

Week 11-12 Lecture Material Introduction to Stochastic Simulation

1. Why Simulation?

- preserves the flavor of real world variability: the extreme values of the regionalized variable, the spatial correlation observed in the data
- if simulation is "conditional", it honors the observed data exactly, without the attendant smoothing of the interpolated estimates, as in kriging
- as in kriging, permits incorporation of ancillary data to constrain the interpolation, and provides visual and quantitative measures of the uncertainty of the estimated variable ("cosimulation", analogous to cokriging)

- note: data reproduction vs. honoring - data are honored where the underlying trend or pattern is reproduced (eg: as in kriging), whereas data are reproduced where predicted values are forced to be equal to observed values

2. Difference between Simulation and Kriging

- kriging produces just one map of estimates which is "best" (and hence unique) in a statistical sense, but which does not reproduce global statistics (histogram, variance, covariance)
- kriging is a global estimator, in that its estimate represents all the data within a defined area (the local neighborhood); it honors data points but does not exactly reproduce data values
- kriged maps are good to show smooth variations and underlying trends but not the inherent local variability of the data
- kriging provides an incomplete measure of local accuracy (ie. kriging "variance" may not represent a Gaussian distribution of errors) and no measure of joint accuracy when several locations are considered together

- simulation is a local estimator, which obeys a global spatial correlation function but also exactly reproduces data; the goal is reproduction of global patterns of spatial continuity and global statistics (histogram, covariance), rather than local accuracy
- simulation is good at showing the nature of local variability and global patterns of local variability, such as the degree of connectedness of extreme-valued regions
- simulation produces any number of statistically equivalent maps which, when taken together, define the local estimation uncertainty as well as the global pattern of uncertainty

3. Type of Simulation Approaches

- simulated variables can be continuous or categorical
- conditional simulation reproduces observed data
- unconditional simulation only reproduces global statistics (histogram, variogram, etc.)
- categorical variables can be simulated to honor specific geometrical patterns (object-based)
- continuous variables can be simulated to honor a specified covariance model
- variables can be simulated through optimization processes starting from scratch or from previously-simulated images, to honor additional, external, constraining data

- hybrid simulation approaches can be implemented to model different styles of spatial variability and to impose external conditional constraints (eg: based solely on geologic conceptual models, a categorical, object-based simulation could be used to model the large-scale reservoir architecture and spatial distribution of lithofacies types; this could be followed by a continuous variable, unconditional, simulation of permeability within each lithofacies separately; finally, a conditional optimization simulation could be performed on the simulated permeability so as to modify it to honor well-test pressure / drawdown data)

- the GSLIB library contains a number of different simulation algorithms, no single one of which is sufficiently flexible to be a panacea for simulation problems

- three types of simulation algorithms will be discussed: sequential Gaussian (SGSIM), sequential indicator (SISIMPDF), and simulated annealing (SASIM)

4. Gaussian Simulation

- the ease and convenience (both theoretical and practical) with which Gaussian distributions can be described and handled statistically makes a Gaussian R.F. model very appealing

- this class of simulation algorithms is restricted to continuous (ie. non-categorical) variables

- if the pdf's of all variables that constitute the physical process (eg: a regionalized variable plus its local uncertainty or variance) are normally distributed and independent, a Gaussian R.F. model can be invoked under certain conditions (see Deutsch and Journel, 1997, p. 139-140), the most important of which is that the variables are univariate and bivariate normal and independent (ie. extreme values not correlated in space, no proportional effect)

- the assumption of normality is not restrictive for real-world data, since a normal-score transform can be applied to make any data set normally-distributed

- the most serious limitation for real data is that a regionalized variable at a given location is often not independent eg: it may be correlated with its local variance (via a proportional effect) and/or it may have strong local spatial correlation (eg: interconnected high or low values); a method to check on the latter characteristic is described in Deutsch and Journel (1997, p. 142-143) and is strictly necessary if the Gaussian simulation approach is to be defended; even if a proportional effect exists, it is not a problem if the proportional effect is weak or nonexistent within the moving search neighborhood

