

Introduction to Modeling Spatial Processes Using Geostatistical Analyst

*Konstantin Krivoruchko, Ph.D.
Software Development Lead, Geostatistics
kkrivoruchko@esri.com*

Geostatistics is a set of models and tools developed for statistical analysis of continuous data. These data can be measured at any location in space, but they are available in a limited number of sampled points.

Since input data are contaminated by errors and models are only approximations of the reality, predictions made by Geostatistical Analyst are accompanied by information on uncertainties.

The first step in statistical data analysis is to verify three data features: dependency, stationarity, and distribution. If data are independent, it makes little sense to analyze them geostatistically. If data are not stationary, they need to be made so, usually by data detrending and data transformation. Geostatistics works best when input data are Gaussian. If not, data have to be made to be close to Gaussian distribution. Geostatistical Analyst provides exploratory data analysis tools to accomplish these tasks. With information on dependency, stationarity, and distribution you can proceed to the modeling step of the geostatistical data analysis, kriging.

The most important step in kriging is modeling spatial dependency, semivariogram modeling. Geostatistical Analyst provides large choice of semivariogram models and reliable defaults for its parameters.

Geostatistical Analyst provides six kriging models and validation and cross-validation diagnostics for selecting the best model.

Geostatistical Analyst can produce four output maps: prediction, prediction standard errors, probability, and quantile. Each output map produces a different view of the data.

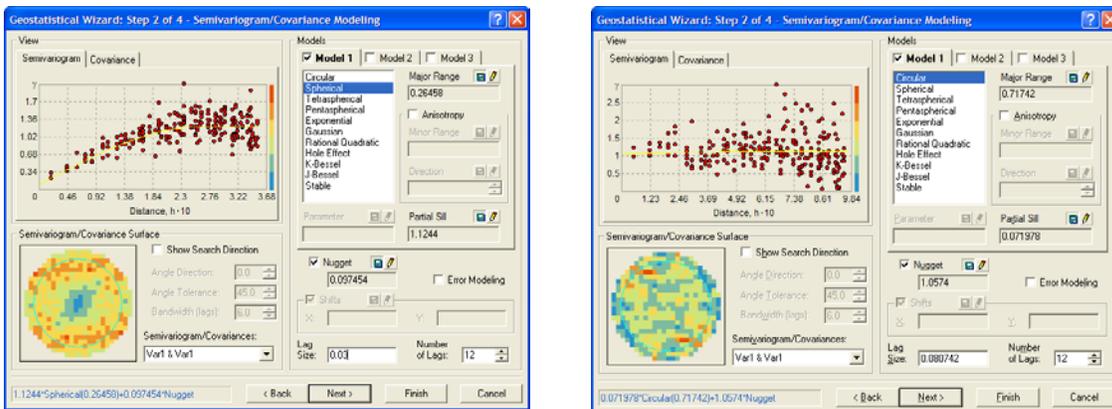
This tutorial only explains basic geostatistical ideas and how to use the geostatistical tools and models in Geostatistical Analyst. Readers interested in a more technical explanation can find all the formulas and more detailed discussions on geostatistics in the [Educational and Research Papers](#) and [Further Reading](#) sections. Several geostatistical case studies are available at [Case Studies](#).

Spatial dependency

Since the goal of geostatistical analysis is to predict values where no data have been collected, the tools and models of Geostatistical Analyst will only work on spatially dependent data. If data are spatially independent, there is no possibility to predict values between them. Even with spatially dependent data, if the dependency is ignored, the result of the analysis will be inadequate as will any decisions based on that analysis.

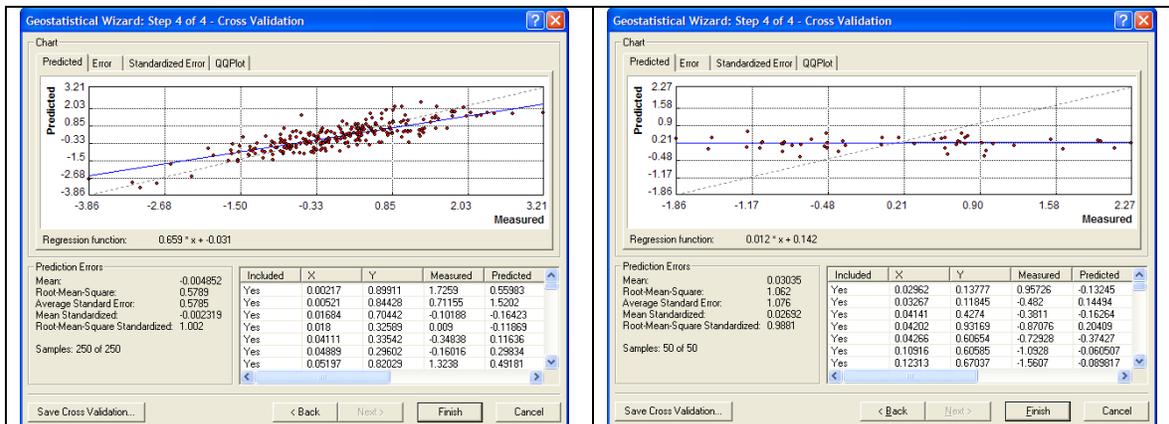
Spatial dependence can be detected using several tools available in the Geostatistical Analyst's Exploratory Spatial Data Analysis (ESDA) and Geostatistical Wizard. Two examples are presented below.

The figure below shows a semivariogram of data with strong spatial dependence, left, and with very weak spatial dependence, right.



In the cross-validation diagnostic, one sample is removed from the dataset, and the value in its location is predicted using information on the remaining observations. Then the same procedure is applied to the second, and third, and so on to the last sample in the database. Measured and predicted values are compared. Comparison of the average difference between predicted and observed values is made.

The cross-validation diagnostics graphs for the datasets used to create the semivariograms above look very different:



If data are correlated, one can be removed and a similar value predicted at that location, left, which is impossible for spatial data with weak dependence: predictions in the figure to the right are approximated by a horizontal line, meaning that prediction for any removed sample is approximately equal to the data's arithmetic average.

Analysis of continuous data using geostatistics

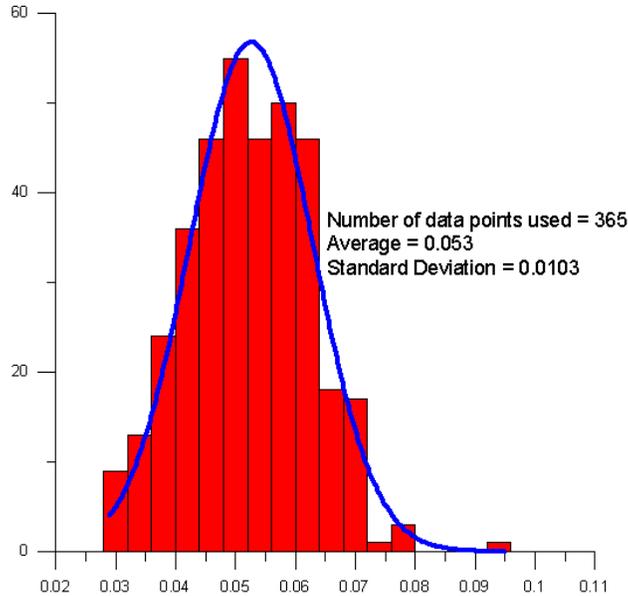
Geostatistics studies data that can be observed at any location, such as temperature. There are data that can be approximated by points on the map, but they cannot be observed at any place. An example is tree locations. If we know the diameters and locations of two nearby trees, prediction of diameter between observed trees does not make sense simply because there is no there. Modeling of discrete data such as tree parameters is a subject of point pattern analysis.

The main goal of the geostatistical analysis is to predict values at the locations where we do not have samples. We would like our predictions to use only neighboring values and to be optimum. We would like to know how much prediction error or uncertainty is in our prediction.

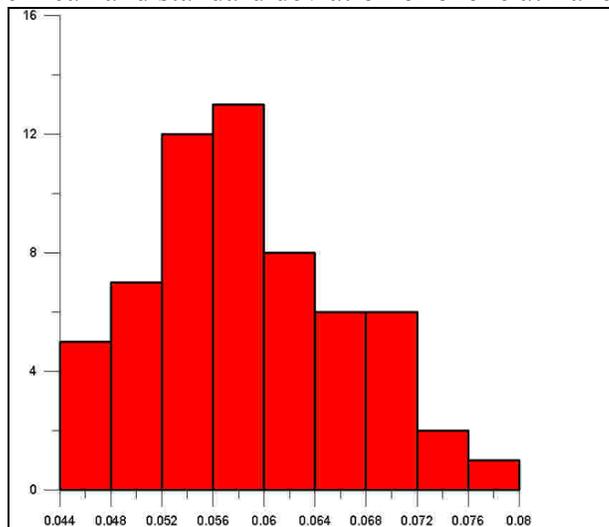
The easiest way to make predictions is to use an average of the local data. But the only kind of data that are equally weighted are independent data. Predictions that use an average are not optimum, and they give over-optimistic estimates of averaging accuracy. Another way is to weight data according to the distance between locations. This idea is implemented in the Geostatistical Analyst's Inverse Distance Weighting model. But this is only consistent with our desire to use neighboring values for prediction. Geostatistical prediction satisfies all over.

