Geostatistics FAQ — Frequently Asked Questions

Last changes: Chapter 10.3 rewritten on 21st of February 2003

Table of Contents:

---

[0] Preface

(Part of *Geostatistics FAQ* , Copyright © 1996-97, syed@spemail.org)

[0.1] Preface

This is the third issue of the AI-GEOSTATS FAQ list. Comments and suggestions to improve it are very much appreciated.

AI-GEOSTATS is general purpose list to discuss issues related to spatial data analysis and geostatistics. Questions from the most basic to the most complex are discussed, and subscribers come from a variety of backgrounds, including mining, petroleum, ecology, soil science, environmental engineering, statistics, and many others. AI stands initially for Arc-Info, a popular GIS used for spatial data analysis. GEOSTATS stands for geostatistics, an applied statistical discipline for the analysis of geostatistical data.

The AI-GEOSTATS mailing list has a World Wide Web home page at http://www.ai-geostats.org/

The moderator is Grégoire Dubois. These pages include subscriber information, software information (both public domain and commercial), archives of the discussions, links to papers published online, information on books and other information that pertain to spatial data analysis and geostatistics.

Users can subscribe to the mailing list by sending mail to majordomo@unil.ch and writing the following as the message:

- subscribe ai-geostats [your email address]
- To unsubscribe, send the following message to the same address:
- unsubscribe ai-geostats [your email address]
- Comments and suggestions concerning the concept of this service can be directed to Grégoire Dubois (gregoire.dubois@usa.net).

The following questions are answered in this FAQ list. If you have a query that is not in this FAQ, feel free to Email to the mailing list. If you see errors in this FAQ, please Email myself or Gregoire. The answers to the questions reflect my own biases and prejudices after some years working in the petroleum industry. I don't pretend to know much about other disciplines, so if you notice solutions that are not applicable to your problem domain, do point them out and share your experiences!

Back to top

[1] Introduction

[1.1] What is geostatistics?

Geostatistics involves the analysis and prediction of spatial or temporal phenomena, such as metal grades, porosities, pollutant concentrations, price of oil in time, and so forth. The prefix geo- is usually associated with geology, since geostatistics has its origins in mining. Nowadays, geostatistics is just a name associated with a class of techniques used to analyze and predict values of a variable distributed in space or time. Such values are implicitly assumed to be correlated with each other, and the study of such a correlation is usually called a "structural analysis" or "variogram modeling." After structural analysis, predictions at unsampled locations are made using "kriging" or they can be simulated using "conditional simulations." Briefly, the steps in a geostatistical study include: (a) exploratory data analysis (b) structural analysis (calculation and modeling of variograms) (c) making predictions (kriging or simulations)

[1.2] What is spatial data analysis?

Geostatistics is sometimes referred to as a set of techniques for the spatial analysis of "geostatistical data", or data with continuous spatial index. Included in the "family" of spatial data types are lattice data (data with a countable collection of spatial sites, e.g. the distribution of infant mortalities in different counties or towns) and spatial point pattern data (data where both location and magnitude are random, e.g. a realization of geological sedimentary bodies in space). This FAQ is currently only limited to the analysis of geostatistical data. However, some overlap of techniques do exist, e.g. the use of simulated annealing for the stochastic conditional simulation of "geostatistical data" on a regular grid as opposed to lattice data.

Back to top

[2] Books and journals

[2.1] Which books should I read in order to get started with geostatistics?

There has been a recent increase in the number of books on the subject of geostatistics, but a good beginner book to read would still be "An Introduction to Applied Geostatistics" by R.M. Srivastava and E. Isaaks (Oxford University Press). For more advanced mathematical treatments, one can read "Statistics for Spatial Data" by Noel A. Cressie (John Wiley) and "Mining Geostatistics" by A.G. Journel and Ch.J Huijbregts (Academic Press). Mining engineers can also read "Handbook of Applied Advanced Geostatistical Ore Reserve Estimations" by M. David (Elsevier).

The Proceedings of the Geostatistics Congresses (held every four years) are published by Kluwer Academic Publishers in Holland. The papers are usually of advanced mathematical nature, however.

Readers interested in the current state-of-the-art in non-linear geostatistics can read Rivoirard's "Introduction to Disjunctive Kriging and Non-linear Geostatistics" (Oxford).

Readers interested in multivariate analysis can read "Multivariate Analysis" by Hans Wackernagel and published by Springer-Verlag.

"Geostatistics for natural resources characterization" by Pierre Goovaerts (Oxford) provides a good bridge between GSLIB applications and their underlying theory and practice.

"Geostatistical Error Management" by Jeff Myers (Van Nostrand and Reinhold) provides a good introduction to the applications of geostatistics for the environmetal sciences.