- despite such caveats, the Gaussian simulation approach is widely used; its intuitive and algorithmic simplicity define its widespread appeal

5. Sequential Simulation Algorithm (see Deutsch and Journel, 1979, p. 144-146)

- sequential simulation is a popular class of algorithms that is in widespread use; the approach is used in many simulation approaches, not just the Gaussian version

- the sequential simulation approach can be broken down conceptually into three parts: A) the initialization process; B) a global random walk process; and C) a local search and conditional random estimation process

A. Initialization

- determine the univariate cdf from the entire study area, using declustering if necessary
- perform a normal-score transform of the data using a standard-normal cdf
- define a regular grid network of nodes at which estimates will be made; data points may be re-located to these grid nodes (for faster computation) or not (for maximum accuracy); typically, this grid network will be much finer than was used in kriging, because simulation attempts to reproduce the nature of local variability and uncertainty

B. Global random walk process

- using a random number seed, start at a random location on a grid network
- with the random number generator, move to subsequent random locations after C) has been completed, but never visiting the same grid node twice

C. Local search and conditional estimation process

- within a prescribed search neighborhood, look for a prescribed number of nearest neighbors comprising both original data and/or previously simulated values
- within the search neighborhood, use the variogram model of the normal-score data with simple kriging to estimate the mean and variance at the grid node location; this defines the cdf of the regionalized variable at that location which is said to be conditional to the spatial correlation structure and any data or previously simulated values that occur in the search neighborhood; if neither data or previously simulated nodes exist within the search neighborhood, the cdf is conditional only to the global histogram so the value estimated at such a grid node will be a value drawn at random from the cdf estimated from the global histogram
- draw a simulated value at random from the conditional cdf (ccdf) and assign it to the grid node
- go to the next randomly chosen grid node and repeat C) until all grid nodes have been filled

- at the end of this process, the simulated normal-score values are back-transformed into simulated values for the original variable
- if multiple simulations are to be performed, the above algorithm is repeated, starting at a different initial grid location and visiting grid nodes in different order each time

6. Limitations of Gaussian Simulation

- in addition to the caveats discussed in Section 4, above, Gaussian simulation tends to overestimate local spatial entropy, creating more apparent variability than is actually present
- it is also not sufficiently flexible to handle a mixture of different normal populations; because of this, it does not permit cosimulation of multiple variables (a process analogous to cokriging); furthermore, it does not allow the incorporation of soft information (such as categorical data) during simulation (except for external constraints which can always be imposed on the variogram or global histogram)

- to go beyond the limitations of the Gaussian simulation approach, indicator simulation is used

7. Problem Set X. Introduction to SGSIM - Gaussian Simulation (continuous variables)

Week 13 Lecture Material Indicator Variables, Probability Kriging, Indicator Simulation

1. Applications of Indicator Variables

- indicator variables are a special form of a binary categorical variable that can only have values of 0 or 1; it can be regarded as a transformed categorical variable (eg: if rock type is granite, its indicator value is 1, otherwise for all other rock types it is 0), or as a transformation of a continuous variable around a specified threshold
- indicator variables permit rigorous statistical treatment of any categorical geologic variable not collected on interval or ratio scales eg: describing the distribution of limestone clasts in a sedimentary deposit; modeling the probability of caliche occurrence in a sediment; predicting the spatial distribution of lithofacies types in an aquifer
- using indicator variables, correlation structure analysis can be performed on any type of nominal or ordinal data that may show different spatial continuity among classes (eg: spatial distribution of permeability in gravel bars vs. overbank deposits; high and low ore grades in vein deposits)
- a very important application of indicator variables is in the estimation of non-Gaussian distributions
- indicator simulation is always preferred over Gaussian simulation where extreme values (high or low) are expected to be more spatially correlated than medium values
- indicator kriging and simulation are used to model the spatial distributions of various categories of a categorical variable, but are much more important in modeling a continuous variable for which a Gaussian R.F model is inappropriate

