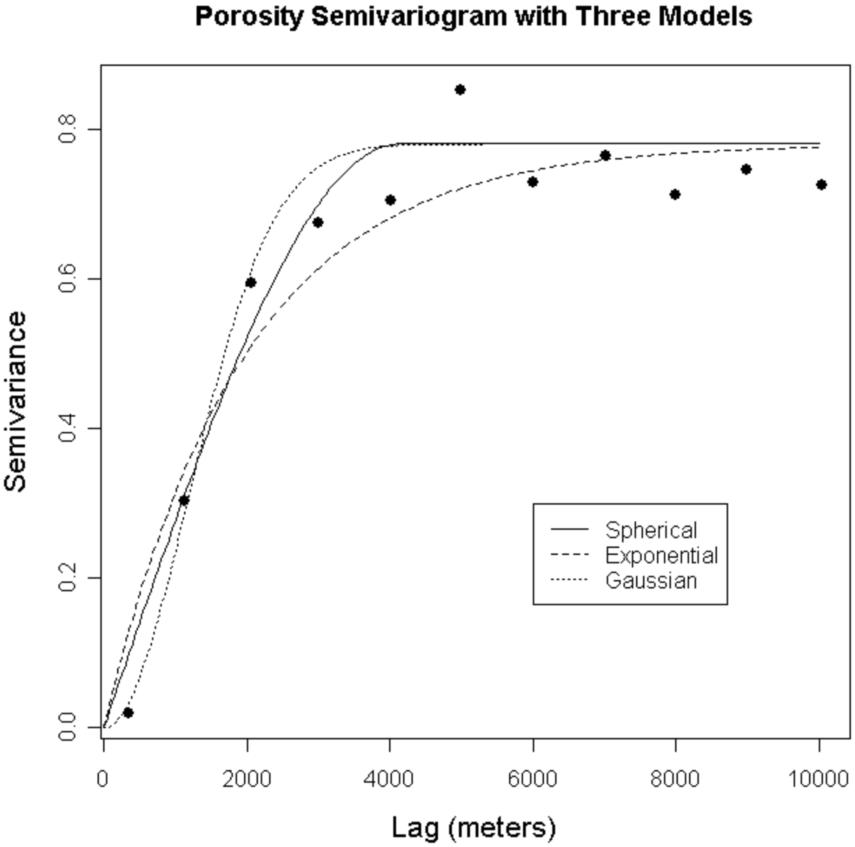
The Gaussian model, with its parabolic behavior at the origin, represents very smoothly varying properties. (However, using the Gaussian model alone without a nugget effect can lead to numerical instabilities in the kriging process.) The spherical and exponential models exhibit linear behavior the origin, appropriate for representing properties with a higher level of short-range variability.

Models with a finite sill, like the Gaussian, exponential, and spherical, are referred to as *transition* models and have corresponding covariance functions given by $\text{cov}(h) = c - g(h)$. The power model does not reach a finite sill and does not have a corresponding covariance function. Power-law semivariogram models are appropriate for properties exhibiting fractal behavior.

Linear combinations of licit semivariogram models are also licit models, so that more complicated models may be built by adding together the basic models described above with different ranges and sills, leading to the second meaning of "sill" discussed above. For example, a variogram model might consist of a nugget of 0.1, a Gaussian component with a range of 1500 m and a sill of 0.2 and an exponential component with a range of 8000 m and sill of 0.6, giving an overall sill of 0.9:



The actual process of fitting a model to an empirical semivariogram is much more of an art than a science, with different authorities suggesting different methods and protocols. Since empirical semivariograms are often quite noisy, quite a bit of subjective judgment goes into selecting a good model. Here is the semivariogram for our example porosity data, with three fitted models. In each case the sill value was fixed at the overall variance of 0.78 and the range was estimated using weighted nonlinear regression (weighting by number of data pairs for each lag):



The fitted ranges for the three models are 4141 m for the spherical, 5823 m for the exponential, and 2884 m for the Gaussian. The Gaussian model gives the best fit, but the spherical is a close second.