[2.2] What about journals?

Expect a diverse literature search for geostatistical techniques and case studies. E.g., for geological data, a major publication would be the Journal of Mathematical Geology, published by the International Association for Mathematical Geology. Also published by the same body is Computers and Geosciences, which focuses on algorithmic implementations. Other focus areas would include mining, environmental applications, social sciences, remote sensing, petroleum engineering, and so forth. Some techniques would normally overlap between such focus areas, while specific advances would be particular to a certain problem domain (e.g. the prominence of petroleum industry initiatives for stochastic conditional simulation of reservoir properties and multivariate data integration).

Back to top

[3] Software: public domain and commercial

[3.1] Where can I get public domain and commercial geostatistical software?

A lot of public domain and commercial software (as well as how to download them) are listed in the software section of AI-GEOSTATS, which is obtainable from http://www.ai-geostats.org/. One

of the most popular public domain software is GSLIB, which is a collection of Fortran 77 programs for kriging, cokriging, variogram calculations, stochastic simulations, and basic statistical plotting (histograms, variograms, 2D grid maps). You would need a Fortran compiler to use these programs. The manual for these programs is published as a textbook by Oxford University Press, "GSLIB: Geostatistical Software Library and User's Guide" by C.V. Deutsch and A.G. Journel. The programs can be obtained by anonymous ftp from geostat1.stanford.edu. Two other popular DOS programs with graphical facilities are Geostat Toolbox and GEOEAS, both in the public domain. Consult the SOFTFAQ for more details. However, both do not have a simulation capability. Variowin for Windows is a freeware to do interactive variogram modeling only. Again, consult the AI-GEOSTATS web site for more details.

Back to top

[4] The Big Picture: issues and problems

[4.1] Why should I use kriging instead of simple interpolation?

In a nutshell:

Deterministic interpolation techniques such as inverse distance and triangulation do not take into account a model of the spatial process, or the variogram. You might be able to get a map and think that you're modeling the spatial process, but then again, you might not. Furthermore, kriging allows you to quantify the quality of your predictions via the kriging variance. You can do the same for deterministic techniques, but it's quite tedious. You'd still need to model the variogram and derive the estimation weights. Another advantage of kriging is that you can take into account clustering (redundant data) and possible anisotropies much more easily than, say, inverse distance techniques. Furthermore, more advanced geostatistical techniques such as indicator kriging and simulations allow you to quantify soft" or qualitative information in a quantitative manner. To do the same using triangulation, for example, would require a lot of tedious trial and error and a lot of dummy data.

Of course, kriging implies more work. You'd need to model the variogram, which is usually time consuming. Nevertheless, the results from kriging are generally of higher quality and more realistic compared to techniques such as inverse distance and triangulation.

[4.2] What is the theory of regionalized variables?

Geostatistics as a discipline was formalized by G. Matheron at the Ecole des Mines. He calls it the theory of regionalized variables. A regionalized variable is any variable distributed in space (or time). The theory says any measurement can be viewed as a realization of a random function (or random process, or random field, or stochastic process). This theory forms the underpinnings of geostatistics.

[4.3] What's wrong with the random function model?

Nothing. Although some earth scientists are uncomfortable with the idea. To them earth science data sets are not random, they are merely complicated, and because of lack of data, we do not have a clearer picture of the true population. They prefer to think that an earth science data set is essentially deterministic.

Nevertheless, the random function model _works_ in pratice. It is an effective theoretical crutch to incorporate our lack of knowledge of the true underlying population. The measurements can be viewed as a realization (drawn by God?) of the random function model. It is only a model. There is nothing wrong with it. And it has shown to be an effective model that we can use to characterize uncertainty. The least we can do is understand the model.

[4.4] What is the region of stationarity?

To put it in simple terms, the region of stationarity is the region where you want to make your estimates. It has to coincide with the extent of the population that you are modeling, but there is no test that can prove or disprove this. The smaller the data set, by necessity, the bigger the region of stationarity; otherwise, one wouldn't be able to make any estimates at all.

Briefly, geostatistics assumes second order stationarity in the data. It means that the mean has to exist and is constant and independent of location within the region of stationarity, and that the covariance has to exist, and is only dependent on the distance between any two values, and not on the locations. Or, if the covariance doesn't exist, the variogram is assumed to exist, and the increments depend only on the difference between any two points, and not on the locations (this is called the intrinsic hypothesis).

Choosing a correct region of stationarity ensures that your estimates (whether of the variogram or of the value at any unsampled location itself) is valid. If it's not valid (e.g., you're mixing two different populations), then so would your estimates.

Back to top

[5] Spatial correlation: general

[5.1] What is a variogram?