Random variable

The figure below shows an example of environmental observations. It is a histogram of the maximum monthly concentration of ozone (parts per million, ppm) at Lake Tahoe, from 1981-1985:



These data can be described by the continuous blue line. This line is called normal distribution in statistics. Assuming that the data follow normal distribution, with mean and standard deviation parameters estimated from the data, we can randomly draw values from the distribution. For example, the figure below shows a histogram of 60 realizations from the normal distribution, using the mean and standard deviation of ozone at Lake Tahoe:

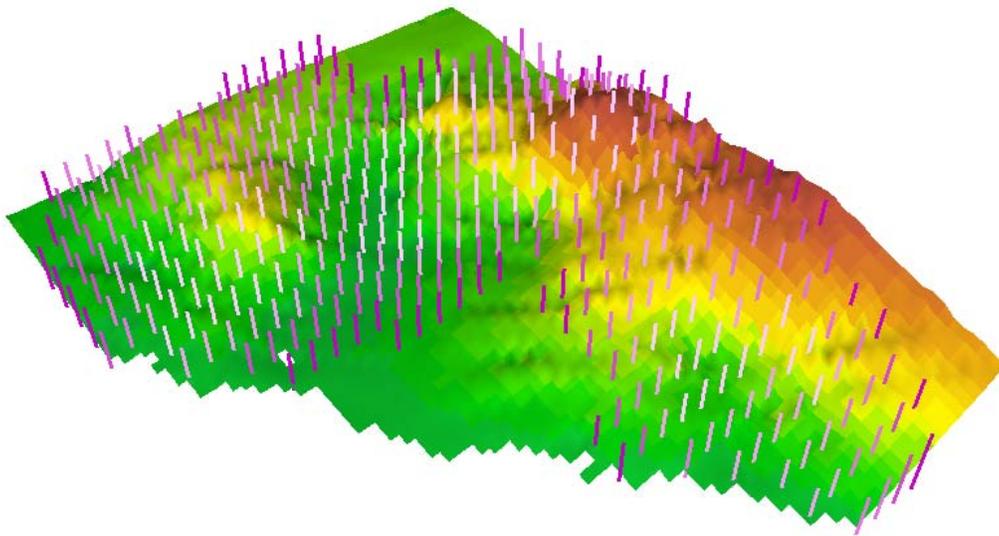


We would not be surprised to find that such values had actually been measured at Lake Tahoe.

We can repeat this exercise of fitting histograms of ozone concentration to normal distribution lines and drawing reasonable values from it in other California cities. If measurements replications are not available, we can assume that data distribution is known in each datum location. Some contouring algorithm applied to the sequence of real or simulated values at each city location results in a sequence of surfaces. Instead of

displaying all of them, we can show the most typical one and a couple that differ the most from the typical one. Then various surfaces can be represented by the most probable prediction map and by estimated prediction uncertainty in each possible location in the area under study.

The figure below, created in 3D analyst, illustrates variation in kriging predictions, showing the range of prediction error (sticks) over kriging predictions (surface). A stick's color changes according to the prediction error value. As a rule, prediction errors are larger in areas with a small number of samples.



Predictions with associated uncertainties are possible because certain properties of the random variables are assumed to follow some laws.

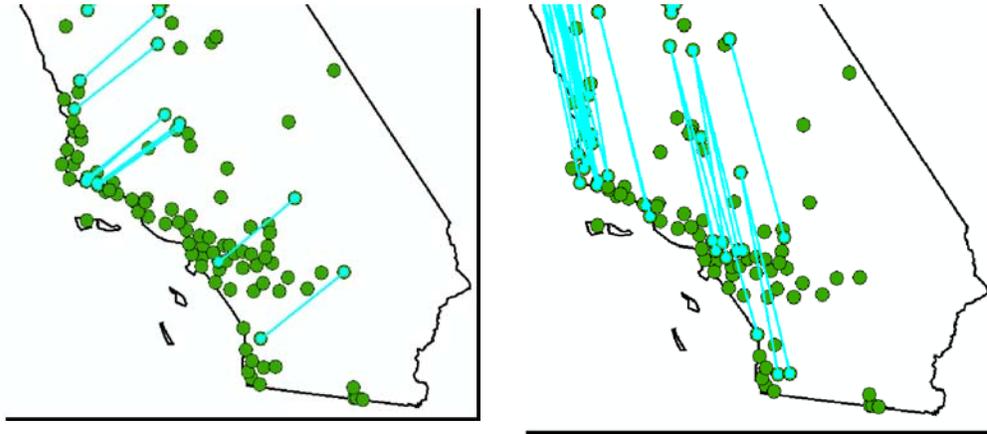
Stationarity

The statistical approach requires that observations be replicated in order to estimate prediction uncertainty.

Stationarity means that statistical properties do not depend on exact locations. Therefore, the mean (expected value) of a variable at one location is equal to the mean at any other location; data variance is constant in the area under investigation; and the correlation (covariance or semivariogram) between any two locations depends only on the vector that separates them, not their exact locations.

The figure below (created using Geostatistical Analyst's Semivariogram Cloud exploratory tool) shows many pairs of locations, linked by blue lines that are approximately the same length and orientation. In the case of data stationarity, they have

approximately the same spatial similarity (dependence). This provides statistical replication in a spatial setting so that prediction becomes possible.

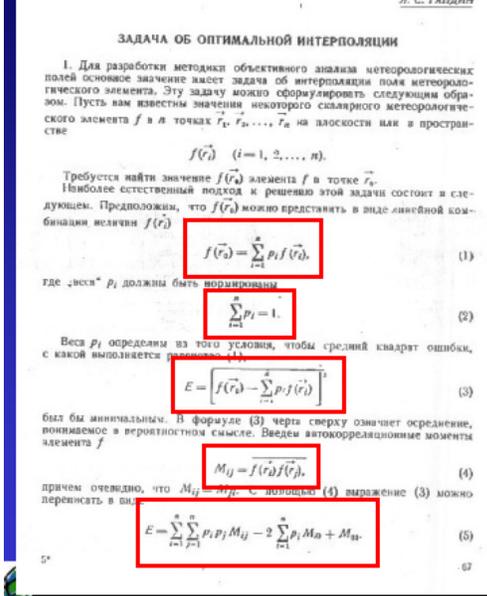


Spatial correlation is modeling as a function of distance between pairs of locations. Such functions are called covariance and semivariogram. Kriging uses them to make optimum predictions.

Optimum interpolation or “kriging”

Kriging is a spatial interpolation method used first in meteorology, then in geology, environmental sciences, and agriculture, among others. It uses models of spatial correlation, which can be formulated in terms of covariance or semivariogram functions. These are the first two pages of the first paper on optimum interpolation (later called “kriging”) by Lev Gandin. It is in Russian, but those of you who have read geostatistical papers before can recognize the derivation of ordinary kriging in terms of covariance and then using a semivariogram. Comments at the right show how the kriging formulas were derived. The purpose of this exercise is to show that the derivation of kriging formulas is relatively simple. It is not necessary to understand all the formulas.

(First article on kriging by Lev Candia, 1959)



1) Prediction is a *linear* combination of the measurements

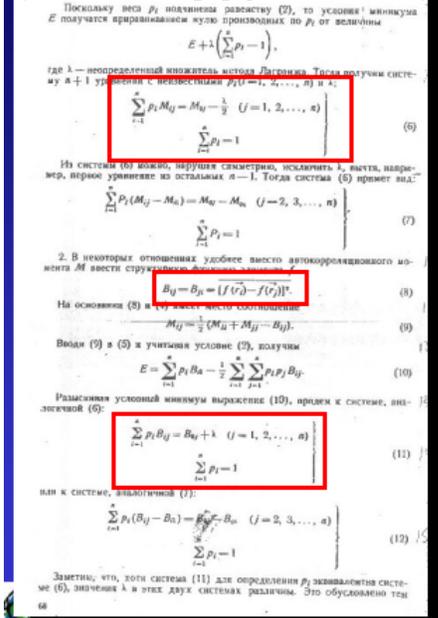
2) Sum of weights is equal to one

3) Use minimum of the expected square difference between the actual value and its prediction

4) Define covariance

5) Prediction error depends on known covariance

(First article on kriging by Lev Candia, 1959)



6) “Ordinary kriging” system using covariance

8) Define variogram

11) “Ordinary kriging” system using variogram

Kriging prediction and prediction uncertainty depend on covariances (or semivariograms)

Kriging uses a weighted average of the available data. Instead of just assuming that the weights are functions of distance alone (as with other mapping methods like inverse-distance weighting), we want to use the data to tell us what the weights should be. They are chosen using the concept of spatial stationarity and are quantified through the covariance or semivariogram function. It is assumed that the covariance or semivariogram function is known.

Kriging methods have been studied and applied extensively since 1959 and have been adapted, extended, and generalized. For example, kriging has been generalized to classes of nonlinear functions of the observations, extended to take advantage of covariate information, and adapted for non-Euclidean distance metrics. You can find a discussion on modern geostatistics in this paper, below, and in the [Educational and Research Papers](#) section.