2. Categorical vs Indicator Variable

- a categorical variable is a label (eg: rock types are represented by their names, such as granite, diorite, etc. or with arbitrary categories of 1, 2, etc.); like the non-parametric statistical methods based on category counts and category-specific probabilities (eg: chi-square contingency tables), the spatial distribution of categorical variables can be quantitatively modeled by defining category counts and category-specific probabilities (indicator statistics); see Figure 1.a
- an indicator variable is a binary categorical variable that can only be valued as 1 or 0
- any categorical or continuous variable can be represented by an indicator variable, which can be thought of as a type of transform
- if the original variable is already a binary categorical variable (eg: rock type at a location is either siltstone or it is not), the indicator transform is simply 1 (for siltstone) or 0 (not siltstone)
- for a categorical variable that has more than two categories, an indicator transform is performed separately on each category; for example, in the case of three categories (eg: basalt, sediment, limestone), an indicator variable, I_b , is created for the basalt category (= 1 if basalt, 0 otherwise), and indicators I_s and I_l for the other two categories; each indicator variable is defined as 1 or 0 at all data locations, and the order of categories is irrelevant other than the ordering is kept constant
- a continuous variable can also be indicator transformed; in this case, the variable's range is divided into N (= 1, 2, ... N) class intervals of any arbitrary size, defined by a series of thresholds or cutoffs, z_c (eg: 10, 25, 50 and 90 ppm), each of which are used to apply an indicator-transform to the data separately into N indicator variables; the difference is that the sequential ordering information of the original class intervals is retained, and the indicator is defined as 1 if the

variable does not exceed the upper class interval boundary; thus, the indicator transform for class 1 ($z_c = 10$) is 1 if the variable is less than 10 ppm, 0 otherwise; for class 2 ($z_c = 25$) as 1 if the variable is less than 25, 0 otherwise; and so on (see Deutsch and Journel, 1997, p. 152)

- each of the N indicator variables is then treated like a continuous variable eg: by calculating the mean or variogram of the indicator, estimation via kriging or simulation, etc. The principle difference is that each indicator (representing a different class interval or category) can have a different mean or variogram structure, thereby permitting more complex spatial correlation structure to be modeled
- the power of indicator variables lies in their ability to quantify and estimate non-Gaussian continuous variables through non-parametric methods, where parametric methods such as Gaussian simulation algorithms fail to adequately represent non-Gaussian spatial characteristics

3. The Indicator Transform as an Estimator of the cdf of a Continuous Variable

- kriging of a single, binary categorical variable (= 1 if a condition exists, 0 otherwise), produces an estimate of the probability of occurrence of the category (Figure 1.b)
- thus, an indicator transform of a continuous variable divided into in K class intervals produces K estimates of the probability that the variable is below the threshold defined at each class interval boundary; taken together, these probability estimates define the discretized cumulative probability distribution function (cdf) of the continuous variable (see Figure 1.c)
- a sequence of indicator variables, each defined for a successively higher threshold on the cdf, can discretize any cdf by representing it in non-parametric form; because each indicator variable represents only a single threshold on the cdf, it can be kriged or simulated separately (as a binary categorical variable, to produce a probability of occurrence), or as an ensemble with all other indicator thresholds (as a continuous, non-Gaussian variable)
- the advantage is that a cdf of any shape can be modeled with indicators, and that by discretizing data into a limited number of class intervals, the population cdf of any arbitrarily complex unimodal distribution can be quantitatively estimated