The variogram (or its equivalent, the covariance) underpins all of geostatistics. You use the variogram to model the way two values in space or time are correlated. Most people intuitively know that two values in space that are close together tend to be more similar than two values farther apart. Univariate statistics cannot take this into account. Two distributions might have the same mean and variance, but differ in the way they are correlated with each other. Geostatistics allows one to quantify the correlation between any two values separated by a distance h (usually called the lag distance) and uses this information to make predictions at unsampled locations. Variogram modeling is a prerequisite for kriging, or making predictions. Variogram modeling is a modeling of such a spatial correlation structure.

Variograms are usually described using two parameters. The range is the lag distance at which all successive values are independent of each other. The sill is the variogram value corresponding to the range. When we want to model the variogram, we have to specify a type of model as well.

[5.2] How is the variogram related to the covariance?

The variogram is related to the covariance in the following manner:

- gamma(h) = covariance(0) - covariance(h)

The user can therefore derive one from the other. Of course, the covariance has to exist (strict stationarity). In general, the larger the lag distance, the smaller the covariance. For the variogram, the larger the lag distance, the larger the variogram value.

If it doesn't exist (the variable has an infinite capacity for dispersion), then we assume that only the variogram exists and is stationary (the intrinsic hypothesis). Existence of the covariance implies existence of the variogram, but not vice versa. Brownian motion is an example of a variable that has an infinite capacity for dispersion. Use of the variogram is not popular outside of geostatistics.

[5.3] Why is it sometimes advantageous to use the covariance?

Because you can take into account possibly changing means (non-ergodic relationships). Furthermore, use of the standardized covariance (the correlogram) is useful to take into account the lag variances. In a cross-covariance, you can use more data to calculate the means, if one of the data is more abundantly sampled. Because of the above reasons, use of the covariance results in generally more stable spatial correlation structures.

[5.4] What is a correlogram?

The covariance measure can be normalized to a value of 1 if divided by the variance of the data. This gives us the correlogram, a measure traditionally used by time series analysts. Since the covariance is normalized, it is also a good means to compare different spatial correlation structures which have different magnitudes of values. Any other advantage of the covariance also applies to the correlogram, which includes possibly taking into account non-ergodic relationships in the correlation measure (see Srivastava), and the use of more data to calculate the mean.

[5.5] What is a madogram?

Instead of taking the mean squared differences, one can also calculate the mean absolute difference and obtain the madogram. Avoiding the squared term results in a more robust measure of the correlation structure. Other measures include the use of the median squared difference, or the median absolute difference, instead of using the mean.

[5.6] I have no spatial correlation, should I use geostatistics?

An absence of spatial correlation or a pure nugget effect is quite _rare_ for spatial data sets. In most cases, due to large sampling distances (for example, in an offshore oil field), such a correlation structure cannot be resolved (range smaller than the average sample spacing). The question is not so much "should I use geostatistics" but more of "what kind of variogram model should I use to make my predictions?" In some cases, the variograms do not exhibit a _pure_ nugget, but an _apparent_ nugget, e.g. due to small number of pairs at each lag distance. In both cases, a variogram model can still be assumed, from a subjective experience or from analogous situations (e.g., an outcrop in petroleum sciences). Cross-validation is a tool that can help pinpoint the most suitable structure for prediction purposes.

[5.7] What is an anisotropy?

If the variable exhibits different ranges in different directions, then there is a geometric anisotropy. For example, in an aeolian deposit, permeability might have a larger range in the wind direction compared to the range perpendicular to the wind direction.

If the variable exhibits different sills in different directions, then there is a zonal anisotropy. For example, a variogram in a vertical wellbore typically shows a bigger sill than a variogram in the horizontal direction.

Some variograms are a combination of both geometric and zonal anisotropies.

Anisotropies can be accounted for using anisotropic variograms. Typically, to detect anisotropies, variograms are calculated in different directions, and a rose diagram (a diagram of ranges in the different directions) is plotted out. The anisotropy ratio is the ratio between the smallest range and biggest range (these directions are assumed to be approximately perpendicular to each other). A ratio of one denotes an isotropic variogram (same variogram in all directions).

Back to top

[6] Spatial correlation: implementation

[6.1] How do I calculate the variogram?

Define a lag increment (or spacing between any two points in the data). Say 100 ft. From the measurements, take all pairs of values separated 100 ft. apart. For each pair calculate the difference, and then square it. Sum up all the differences and divide by twice the number of pairs. This gives you the value of the variogram for that particular lag increment or distance.

Do the same for other lag distances, say 200 ft, 300 ft, 500 ft, and so on. Plot out the variogram value versus the lag distance. What you get is called the experimental variogram, or simply the variogram.