Semivariogram and covariance

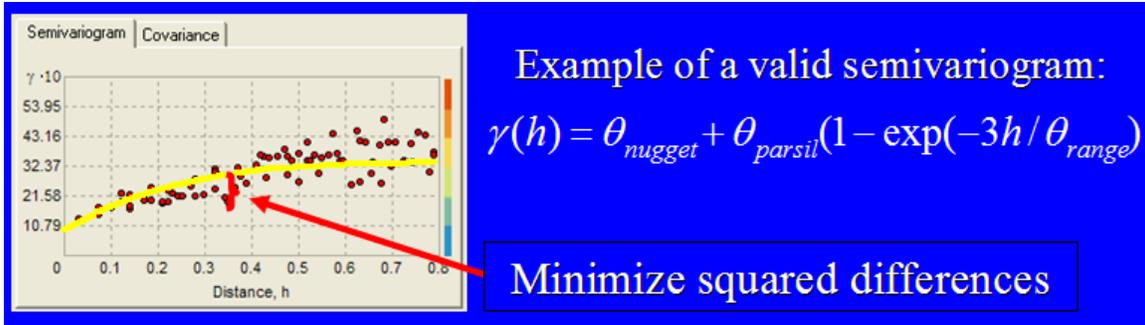
Creating an empirical semivariogram follows four steps:

- Find all pairs of measurements (any two locations).
- Calculate for all pairs the squared difference between values.
- Group vectors (or lags) into similar distance and direction classes. This is called binning.
- Average the squared differences for each bin. In doing so, we are using the idea of stationarity: the correlation between any two locations depends only on the vector that links them, not their exact locations.

Empirical Semivariogram	Empirical Covariance
<p>Semivariogram(distance h) =</p> <p>$\frac{1}{2}$ average[(value at location i – value at location j)²]</p> <p>for all pairs of locations i and j separated by distance h.</p>	<p>Covariance(distance h) =</p> <p>average[(value at location i – mean)*(value at location j – mean)]</p> <p>for all pairs of locations i and j separated by distance h.</p>

Estimation of the covariance is similar to the estimation of semivariogram, but requires the use of the data mean. Because the data mean is usually not known, but estimated, this causes bias. Hence, Geostatistical Analyst uses semivariogram as default function tool to characterize spatial data structure.

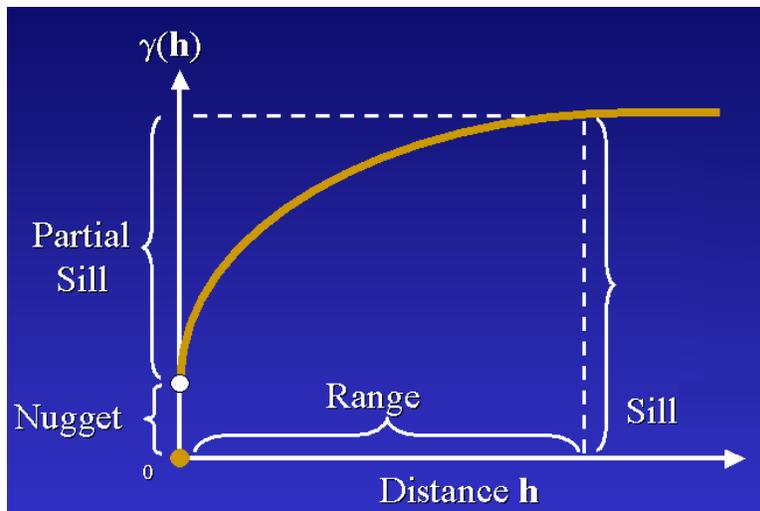
In Geostatistical Analyst, the average semivariogram value in each bin is plotted as a red dot, as displayed in the figure below, left.



The next step after calculating the empirical semivariogram is estimating the model that best fits it. Parameters for the model are found by minimizing the squared differences between the empirical semivariogram values and the theoretical model. In Geostatistical Analyst, this model is displayed as a yellow line. One such model, the exponential, is shown in the figure above, right.

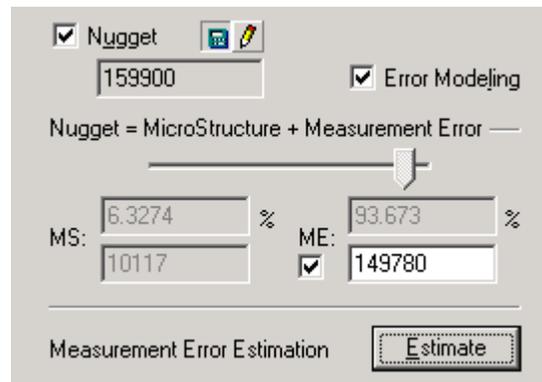
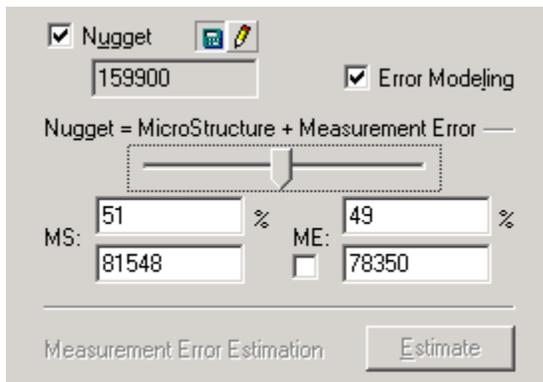
The semivariogram function has three or more parameters. In the exponential model, below,

- Partial sill is the amount of variation in the process that is assumed to generate data
- Nugget is data variation due to measurement errors and data variation at very fine scale, and is a discontinuity at the origin
- Range is the distance beyond which data do not have significant statistical dependence



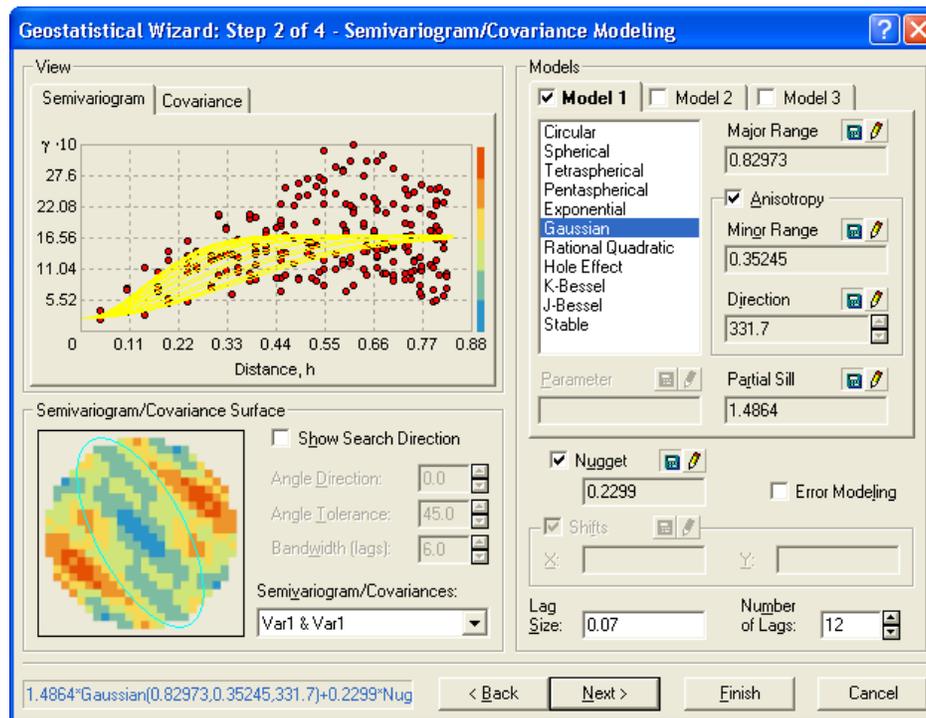
The nugget occurs when sampling locations are close to each other, but the measurements are different. The primary reason that discontinuities occur near the origin for semivariogram model is the presence of measurement and locational errors or variation at scales too fine to detect in the available data (microstructure).

In Geostatistical Analyst, the proportion between measurement error and microstructure can be specified, as in the figure below, left. If measurement replications are available, the proportion of measurement error in the nugget can be estimated, right.



Not just any function can be used as covariance and semivariogram. Geostatistical Analyst provides a set of valid models. Using these models avoids the danger of getting absurd prediction results.

Because we are working in two-dimensional space, we might expect that the semivariogram and covariance functions change not only with distance but also with direction. The figure below shows how the Semivariogram/Covariance Modeling dialog looks when spatial dependency varies in different directions.



The Gaussian semivariogram model, yellow lines, changes gradually as direction changes between pairs of points. Distance of significant correlation in the north-west direction is about twice as large as in the perpendicular direction.

The importance of Gaussian distribution

Kriging predictions are best among all weighted averaged models if input data are Gaussian. Without an assumption about the kriging prediction's Gaussianity, a prediction error map can only tell us where prediction uncertainty is large and where it is small. Only if the prediction distribution at each point is Gaussian can we tell how credible the predictions are.