4. Kriging of a Single, Binary Categorical Variable

- indicator transforms are very useful for kriging and simulating categorical variables
- in the simplest case, only one indicator variable is involved, and it is assigned a value of 1 if a particular attribute (eg: rock type) is present, 0 otherwise
- the indicator can be viewed as a continuous variable which just happens to have been sampled only where it was valued either 1 or 0; in that case it can be viewed as a continuous variable; ordinary kriging of the indicator variable thereby produces estimates of the probability of occurrence of the attribute (eg: see Johnson and Dreiss, 1989 as an example of this approach)
- direct indicator kriging or indicator simulation, on the other hand, of a binary (0 or 1) indicator variable only produces estimates of the presence (1) or absence (0) of the attribute (Figure 1.b)
- indicator kriging is very similar to kriging of a continuous variable, except that all data and estimates are valued as 0 or 1; kriging returns an estimate of probability in the interval $[0,1]$ which, assuming a Gaussian cdf, allows classification of the estimated value as either 1 ($p > .5$) or 0 ($p < .5$)

5. Indicator Representation of the cdf

- indicator variables can be used to estimate the cdf of a continuous variable, as well as to describe the distribution of multiple categories of a categorical variable
- the cdf of a continuous variable can be estimated by representing it with two or more indicator variables defined at two or more thresholds or category boundaries
- the distribution of a categorical variable is represented by a cumulative histogram or cumulative frequency distribution ("cfd")

- for a continuous variable, the range of the variable $z(x)$ is divided into $i = 1, 2, \dots, n$ classes, each defined by a threshold value, z_c^i ; the value of the indicator variable $I(i)$ representing the i th threshold, z_c^i , is 1 if $z(x) < z_c^i$ and 0 if $z(x) > z_c^i$
- for each threshold, z_c^i , the *mean value of the indicator variable* $I(i)$ defined over all data locations represents the *probability of not exceeding that threshold* (Figure 1.c)
- the estimated cdf is constructed from all thresholds and their indicator means, and represents the probability of being at or below a given z value
- by drawing a random number between 0 and 1, a probabilistic value of $z(x)$ can be selected from the discrete cdf, just as a value was drawn from the Gaussian pdf in Gaussian simulation

- for a categorical variable, the indicator for each category $I(i)$ can only be 1 or 0
- for each category, i , the *mean value of the indicator variable* for that category, $I(i)$ (defined over all data locations), represents the *probability of being in that category*
- a cumulative relative frequency distribution ("cfd") is constructed, in which the non-cumulative relative frequency of each category represents the probability of being in a particular category
- by drawing a random number between 0 and 1, a category can be selected from the "cfd" just as a value was drawn from the Gaussian pdf in Gaussian simulation; in this case, however, the simulated value of the variable, $z(x)$, can be one of only a discrete number of categories

- the indicator approach makes it possible to handle any type of statistical distribution (non-Gaussian) in a non-parametric fashion
- an indicator-discretized cdf also makes it possible to describe the spatial continuity of each indicator threshold or category separately, with different variogram structures, permitting greater flexibility in modeling of complex spatial distributions than Gaussian simulation could achieve
- because of this property, indicator simulation can also be used to model a regionalized variable that is drawn from mixed populations (see Deutsch and Journel, 1997, p.152) eg: if a continuous variable such as ore grade differs among several rock types, strictly it should be segregated and modeled separately within each lithofacies; however, if the variable can be segregated into class intervals that correspond to different rock types, then the indicator variables defined at each of these intervals can be used to estimate a global cdf in which the class intervals loosely represent the different populations; the modeling of spatial continuity is done separately for each rock type, but the overall simulation is performed on the global data set using an indicator formalism
- it is important to note that indicator thresholds classes need not be of uniform width and can be defined on any basis; they can represent physically-meaningful divisions (eg: rock type) or arbitrary divisions made simply to discretize the data (eg: logarithmically-spaced permeability classes)