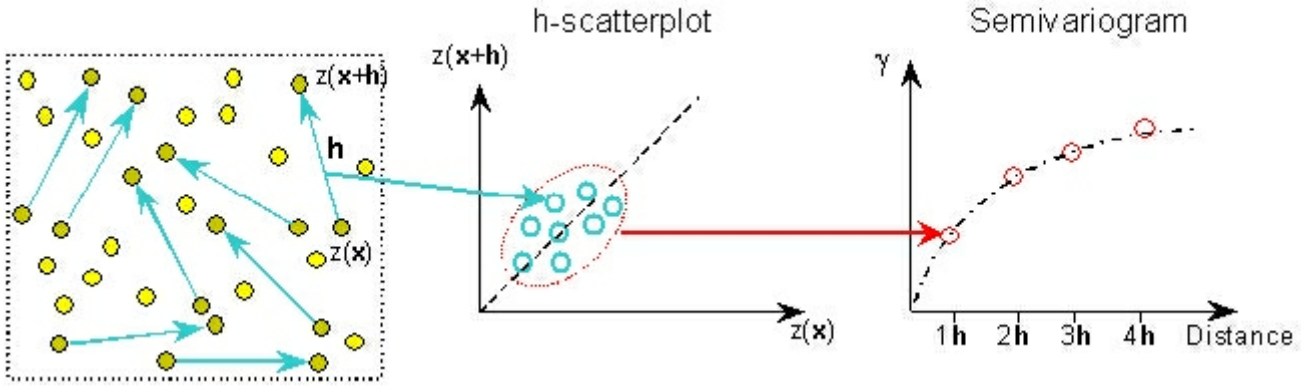

Fig.1. *From the sample map to the experimental variogram*

In general, at smaller lag distances, the value of the variogram would also be smaller. For larger lag distances, the value of the variogram would be larger. This is because, values separated at small distances tend to be more similar compared to values separated at a larger distance. However, at a lag distance called the "range" the variogram would typically stabilize. The value of the variogram at this point is called the sill. At distances greater than the range, any two pairs of values are independent of each other.

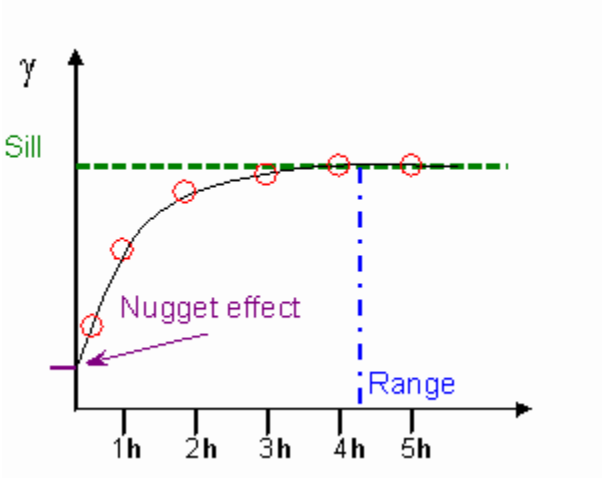

Fig.2. *Parameters of the semivariogram*

[6.2] What is a lag tolerance?

If the data are irregularly spaced, there's no way you can get pairs of values exactly 100 ft apart or exactly 200 ft apart. What you do is you define a lag tolerance, normally taken to be half the lag increment used. For a lag increment of 100 ft, the lag tolerance would be +/- 50 ft, or one simply takes all pairs separated 100 ft +/- 50 ft apart. This ensures that you can get enough pairs for a particular lag distance. When you plot out the variogram, the lag distance will be the average of all the distances between pairs falling within that tolerance, so it wouldn't be 100 ft. The bigger your tolerance, the more pairs you can define, and the smoother looking the variogram. The smaller the tolerance, the fewer the number of pairs and the variogram wouldn't look as smooth.

[6.3] What is a direction tolerance?

By defining a tolerance for the direction, one can calculate directional variograms. Normally, one chooses an azimuth or direction to analyze, say 45 degrees. Then all pairs in this direction, plus some tolerance on either side of the azimuth (say 30 degrees) are taken and the variogram calculated. A tolerance of 90 degrees simply means an omnidirectional variogram, regardless the azimuth. The bigger the tolerance, the more pairs that you would get.

[6.4] What are some of the techniques I can use to clean up my variograms?

Variogram modeling requires experience and practice. Practical tips one can try out differ from data to data, but they include:

- Checking for enough number of pairs at each lag distance (from 30 to 50).
- Removal of outliers using an H scatterplot (see Figure 1).
- Truncating at half the maximum lag distance to ensure enough pairs.
- Using a larger lag tolerance to get more pairs and a smoother variogram.
- Starting with an omnidirectional variogram before plunging into directional variograms. There is no reason to expect structure in the directional variograms if the omnidirectional variogram looks awful.
- Using other variogram measures to take into account lag means and variances (e.g., inverted covariance, correlogram, or relative variograms).
- Using transforms of the data for skewed distributions (e.g. logarithmic transforms).