Anisotropy

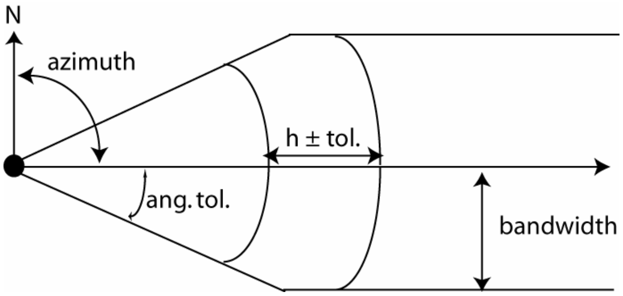
The above discussion has assumed that the spatial correlation structure is the same in all directions, or *isotropic*. In this case the covariance function, correlogram, and semivariogram depend only on the magnitude of the lag vector, $h = |\mathbf{h}|$, and not the direction, and the empirical semivariogram can be computed by pooling data pairs separated by the appropriate distances, regardless of direction. Such a semivariogram is described as *omnidirectional*.

In many cases, however, a property shows different autocorrelation structures in different directions, and an *anisotropic* semivariogram model should be developed to reflect these differences. The most commonly employed model for anisotropy is geometric anisotropy, with the semivariogram reaching the same sill in all directions, but at different ranges. In geological settings, the most prominent form of anisotropy is a strong contrast in ranges in the (stratigraphically) vertical and horizontal directions, with the vertical semivariogram reaching the sill in a much shorter distance than the horizontal semivariogram. In some settings, there may also be significant lateral anisotropy, often reflecting prominent directionality in the depositional setting (such as, along and perpendicular to channels).

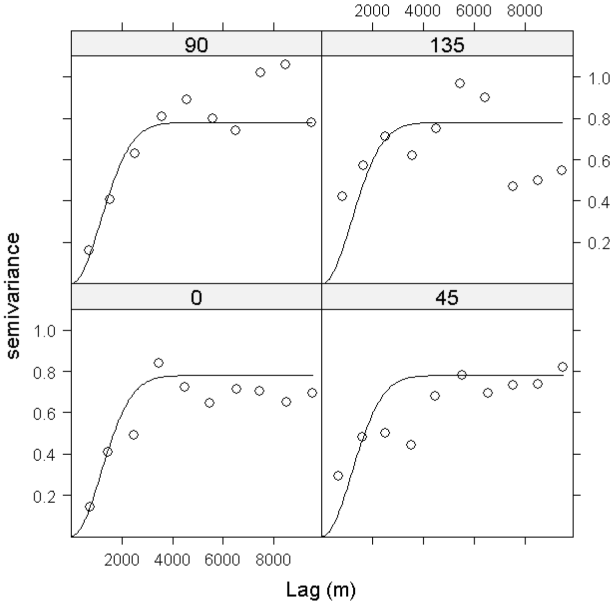
The most common approach to modeling geometric anisotropy is to find ranges, a_x, a_y, and a_z, in three principal, orthogonal directions and transform a three-dimensional lag vector $\mathbf{h} = (h_x, h_y, h_z)$ into an equivalent isotropic lag using:

$$h = \sqrt{(h_x/a_x)^2 + (h_y/a_y)^2 + (h_z/a_z)^2}$$

To check for directional dependence in an empirical semivariogram, we have to compute semivariance values for data pairs falling within certain directional bands as well as falling within the prescribed lag limits. The directional bands are specified by a given azimuthal direction, angular tolerance, and bandwidth:



Here is the porosity semivariogram in the directions N 0° E, N 45° E, N 90° E, and N 135° E with angular tolerance of 22.5 and no bandwidth limit, together with the omnidirectional Gaussian model:



The directional semivariograms are noisier due to the reduced number of data pairs used for estimation. They do not show overwhelming evidence of anisotropy.