Geostatistical Analyst's detrending and transformation options help to make input data close to Gaussian. For example, if there is evidence that data are close to lognormal distribution, you can use the lognormal transformation option in the Geostatistical Analyst modeling wizard. If input data are Gaussian, their linear combination (kriging prediction) is also Gaussian.

Because features of input data are important, some preliminary data analysis is required. The result of this exploratory data analysis can be used to select the optimum geostatistical model.

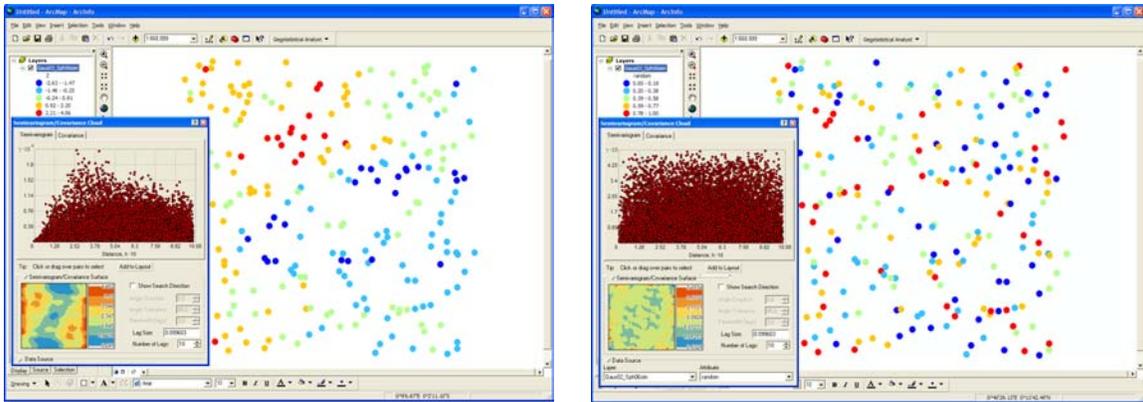
Exploratory Spatial Data Analysis

Geostatistical Analyst's Exploratory Spatial Data Analysis (ESDA) environment is composed of a series of tools, each allowing a view into the data. Each view is interconnected with all other views as well as with ArcMap. That is, if a bar is selected in the histogram, the points comprising the bar are also selected on any other open ESDA view, and on the ArcMap map.

Certain tasks are useful in most explorations, defining the distribution of the data, looking for global and local outliers, looking for global trends, and examining spatial correlation and the covariation among multiple datasets.

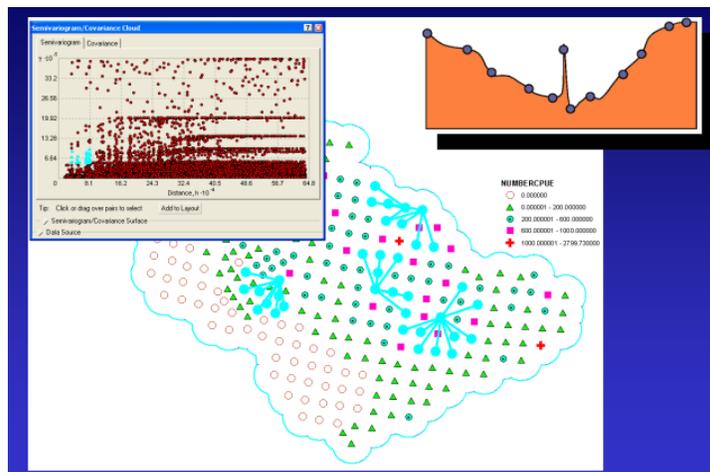
The semivariogram/covariance cloud tool is used to examine data spatial correlation, figure below. It shows the empirical semivariogram for all pairs of locations within a dataset and plots them as a function of the distance between the two locations. In addition, the values in the semivariogram cloud are put into bins based on the direction and distance between a pair of locations. These bin values are then averaged and smoothed to produce the surface of the semivariogram. The semivariogram surface shows semivariogram values in polar coordinates. The center of the semivariogram surface corresponds to the origin of the semivariogram graph.

There is a strong spatial correlation in the data displayed in the figure below, left, but data are independent in the right.



The semivariogram surface in the figure to the right is homogeneous, meaning that data variation is approximately the same regardless of the distance between pairs of locations. The semivariogram surface in the figure to the left shows a clear structure of spatial dependence, especially in the east-west direction.

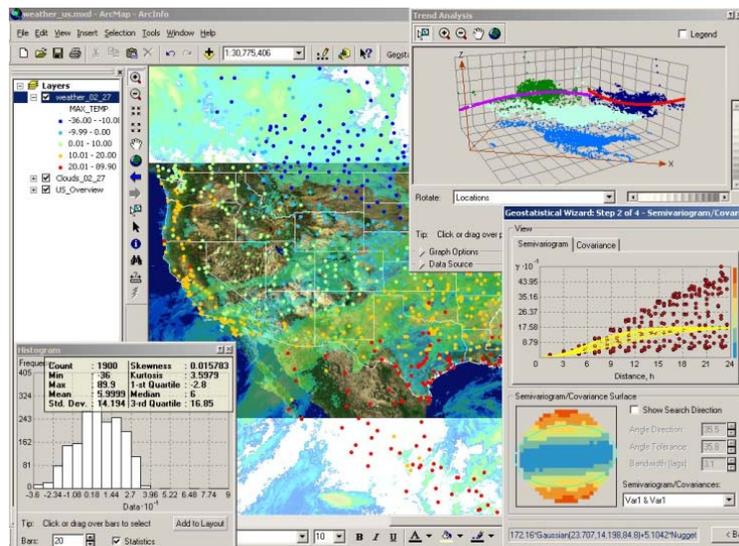
The semivariogram/covariance cloud tool can be used to look for data outliers. You can select dots and see the linked pairs in ArcMap. If you have a global outlier in your dataset, all pairings of points with that outlier will have high values in the semivariogram cloud, no matter what the distance. When there is a local outlier, the value will not be out of the range of the entire distribution but will be unusual relative to the surrounding values, as illustrated in the top right side of the figure below. In the semivariogram graph, top left, you can see that pairs of locations that are close together have a high semivariogram value (they are to the far left on the x-axis, indicating that they are close together, and high on the y-axis, indicating that the semivariogram values are high).



When these points are selected, you can see that all of these points are pairing to just four locations. Thus, the common centers of the four clusters in the figure above are possible local data outliers, and they require special attention.

The next task after checking the data correlation and looking for data outliers is investigating and mapping large-scale variation (trend) in the data. If trend exists, the mean data value will not be the same everywhere, violating one of the assumptions about data stationarity. Information about trend in the data is essential for choosing the appropriate kriging model, one that takes the variable data mean into account.

Geostatistical Analyst provides several tools for detecting trend in the data. The figure below shows ESDA Trend Analysis, top right, and the Wizard's Semivariogram Dialog, bottom right, in the process of trend identification in meteorological data. They show that data variability in the north-south direction differs from that in the east-west.



The Trend Analysis tool provides a three-dimensional perspective of the data. Above each sample point, the value is given by the height of a stick in the Z dimension with input data points on top of the sticks. Then the values are projected onto the XZ plane and the YZ plane, making a sideways view through the three dimensional data. Polynomial curves are then fit through the scatter plots on the projected planes. The data can be rotated to identify directional trends.

There are other ESDA tools for investigation of the data variability, the degree of spatial cross-correlation between variables, and closeness of the data to Gaussian distribution.

After learning about the data using interactive ESDA tools, model construction using the wide choice of Geostatistical Analyst's models and options becomes relatively straightforward.

Interpolation models

Geostatistical Analyst provides deterministic and geostatistical interpolation models. Deterministic models are based on either the distance between points (e.g., Inverse Distance Weighted) or the degree of smoothing (e.g., Radial Basis Functions and Local