- Using the mean absolute difference or median absolute difference to derive the range

Back to top

[7] Kriging: general

[7.1] What is (co)kriging?

Kriging is a geostatistical estimation technique. It uses a linear combination of surrounding sampled values to make such predictions. To make such predictions, we need to know the weights applied to each surrounding sampled data. Kriging allows you to derive weights that result in optimal and unbiased estimates. Within a probabilistic framework, kriging attempts to: (a) minimize the error variance; and (b) systematically set the mean of the prediction errors to zero, so that there are no over- or under-estimates.

There are some cases when other, usually more abundantly sampled data, can be used to help in the predictions. Such data are called secondary data (as opposed to primary data) and are assumed to be correlated with the primary data. For example, we can predict porosities based not only on the well measured porosities but also on seismically derived porosities. In this situation, we can try a cokriging. To perform cokriging, we need to model not only the variograms of the secondary and primary data, but also the cross- variograms betwen the primary and secondary data.

Kriging is expressed in a so-called "kriging system", which relates the covariance between the samples, the covariance between each sample to the location to be estimated, and the unknown weights. The covariance matrix is inverted to solve for the weights.

[7.2] What is an unbiasedness constraint?

In a nutshell:

In kriging, the expected value of the prediction errors is systematically set to zero so that there is no over- or under-estimation. If there is, then our predictions will be biased. In ordinary kriging, this unbiasedness constraint is equivalent to letting the sum of the weights add up to one.

[7.3] Why would kriging a log-transformed variable introduce a bias?

Simply kriging a log-transformed would introduce a bias because in the kriging, we're assuming that the prediction is the mode/mean/median of our minimized error distribution. But, by transforming the transformed value back to its original value, it wouldn't correspond to the mean anymore (it'll probably be near the median)! Therein lies the bias.

[7.4] What is universal kriging?

Universal kriging is used if there is a gradual trend in the data, say, the top or roof of a petroleum reservoir. If there is a trend, then the mean is no longer constant, and the variogram or covariance is no longer the appropriate tool to model the spatial correlation structure. What we need is a variogram of the residuals, and a model to describe the shape of the trend (we don't need to know the trend itself).

The variogram of residuals can be approximated from the raw variogram, provided it is:

- derived from a location with a small trend component
- or only the short scale behavior is used

In universal kriging, we assume that we know the shape of the trend (usually expressed as a function of the coordinates, either first order -- linear drift -- or second order -- quadratic drift). Kriging is then performed on the residuals.

[7.5] What is IRF-k kriging?

The intrinsic random function of order k can also be used to make spatial predictions of variables with a trend component. The idea is to use higher order differences to filter out polynomials of degrees 1, 2, 3, ..., n. The function for such differences is called the generalized covariance of order k.

The main disadvantage of IRF-k theory is that it is not graphical anymore, and would have to be performed using an automatic statistical fitting procedure. Furthermore, not many software packages offer this facility. However, the results, theoretically at least, are usually more accurate.

[7.6] Is kriging an exact interpolator?

Yes, if you're willing to assume that there is no measurement error (replicate measurements) and that any sharp discontinuity at the origin (the nugget effect) is caused by small scale variations that cannot be resolved by the sample spacing.

[7.7] Effect of the range on predictions.

A larger range means more continuous behavior. The variable is very well correlated in space. Predictions result in fairly smooth maps of the variable of interest.

[7.8] Effect of the model on predictions.

Think in terms of the shape of the variogram for early lags. An Gaussian model is more continuous than a spherical. Same effect as in 6h, i.e., the better correlated the variable, the smoother looking the resulting maps.

[7.9] Effect of the sill on predictions.

Rescaling the sill by 10 or 100, for example, would not change the values of the predictions. Your maps would still look the same. What changes, though, are your prediction variances. The higher the sill value, the higher the prediction variances.

[7.10] Kriging with nugget model.

Using a 100% nugget model is similar to regression with uncorrelated independent variables, i.e., a kind of trend surface mapping. Predictions, however, are conditioned to the actual data at the data locations. Maps obtained with a nugget model look continuous, except at the data locations where discontinuous jumps to the actual data values occur. This is because kriging satisfies the exactitude property, or data are honored in the predictions.

[7.11] What is a cross validation?