6. The Sequential Indicator Simulation Algorithm (Goovaerts, 1997, p.395)

- as in the Gaussian case, sequential simulation is fast, robust and conceptually easy to understand
 - it is the most widely-used non-Gaussian simulation technique
 - it is based on kriging estimation in a nearest-neighbor sense, to estimate the local probability distribution while honoring global statistics and variogram structures
 - the approach permits significant spatial correlation between contiguous simulated values
- as in sequential Gaussian simulation (Week 11-12, Section 5), sequential indicator simulation can be broken down conceptually into three parts: A) the initialization process; B) a global random walk process; and C) a local search and conditional estimation process; details of the two algorithms differ in the manner that values at a grid node are estimated (ie. by indicator kriging) and how probabilistic estimates (simulated values) are assigned using the discrete cdf

A. Initialization

- define N appropriate indicator thresholds or categories and perform indicator transforms to assign indicator values at all data locations for each threshold or category
- determine the global cumulative probability for each indicator variable, using declustering if necessary; the cdf or "cdf" so defined represents "prior probabilities" (before any simulated values are created)
- define a regular grid network of nodes at which estimates will be made; data points may or may not be re-located to grid nodes for computational efficiency

B. Global random walk

- using a random number seed, start at a random location on a grid network
- with the random number generator, move to subsequent random locations after C) has been completed, but never visiting the same grid node twice

C. Local search and conditional estimation

- within a local search neighborhood, look for a prescribed number of nearest neighbors comprising both original data and/or previously simulated values
 - within the search neighborhood, use indicator kriging to estimate N probabilities on the local cdf or "cdf" at the grid node location; use within-class interpolation and tail extrapolations to define the entire cdf / "cdf", which represents "posterior probabilities" and is said to be conditional (a ccdf or "ccfd") if previously simulated values were used in their estimation (see Figure 2)
 - draw a random number between 0 and 1 that represents a cumulative probability value
 - based on the ccdf / "ccfd" estimated at the grid node, select the value of $z(x)$ or the category represented by the randomly drawn cumulative probability
 - go to the next randomly chosen grid node and repeat C) until all grid nodes have been filled
- if multiple simulations are to be performed, the above algorithm is repeated, starting at a different initial grid location and visiting grid nodes in different order each time

7. Problem Set XI. Introduction to Sequential Indicator Simulation (indicator variables)

Figure 1.a:

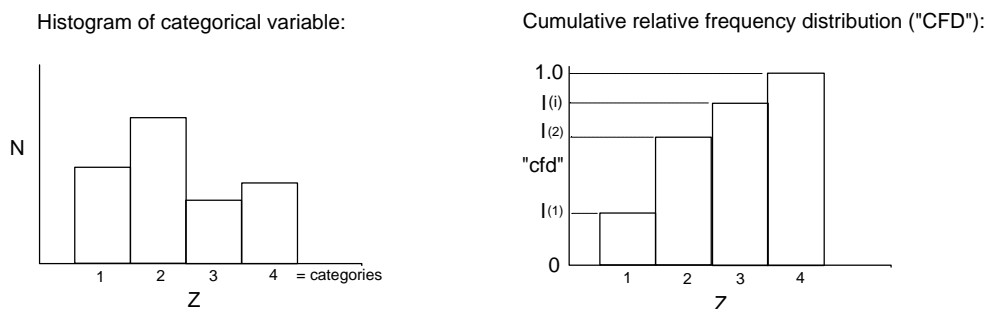
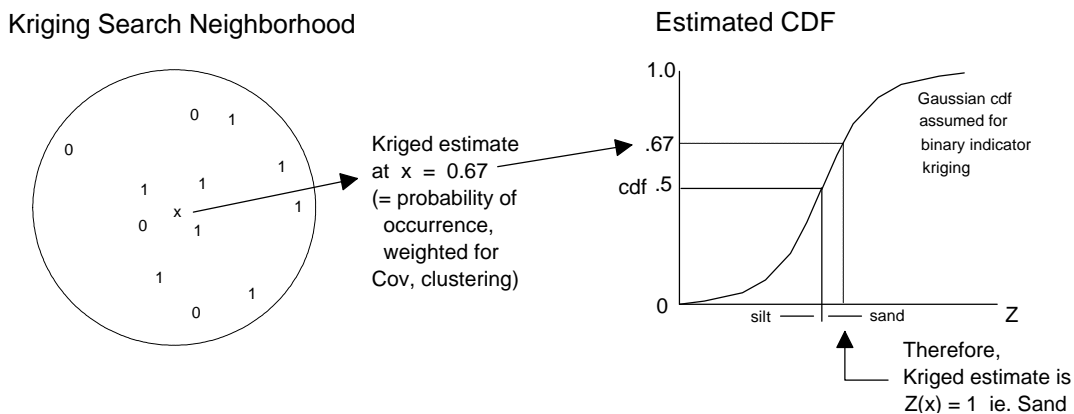