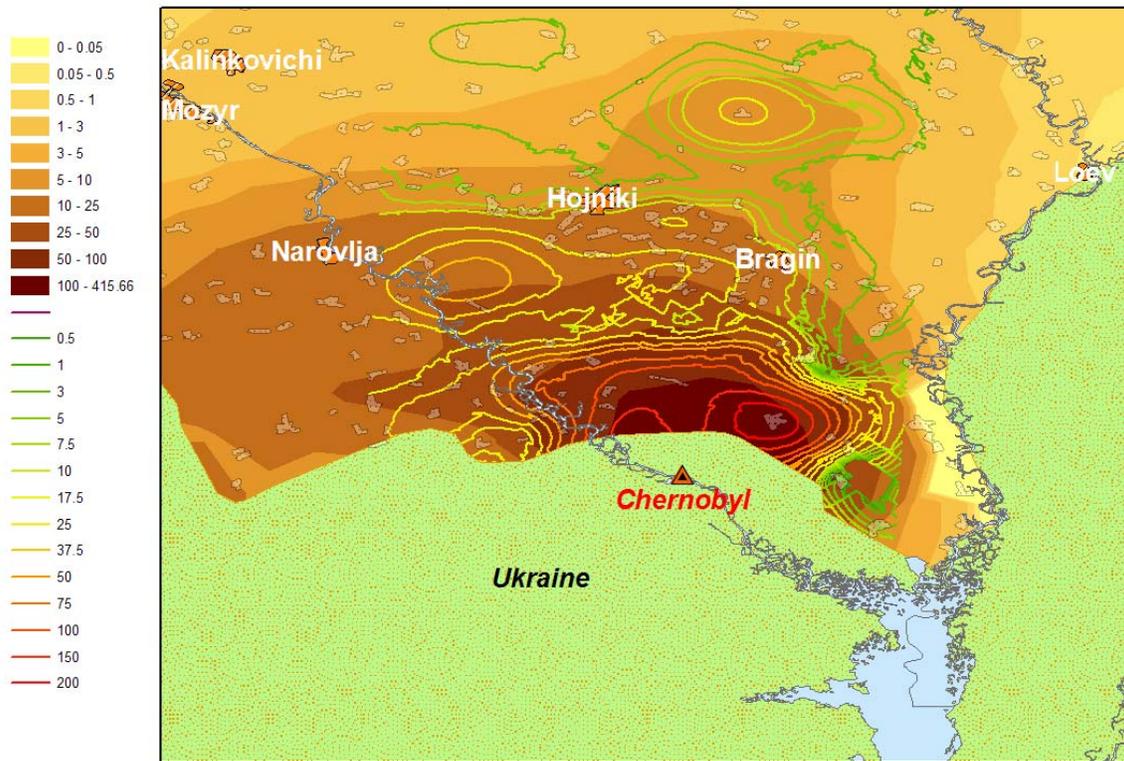
Polynomials). Geostatistical models (e.g., kriging) are based on the statistical properties of the observations.

Interpolators can either force the resulting surface to pass through the data values or not. An interpolation technique that predicts a value identical to the measured value at a sampled location is known as an exact interpolator. Because there is no error in prediction, other measurements are not taken into account. A filtered (inexact) interpolator predicts a value that is different from the measured noisy value. Since the data are inexact, the use of data at other points lead to an improvement of the prediction. Inverse Distance Weighted and Radial Basis Functions are exact interpolators, while Global and Local Polynomial Interpolations are inexact. Kriging can be both an exact and a filtered interpolator.

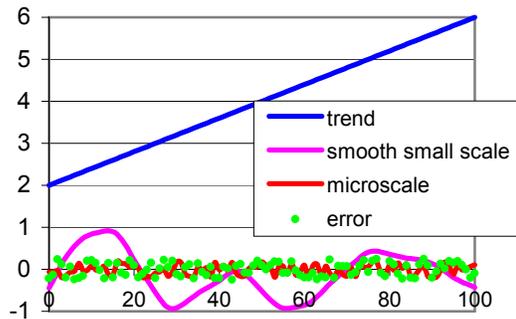
The Geostatistical Analyst's geostatistical model

The geostatistical model for spatial correlation consists of three spatial scales and uncorrelated measurement error. The first scale represents spatial continuity on very short scales, less than the separation between observations. The second scale represents the separation between nearby observations, the neighborhood scale. The third scale represents the separation between non-proximate observations, regional scale.

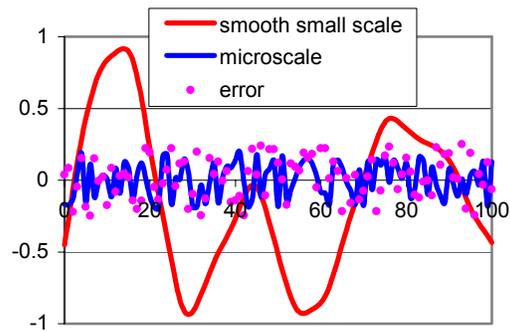
The figure below shows radiocesium soil contamination in southern Belarus as a result of the Chernobyl accident. Units are Ci/sq.km. Filled contours show large scale variation in contamination, which decreases with increasing distance from the Chernobyl nuclear power plant; that is, the mean value is not constant. Using only local data, contour lines display detailed variation in radiocesium soil contamination. Contours of cities and villages with populations larger than 300 are displayed. Only averaged values of radiocesium in these settlements are available, but they are different in different parts of the cities. They belong to the third, micro scale of contamination. Measurement errors are about 20% of the measured values.



The figures below display components of the Geostatistical Analyst's geostatistical model in one dimension.

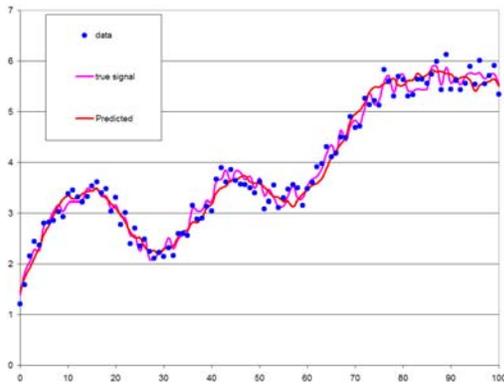


Model components

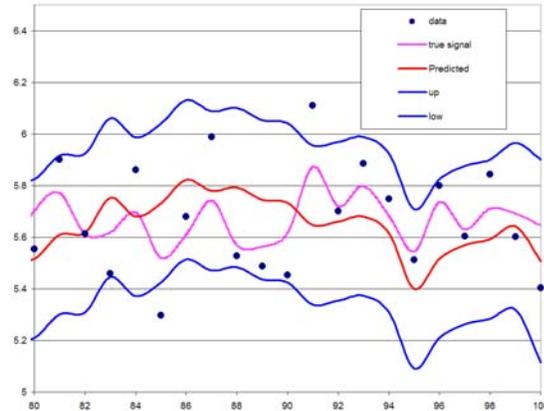


Enlarged components without trend

Kriging can reconstruct both large and small scale variations. If data are not precise, kriging can filter out either estimated measurement errors or specified ones. However, there is no way to reconstruct microscale data variation. As a result, kriging predictions are smoother than data variation, see figure below, right.



Predictions



Predictions, enlarged

The figure above shows the true signal in pink, which we are attempting to reconstruct from measurements contaminated by errors, and the filtered kriging prediction in red. In addition, the 90% confidence interval is displayed in blue.

The true signal is neither equal to the data nor to the predictions. As a consequence, the estimated prediction uncertainty should be presented together with kriging predictions to make the result of the data analysis informative.

Measurement errors and microscale data variation

Extremely precise measurements can be very costly or practically impossible to obtain. While this should not limit the use of the data for decision-making, neither should it give decision-makers the impression that the maps based on such data are exact.

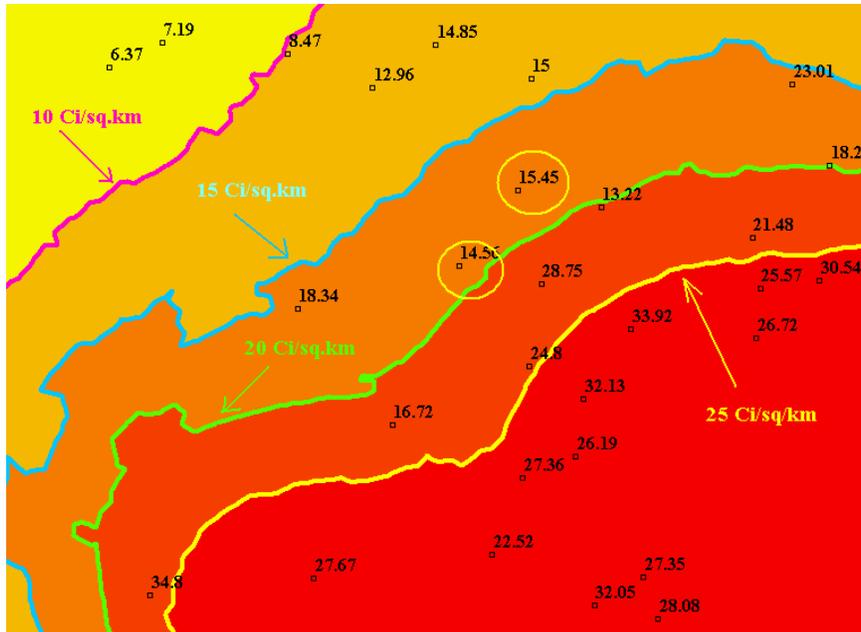
Most measurements in spatial data contain errors both in attribute values and in locations. These occur whenever it is possible to have several different observations at the same location.

In the Geostatistical Analyst, you can specify a proportion of the estimated nugget effect as microscale variation and measurement variation, or you can ask Geostatistical Analyst estimate measurement error for you if you have multiple measurements per location, or you can input a value for measurement variation.

Exact versus filtered interpolation

Exact kriging predictions will change gradually and smoothly in space until they get to a location where data have been collected, at which point the prediction jumps to the measured value. The prediction standard error changes gradually except at the measured locations, where it jumps to zero. Jumps are not critical for mapping since the contouring will smooth over the details at any given point, but it may be very important for prediction of new values to the sampled locations when these predicted values are to be used in subsequent computations.