A cross-validation is a way to check the model assumptions used in the kriging. In a cross-validation, one specifies a variogram model and a search neighborhood. After that, values are kriged at each sampled location, assuming that particular sample is missing. Then the kriged values and the true values are compared. The difference between these two values is called the cross- validated residual. The idea is to use these residuals to assess the performance of the model assumptions. (Is my variogram model appropriate? Is the search neighborhood too small? Am I over- or under-estimating in some regions? Should I incorporate a trend model using universal kriging? Are my prediction errors comparable?) Cross-validation does not prove that the variogram model is correct, merely that it is not grossly incorrect [Cressie, 1990].

Back to top

[8] Kriging: implementation

[8.1] What is a search neighborhood?

Kriging seldom uses all of the sampled data. The more the data used in the kriging, the bigger the matrix that would have to be inverted, and the more time consuming the kriging process. Using a search neighborhood, we are limiting our estimates to just the data within some predefined radius of our point of estimation. We call this the search radius, and furthermore the search neighborhood need not be a circle or sphere; one can have elliptical or ellipsoidal (in 3D) search neighborhoods to account for possible anistropies in the data.

However, in some cases, it would be beneficial to use all of the data. This results in a unique neighborhood (all the data is used, e.g., if number of samples is less than 100). The advantage is that the kriging matrix is inverted only once. For more than 100 samples, it is recommended to use a plain search neighborhood. For kriging with finite ranges, sparse matrix techniques can also be used with a unique neighborhood and with a lot of data without excessive computer cost. For kriging without finite ranges, search neighborhoods would still be required.

Also implicit in the use of the search neighborhood is that we are assuming stationarity only within the search window (quasi-stationarity). For example, this can be useful if there is a trend in the data. By assuming a small enough search window, such a trend component can be ignored in the estimation.

[8.2] Why do I need to use variogram models instead of the real thing?

In a nutshell:

To ensure that kriging the variance is positive or zero (negative variances do not make much sense), the spatial correlation must be based on some "positive definite" theoretical models. Such models are fit to the actual (or experimental) variograms, and include spherical, exponential, Gaussian, etc models. Combinations of such models are also valid. Furthermore, values of the variogram might be needed at lag distances where measurements are not available, so a function which relates the lag distance to the variogram value will therefore be useful.

[8.3] What is the kriging system?

To solve for the kriging weights, a "kriging system" would have to be prepared. In particular, the covariance matrix is inverted and multiplied with the point-to-location covariances (or the right hand side matrix). The covariance matrix models interaction between sampled data (redundancies are automatically catered for). The RHS matrix models the concept of statistical distance between all neighboring data and the point to be estimated.

The kriging system might be familiar to multiple regression modelers, who typically deal with independent predictor variables. In fact, the kriging system can be decomposed to the regression equivalent.

[8.4] What is the linear model of coregionalization?

The linear model of coregionalization is a technique that ensures that estimates derived from cokriging have a positive or zero variance. To ensure this, the sill matrices for each basic structure must be positive definite. As an example, for the two variable case, we can check whether the linear model of coregionalization is honored:

- Gamma_x(h) = 10 + 20 Sph + 40 Exp
- Gamma_y(h) = 20 + 35 Sph + 55 Exp
- Gamma_xy(h) = 7 + 10 Sph + 30 Exp (cross-variogram)

For the nugget structure:

- determinant = (10 x 20) - (7 x 7) = 151 > zero --- OK
- all diagonals are positive (10 and 20) --- OK

For the spherical structure:

- determinant = (20 x 35) - (10 x 10) = 600 > zero --- OK
- all diagonals are positive (20 and 35) --- OK

For the exponential structure:

- determinant = (40 x 55) - (30 x 30) = 1300 > zero --- OK
- all diagonals are positive (40 and 55) --- OK

Since for all structures the determinants are greater than zero and the diagonals are positive, then the linear model of coregionalization is honored.

Back to top

[9] Conditional simulation: general

[9.1] What's the difference between simulation and estimation?

In estimation, we want to make the best possible estimate at the unsampled location. We do this via kriging, by minimizing the error variance. However, there is no guarantee that the map obtained using kriging has the same variogram and variance as the original data (i.e. we're smoothing the map, and not retaining the dispersion characteristics of the original data). Simulation allows us to come up wih theoretically an infinite number of realizations (or renditions)

of the map each of which has approximately the same variogram and variance of the original data. Theoretically, the average of a large number of simulated maps would look like our original kriged map. One important use of simulations is in the petroleum industry. Simulated maps allow the petroleum engineer to come up with a range of fluid flow predictions all of which are plausible given the uncertainty in the distribution of the geological properties. Simulation strives for realism; estimations strive for accuracy.

[9.2] What are some popular simulation techniques?