Figure 1.b:

Indicator Kriging of a Single, Binary Indicator (Categorical Variable, eg: sand lenses in silt)
 1 = sand present 0 = sand absent



Fig

Figure 1.c:

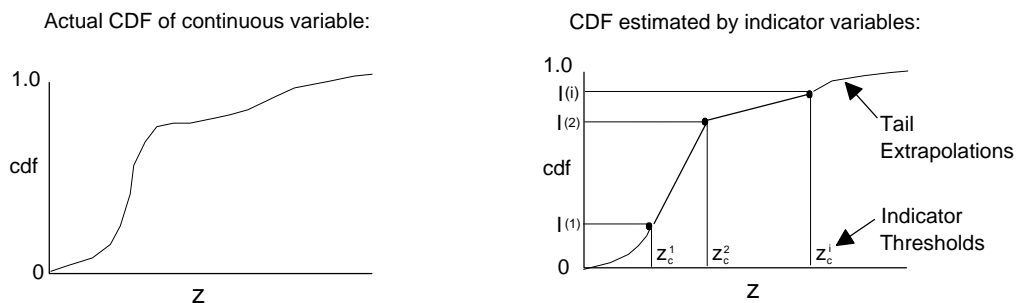
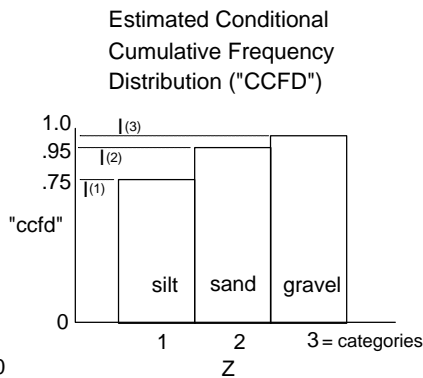
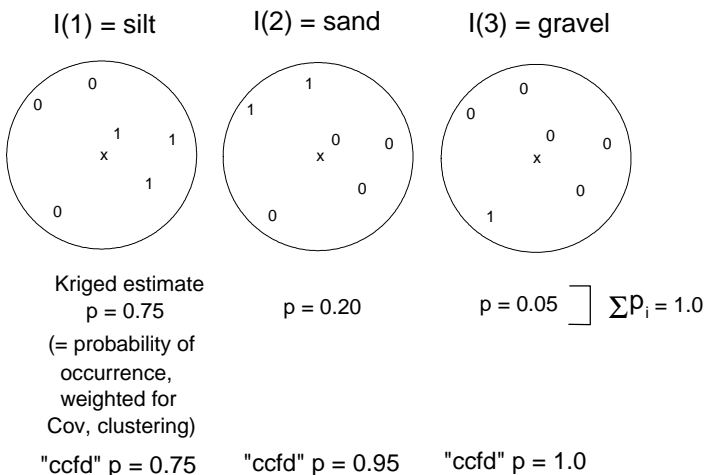