The figure below presents ^{137}Cs soil contamination data in some Belarus settlements in the northeastern part of the Gomel province. Isolines for 10, 15, 20, and 25 Ci/km^2 are shown. The map was created with ordinary kriging using the J-Bessel semivariogram. Two adjacent measurement locations separated by one mile are circled. These locations have radiocesium values of 14.56 and 15.45 Ci/km^2 , respectively. The safety threshold is 15 Ci/km^2 . But the error of ^{137}Cs soil measurements is about 20%, so it would be difficult to conclude that either location is safe. Using only raw data, the location with 14.56 Ci/km^2 could be considered safe.

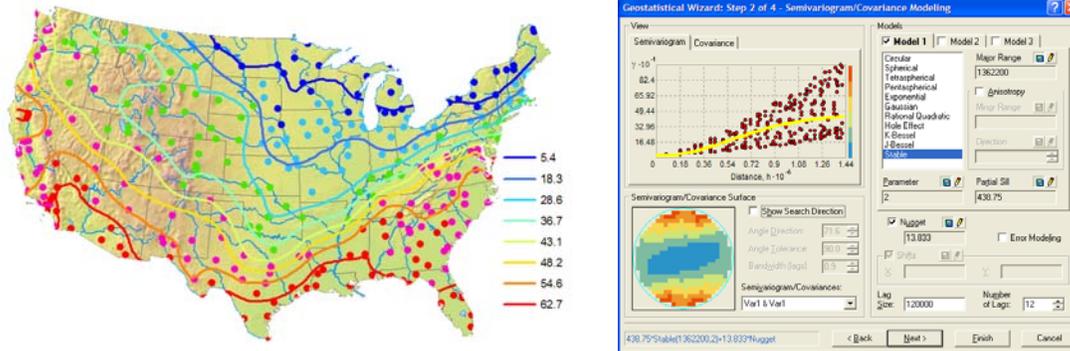


With data subject to measurement error, one way to improve predictions is to use filtered kriging. Taking the nugget as 100% measurement error, the location with the observed value of 14.56 Ci/km^2 is predicted to be 19.53 Ci/km^2 , and the location with 15.45 Ci/km^2 is predicted to be 18.31 Ci/km^2 . Even with the nugget as 50% measurement error and 50% microscale variation, the location with the observed value of 14.56 Ci/km^2 will be predicted to be 17.05 Ci/km^2 , and the location with the value of 15.45 Ci/km^2 will be predicted to be 16.88 Ci/km^2 . Filtered kriging places both locations above 15 Ci/km^2 . People living in and around both locations should probably be evacuated.

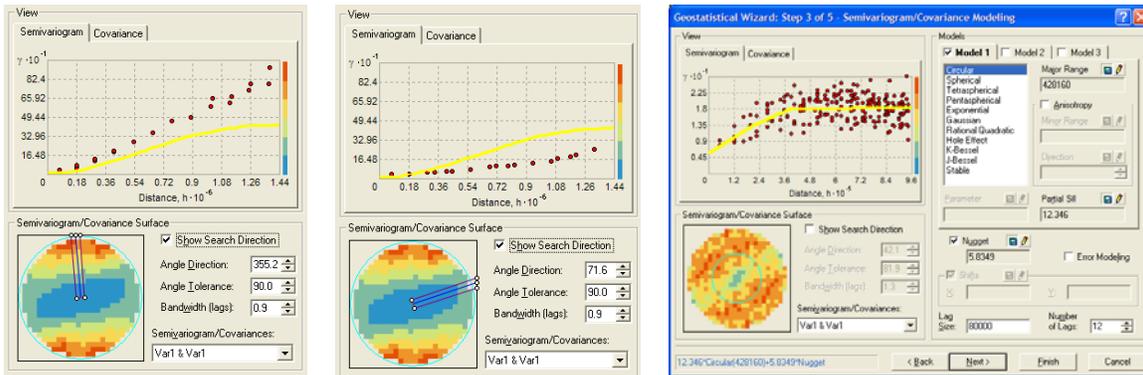
Trend or large scale data variation

Trend is another component of the geostatistical model. North of the equator, temperature systematically increases from north to south and this large-scale variation exists regardless of mountains or ocean. Crop production may change with latitude because temperature, humidity, and rainfall change with latitude.

We will illustrate concept of trend using winter temperature for one particular day in the USA, figure below, left. Mean temperature value is different in the northern and southern parts of the country, meaning that data are not stationary. Systematic changes in the data are recognizable in the semivariogram surface and graph, figure below right: the surface is not symmetric, and there is a very large difference in the empirical semivariogram values.



There is large difference in the semivariogram values in north-south and east-west directions, figure below, left and middle.



In Geostatistical Analyst, large scale variation can be estimated and removed from the data. Then the semivariogram surface will become almost symmetrical, figure above, right, and the empirical semivariogram will behave similarly in any direction, as it should for stationary data.

Kriging neighborhood

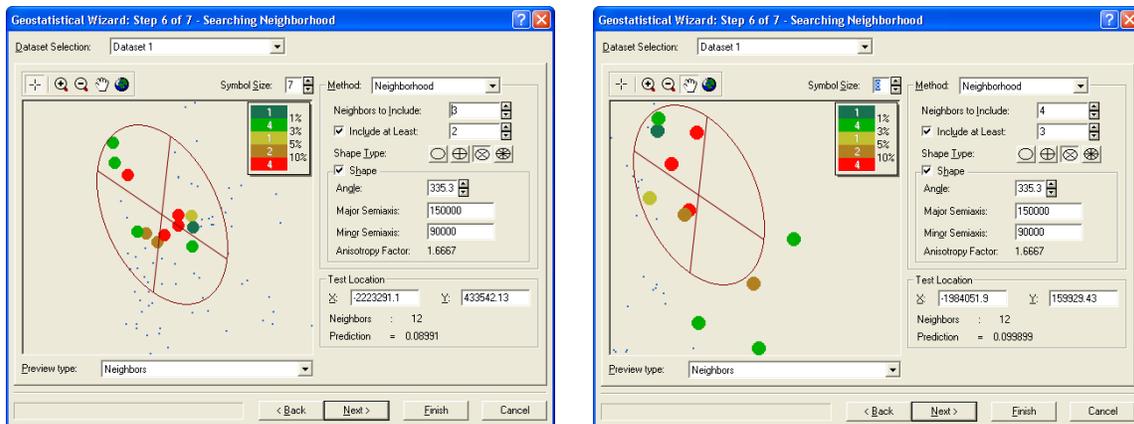
Kriging can use all input data. However, there are several reasons for using nearby data to make predictions.

First, kriging with large number of neighbors (larger than 100-200 observations) leads to the computational problem of solving a large system of linear equations.

Second, the uncertainties in semivariogram estimation and measurement make it possible that interpolation with a large number of neighbors will produce a larger mean-squared prediction error than interpolation with a relatively small number of neighbors. Third, using of local neighborhood leads to the requirement that the mean value should be the same only in the moving neighborhood, not for the entire data domain.

Geostatistical Analyst provides many options for selecting the neighborhood window. You can change the shape of the searching neighborhood ellipse, the number of angular sectors, and minimum and maximum number of points in each sector.

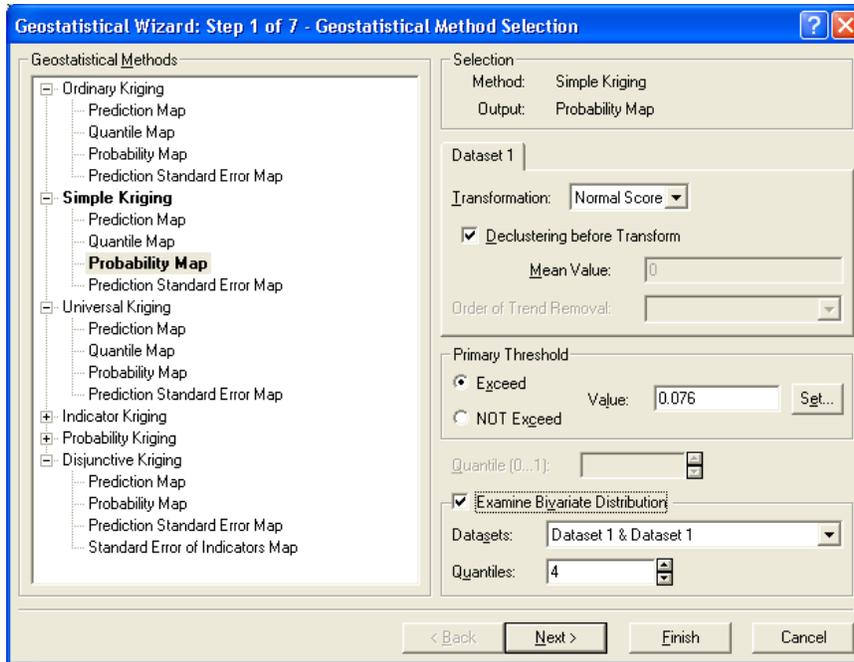
The figure below shows two examples of a kriging searching neighborhood. Colored circles from dark green to red show the absolute values of kriging weights in percents when predicting to the center of the ellipse.



Kriging weights do not depend simply on the distance between points.