Sequential Gaussian simulations and Gaussian simulations. Non-parametric simulation techniques such as sequential indicator simulations and probability field simulations are becoming more and more popular. Some of these algorithms are available in the public domain software package, GSLIB (refer question 3).

Back to top

[10] Conditional simulation: implementation

[10.1] How does Gaussian Simulation work?

Transform the data into a normal distribution. Then perform variogram modeling on the data. Use the variogram model to krige the data at all grid locations. This gives us the "base" map. In simulation, the objective is to generate multiple realizations of the map each of which would honor the actual data (i.e., a conditional map), and approximately the same variogram and distribution. To generate each realization, an unconditional map is simulated (e.g., using the Turning Bands algorithm) which honors the variogram model but not the data at the data locations. Then using the same unconditional map, another map is obtained by kriging the values at the same locations as the actual data. Therefore, at each grid node we would have a simulated error: the difference between the kriged and simulated value. We add this error to the "base" map at each grid location. This gives us the first realization, which can then be back-transformed to its original distribution. Repeat for the other realizations using a different random number sequence to generate multiple realizations of the map.

[10.2] How does Sequential Gaussian Simulation work?

Transform the data into a normal distribution. Then perform variogram modeling on the data. Select one grid node at random, then krige the value at that location. This will also give us the kriged variance. Draw a random number from a normal (Gaussian) distribution that has a variance equivalent to the kriged variance and a mean equivalent to the kriged value. This number will be the simulated number for that grid node. Select another grid node at random and repeat. For the kriging, include all previously simulated nodes to preserve the spatial variability as modeled in the variogram. When all nodes have been simulated, back transform to the original distribution. This gives us the first realization. Repeat for all the other realizations using a different random number sequence to generate multiple realizations of the map.

[10.3] How does Sequential Indicator Simulation work?

In sequential indicator simulation, the variables are divided into several thresholds, for example, corresponding to the 20th, 40th, 60th, and 80th percentiles in the distribution of values. Indicators are then coded from the raw data, where the number one would be assigned if the value falls

below a certain threshold, and zero otherwise. Eventually for each threshold at each sampled location, a distribution of ones and zeros would be obtained. Indicator variograms are then modeled for each of these threshold values, the number of variograms corresponding to the number of thresholds used. Sequential indicator simulation proceeds as follows: a random location is chosen and indicator kriging is performed for each threshold at that location. After correcting for possible order relation violations (due to non-monotonic cumulative probability density functions), the indicator kriging will define a full cpdf at that location. One then draws a simulated value from this cpdf based on a uniform random number generator. The sequential simulation then moves on to another location with this simulated value assumed as "hard" data until all nodes are simulated. This will give us one realization.

[10.3] How does Probability Field Simulation work?

Sequential Indicator Simulation is slow because kriging is repeated from realization to realization. In Probability Field Simulation, the kriging is performed once to estimate the histogram at each grid location. Then using a uniform random number generator, we draw a simulated value at each node using the estimated probability density function (or histogram). The only other constraint to honor is that the distribution of uniform random numbers should honor the variogram of the data, or simply that they should be correlated. This makes for fast conditional simulations compared to Sequential Indicator Simulation.

Back to top

[11] Nonparametric geostatistics

[11.1] What is non-parametric analysis?

In conventional kriging, we make a prediction, derive the kriging variance, assume an error model (usually Gaussian), and then derive uncertainties in our predictions. In other words, we're making a prediction first before assessing the uncertainty. In non-parametric analysis, we try to assess how much we don't know before making a prediction. Furthermore, we do not have to rely on some assumed distribution model for the errors. In essence, we're letting the "data do the talking."

Non-parametric geostatistics is in the realm of non-linear geostatistics, because instead of using linear combinations of the data, we're using linear combinations of _functions_ of the data, usually the indicator function. Indicator coding, indicator variograms, and indicator kriging and simulations have several powerful advantages, including:

- no dependence on assumed models
- inclusion of extreme values in the variogram modeling
- variogram modeling as a function of the magnitude of the data
- joint spatial uncertainty

In a nutshell, in a non-parametric technique, we:

- divide the data into several cutoffs or thresholds corresponding to the classes of the data
- code each data into an indicator value (either zero or one)
- estimate the local histogram at any particular location (based on variograms at each cutoff) using indicator kriging
- use the resulting histograms to derive our estimates or to perform simulations.

The second step is especially amenable for the inclusion of "soft" or qualitative information, i.e., subjective experience, beliefs, data with a lot of error, and so forth.

[11.2] Can I use ordinary kriging to krige indicators?

Yes. Indicator kriging is the ordinary kriging of indicators. Instead of kriging continuous variables, we now krige indicator (binary) variables. The results are generally between 0 and 1, although there is no guarantee that they will fall within this range. These values can be viewed as probabilities of occurrence, with a higher value denoting a higher probability (or vice versa depending on how the indicators have been defined).