The different types of kriging, their uses, and their assumptions

Just as a well-stocked carpenter's toolbox contains a variety of tools, so does Geostatistical Analyst contain a variety of kriging models. The figure below shows the Geostatistical Method Selection dialog. The large choice of options may confuse a novice. We will briefly discuss these options in the rest of the paper.



Simple, ordinary, and universal kriging predictors are all linear predictors, meaning that prediction at any location is obtained as a weighted average of neighboring data. These three models make different assumptions about the mean value of the variable under study: simple kriging requires a known mean value as input to the model (or mean surface, if a local searching neighborhood is used), while ordinary kriging assumes a constant, but unknown mean, and estimates the mean value as a constant in the searching neighborhood. Universal kriging models local means as a sum of low order polynomial functions of the spatial coordinates. This type of model is appropriate when there are strong trends or gradients in the measurements.

Indicator kriging was proposed as an alternative to disjunctive and multiGaussian kriging (that is linear kriging after data transformation), which require a good understanding of the assumptions involved and are difficult to use.

In indicator kriging, the data are pre-processed. Indicator values are defined for each data location as the following: an indicator is set to zero if the data value at the location s is below the threshold, and to one otherwise:

$$I(\mathbf{s}) = I(Z(\mathbf{s}) < \text{threshold}) = \begin{cases} 0, & Z(\mathbf{s}) < \text{threshold}; \\ 1, & Z(\mathbf{s}) > \text{threshold}. \end{cases}$$

Then these indicator values are used as input to the ordinary kriging. Ordinary kriging produces continuous predictions, and we might expect that prediction at the unsampled location will be between zero and one. Such prediction is interpreted as the probability that the threshold is exceeded at location s . A prediction equal to 0.71 is interpreted as having a 71% chance that the threshold was exceeded. Predictions made at each location form a surface that can be interpreted as a probability map of the threshold being exceeded.

It is safe to use indicator kriging as a data exploration technique, but not as a prediction model for decision-making, see [Educational and Research Papers](#).

Disjunctive kriging uses a linear combination of functions of the data, rather than just the original data values themselves. Disjunctive kriging assumes that all data pairs come from a bivariate normal distribution. The validity of this assumption should be checked in the Geostatistical Analyst's Examine Bivariate Distribution Dialog. When this assumption is met, then disjunctive kriging, which may outperform other kriging models, can be used.

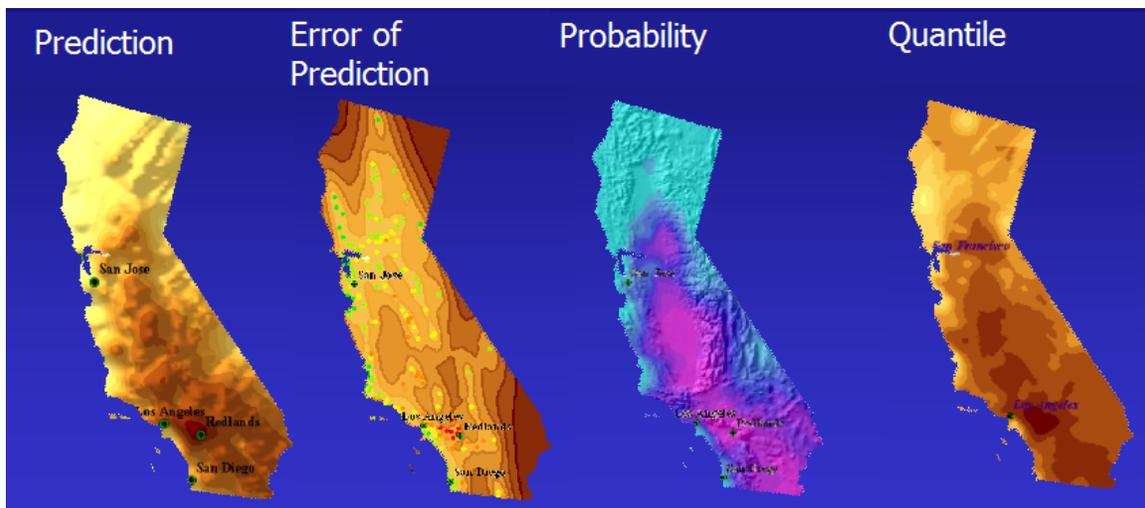
Often we have a limited number of data measurements and additional information on secondary variables. **Cokriging** combines spatial data on several variables to make a single map of one of the variables using information on the spatial correlation of the variable of interest and cross-correlations between it and other variables. For example, the prediction of ozone pollution may improve using distance from a road or measurements of nitrogen dioxide as secondary variables.

It is appealing to use information from other variables to help make predictions, but it comes at a price: the more parameters need to be estimated, the more uncertainty is introduced.

All kriging models mentioned in this section can use secondary variables. Then they are called simple cokriging, ordinary cokriging, and so on.

Types of Output Maps by Kriging

Below are four output maps created using different Geostatistical Analyst renderers on the same data, maximum ozone concentration in California in 1999.



Prediction maps are created by contouring many interpolated values, systematically obtained throughout the region.

Standard Error maps are produced from the standard errors of interpolated values, as quantified by the minimized root mean squared prediction error that makes kriging

optimum.

Probability maps show where the interpolated values exceed a specified threshold.

Quantile maps are probability maps where the thresholds are quantiles of the data distribution. These maps can show overestimated or underestimated predictions.

Transformations

Kriging predictions are best if input data are Gaussian, and Gaussian distribution is needed to produce confidence intervals for prediction and for probability mapping. Geostatistical Analyst provides the following functional transformations: Box-Cox (also known as power transformation), logarithmic, and arcsine. The goal of functional transformations is to remove the relationship between the data variance and the trend. When data are composed of counts of events, such as crimes, the data variance is often related to the data mean. That is, if you have small counts in part of your study area, the variability in that region will be larger than the variability in region where the counts are larger. In this case, the square root transformation will help to make the variances more constant throughout the study area, and it often makes the data appear normally distributed as well.

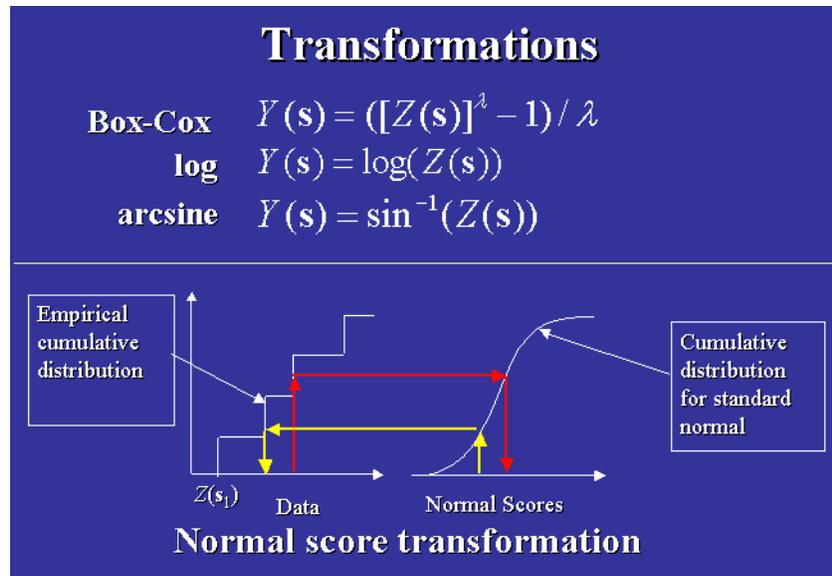
The arcsine transformation is used for data between 0 and 1. The arcsine transformation can be used for data that consists of proportions or percentages. Often, when data consists of proportions, the variance is smallest near 0 and 1 and largest near 0.5. Then the arcsine transformation often yields data that has constant variance throughout the study area and often makes the data appear normally distributed as well.

Kriging using power and arcsine transformations is known as *transGaussian* kriging.

The log transformation is used when the data have a skewed distribution and only a few very large values. These large values may be localized in your study area. The log transformation will help to make the variances more constant and normalize your data.

Kriging using logarithmic transformations called *lognormal* kriging.

The normal score transformation is another way to transform data to Gaussian distribution. The bottom part of the graph in figure below shows the process of transformation of the cumulative distribution function of the original data to the standard normal distribution, red lines, and back transformation, yellow lines.