Back to top

[12] Special concepts: fractals

[12.1] What is fractional Brownian motion or fractional Gaussian noise?

These are fractal models commonly used to simulate properties. *fBm* has an equivalent in geostatistics (the power model). *fGn* has its own variogram model. Fractals assume statistically self-similar distributions of properties. *fBm* models have been used to generate artificial mountains, lakes, clouds, etc and the results look realistic, hence the motivation to use them to simulate geostatistical data. By using a true fractal model, we are assuming that we can determine the variance of the property at one scale based on the variance at any other scale (statistical self-similarity). *fGn* is the approximate derivative of *fBm*, and appears more irregular (more jagged looking mountains). To use fractals, one typically:

- estimates the fractal co-dimension, or H value
- uses the H value to generate artificial traces of the geostatistical property (for example, using the technique of Successive Random Additions, or SRA)

In the petroleum sciences, the H value is usually calculated from data collected from well bores, since these data are usually quite abundant. Most properties in Nature exhibit an H value from 0.6 to 0.9, indicating infinite correlations across all scales.

[12.2] What is fractional Levy motion?

fLm is an even more powerful generalization of *fBm*. Here, the underlying increments are assumed to have a Levy-stable distribution, or tails that decay more slowly than the Gaussian distribution. It's power lies in being able to model sharp or abrupt changes in fnon-stationary sequences (for example, a porosity log). The main difference is that *fLm* modelers assume a particular trace is already *fLm*; they just calculate width of the distribution of increments. Typically, H values obtained using *fLm* techniques are in the order of 0.1 to 0.5.

Back to top

[13] Special concepts: combinatorial optimization

[13.1] What is simulated annealing/genetic algorithms?

Both are combinatorial optimization schemes (COS) that can be used to simulate geostatistical data. In both techniques, the spatial constraints are expressed as an objective function to be minimized. COS are flexible in that any manner of constraints, provided they can be suitable expressed in the objective function, can be incorporated. These constraints include data such as well test permeabilities, connectivity measures, production data, etc.

Because constraints can be several, COS hold a lot of promise. Nevertheless, such algorithms are plagued by their speed; the more complex the objective function, the slower the technique.

In general, for annealing and genetic algorithms, a random distribution of values with a specified histogram is generated. The idea is to "perturb" the system (for example, by swapping values in annealing, or by mutation and cross-over in GAs) with the hope that it would have a spatial correlation structure more similar to the target spatial correlation structure. If this is the case, the system is said to have a lower "energy" and the perturbation is accepted. In most schemes, perturbations that result in an even higher objective function are sometimes accepted to climb out of possibly local minima.

Among other advantages of COS are:

- no Gaussian assumption
- can honor raw instead of theoretical variograms
- can use different variograms in different directions to cater for anisotropies and different spatial behavior

[13.2] How does simulated annealing work?

Generate a random distribution of values with a specified distribution. Define your constraints (usually just the variogram model, but it could be more). The idea is to honor the distribution and the constraints for each realization. To do that, we can "perturb" the system and determine whether the perturbation introduced would bring us nearer to the kind of image we want (i.e., one that honors the specified variogram model). One way of perturbing the system would be to take values at two locations and swapping them. This preserves the histogram of the image. We then calculate the objective function (defined as the difference of the variogram values between the ideal image and the current image). If the objective function decreases, accept the swap. If it increases (the new image is worst than the previous image), the swap can also be accepted, but with a specified probability distribution known as the Metropolis condition. The point here is that although we want to reach global minima as soon as possible, we don't want to reach it too soon such that the solution obtained would be local instead of global minima. The probability of allowing the objective function to increase is related to the magnitude of the increase and the number of iterations that have already been performed. Increases are tolerated early in the procedure, and small increases are tolerated than larger ones. The final realization is obtained when the solution falls under some prespecified tolerance for the objective function. Other constraints can be included as well, but generally the algorithm gets slower the more complex the objective function.

Back to top

[14] Acknowledgments

[14.1] Acknowledgments

Grégoire Dubois - for inventing such a useful geostatistical resource (AI-GEOSTATS) and reviewing the FAQ

Pierre Petitgas - for answering part of question 4a succinctly

Michel Maignan - technical input and cheerleader

Maribeth Milner - for reviewing the FAQ

Subscribers of ai-geostats - for so many interesting discussions that provided material for this FAQ list

Back to top

[15] Copyright

[15.1] Author

Rathsacksvej 1, 5. th.

DK-1862 Frederiksberg C

DENMARK

syed@spemail.org

[15.2] Copyright Notice

[15.3] No Warranty

Back to top