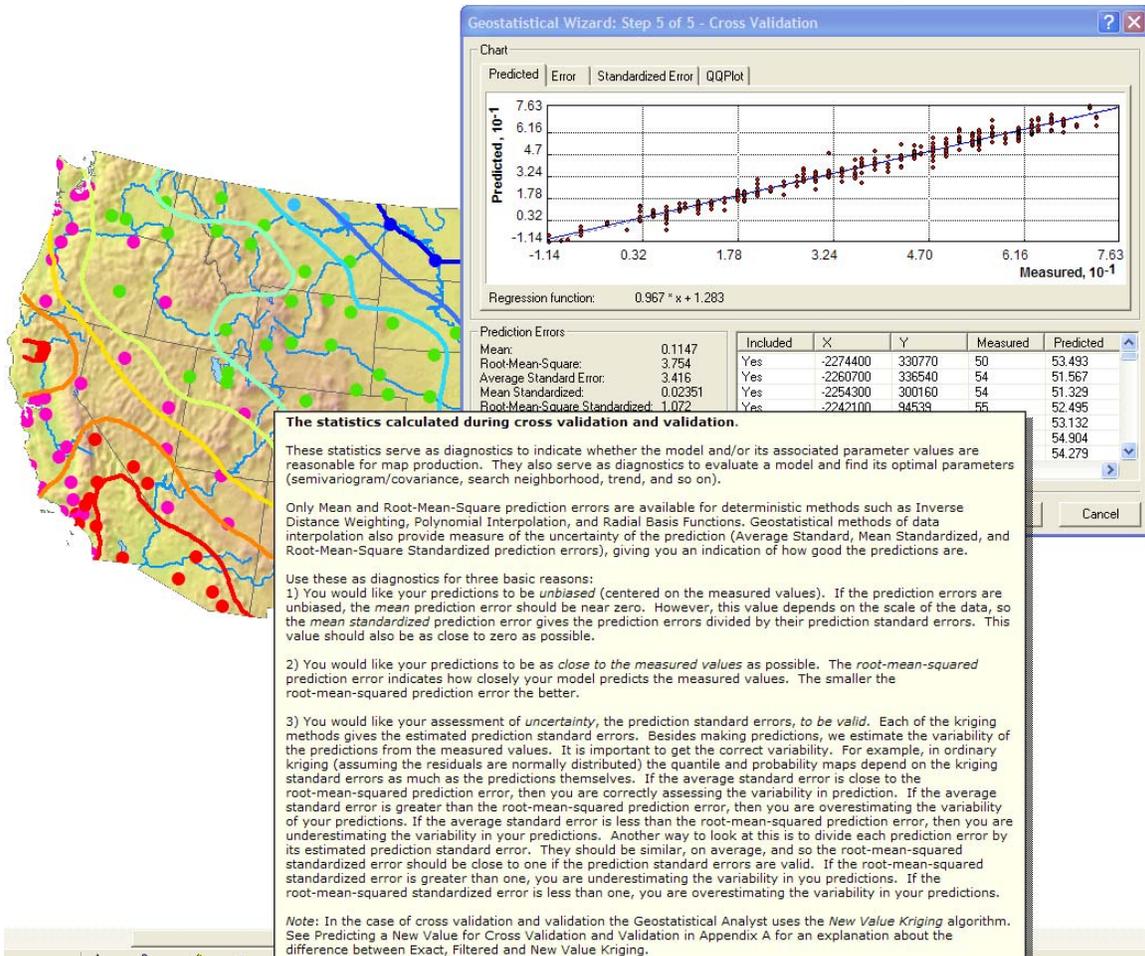


The goal of the normal score transformation is to make all random errors for the whole population be normally distributed. Thus, it is important that the cumulative distribution from the sample reflect the true cumulative distribution of the whole population. The fundamental difference between the normal score transformation and the functional transformations is that the normal score transformation transformation function changes with each particular dataset, whereas functional transformations do not (e.g., the log transformation function is always the natural logarithm).

Diagnostic

You should have some idea of how well different kriging models predict the values at unknown locations. Geostatistical Analyst diagnostics, cross-validation and validation, help you decide which model provides the best predictions. Cross-validation and validation withhold one or more data samples and then make a prediction to the same data locations. In this way, you can compare the predicted value to the observed value and from this get useful information about the accuracy of the kriging model, such as the semivariogram parameters and the searching neighborhood.

The calculated statistics indicate whether the model and its associated parameter values are reasonable. Geostatistical Analyst provides several graphs and summaries of the measurement values versus the predicted values. The screenshot below shows the cross-validation dialog and the help window with explanations on how to use calculated cross-validation statistics.



The graph helps to show how well kriging is predicting. If all the data were independent (no spatial correlation), every prediction would be close to the mean of the measured data, so the blue line would be horizontal. With strong spatial correlation and a good kriging model, the blue line should be closer to the 1:1 line.

Summary statistics on the kriging prediction errors are given in the lower left corner of the dialog above. You use these as diagnostics for three basic reasons:

- You would like your predictions to be unbiased (centered on the measurement values). If the prediction errors are unbiased, the mean prediction error should be near zero.
- You would like your predictions to be as close to the measurement values as possible. The root-mean-square prediction errors are computed as the square root of the average of the squared difference between observed and predicted values. The closer the predictions are to their true values the smaller the root-mean-square prediction errors.
- You would like your assessment of uncertainty to be valid. Each of the kriging methods gives the estimated prediction standard errors. Besides making predictions, we estimate the variability of the predictions from the measurement

values. If the average standard errors are close to the root-mean-square prediction errors, then you are correctly assessing the variability in prediction.

Reproducible research

It is important, that given the same data, another analyst will be able to re-create the research outcome used in a paper or report. If software has been used, the implementation of the method applied should also be documented. If arguments used by the implemented functions can take different values, then these also need documentation.

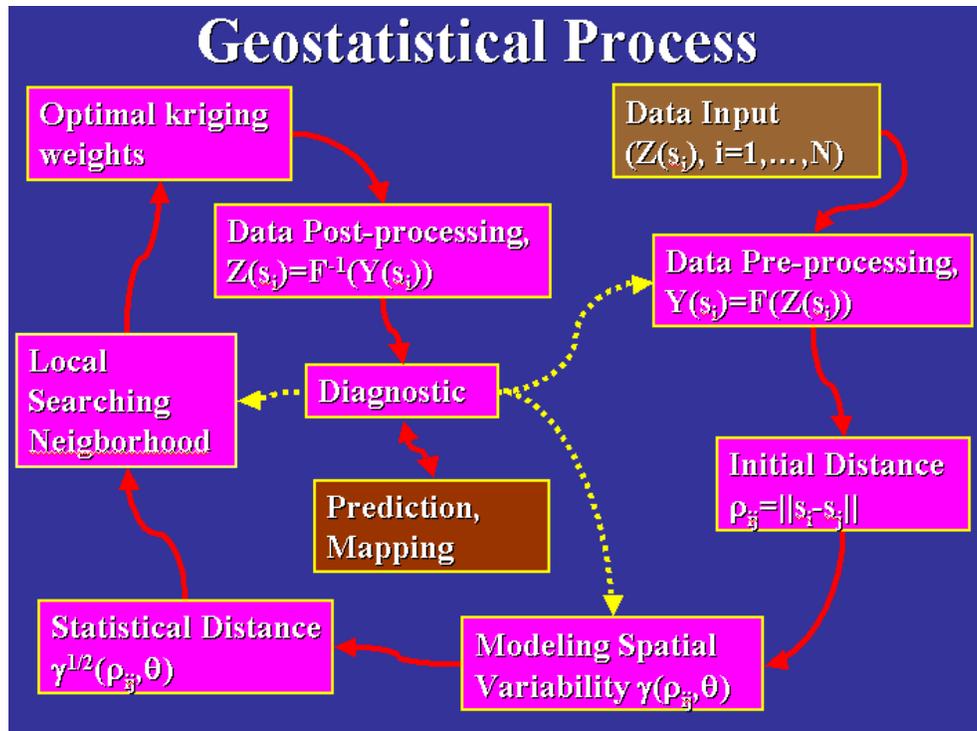
A geostatistical layer stores the sources of the data from which it was created (usually a point feature layers), the projection, symbology, and other data and mapping characteristics, but it also stores documentation that could include the model parameters from the interpolation, including type of or model for data transformation, covariance and cross-covariance models, estimated measurement error, trend surface, searching neighborhood, and results of validation and cross-validation.

You can send a geostatistical layer to your colleagues to provide information on your kriging model.

Summary on Modeling using Geostatistical Analyst

Typical scenario of data processing using Geostatistical Analyst is the following (see also scheme below):

- Looking at the data using GIS functionality
- Learning about the data using interactive Exploratory Spatial Data Analysis tools
- Selecting a model based on result of data exploration
- Choosing the model's parameters using a wizard
- Performing cross-validation and validation diagnostic of the kriging model
- Depending on the result of the diagnostic, doing more modeling or creating a sequence of maps.



Selected References

Geostatistical Analyst manual.

http://store.esri.com/esri/showdetl.cfm?SID=2&Product_ID=1138&Category_ID=121

Educational and research papers available from ESRI online at

http://www.esri.com/software/arcgis/extensions/geostatistical/research_papers.html.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Addison Wesley Longman Limited, Essex.

Cressie, N. (1993). *Statistics for Spatial Data*, Revised Edition, John Wiley & Sons, New York.

Announcement

ESRI Press plans to publish a book by Konstantin Krivoruchko on spatial statistical data analysis for non-statisticians. In this book

- A probabilistic approach to GIS data analysis will be discussed in general and illustrated by examples.
- Geostatistical Analyst's functionality and new geostatistical tools that are under development will be discussed in detail.
- Statistical models and tools for polygonal and discrete point data analysis will be examined.
- Case studies using real data, including Chernobyl observations, air quality in the USA, agriculture, census, business, crime, elevation, forestry, meteorological, and

fishery will be provided to give reader a practice dealing with the topics discussed in a book. Data will be available on the accompanying CD.