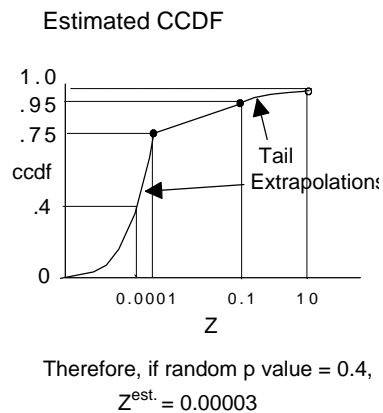
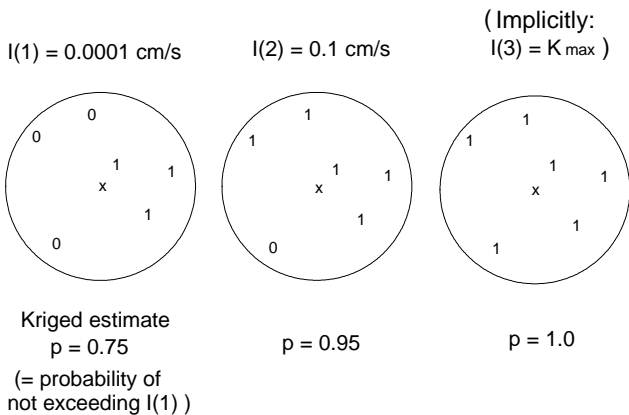
Figure 2:

**Indicator Kriging of N Indicators: Categorical Variable (eg: silt, sand, gravel)
or Continuous Variable (eg: lo-K, hi-K thresholds)**

Categorical Variable: 1 = present 0 = absent



Continuous Variable: 1 if $z \leq \text{threshold}$ 0 if $z > \text{threshold}$



Other Simulation Techniques: Object-Based Simulation and Simulated Annealing

Object-based Simulation:

- used to generate unconditional images of large-scale structures of a categorical variable
- involves distributing geometrical objects in space according to specified probabilities
- the spatial distribution of the object centroids constitutes a point process
- if random processes defining eg: object type, shape or size are attached, a marked point process is generated
- advantages: very flexible; can handle extremely complex geometric shapes
- disadvantages: difficult to condition to local data; calibration is subjective
- see Deutsch and Journel (1997, p.156-157; p.182-183; p. 314-315) for details and example (unconditional) simulations

Simulated Annealing:

- a derivative of optimization algorithms
- used for generating conditional stochastic images of continuous or categorical variables
- combines the ability to reproduce two-point statistics (eg: variograms) and complex geometric conditions
- systematic, node-by-node alteration (perturbation) of a starting image until some prespecified characteristics (an "objective function") are satisfied; perturbations are accepted or rejected depending on whether or not the image changes in the desired way
- the starting image can be random drawn or a previously-simulated image
- in the first case, the objective function comprises a variogram model, indicator variograms, linear correlations with a secondary variable, or other criteria
- in the second case, a stochastic image created by another simulation algorithm is "post-processed" or finished by forcing the perturbed image to conform to the statistical features of a categorical "training image"
- see Deutsch and Journel (1997, p.158-167; p.183-188; p.318-324) for details and example (unconditional) simulations

Week 14 Lecture Material Scaling, Sample Support (Clark, ch.13; Isaaks & Srivastava, ch.19)

1. The problem:

- all geoscience measurements are made on a sample of the earth at a particular scale
 - eg: - petrofabric measurements on core samples, on outcrops, or on aerial photos
 - ore or contaminant concentrations in core samples, pit samples, truckloads
 - hydrograph float recorder ("instantaneous") vs. time-averaged data logger
- the size of the measurement volume (or interval) is known as the sample "support"
- measurements of a regionalized variable made at one scale (eg: core-scale measurements of ore grade or permeability) should not be used to obtain estimates at another scale (eg: truckload or stope-size ore grade), unless an appropriate correction is applied to account for the change of support (scale) effect
- conceptualization of 1-D point process variability vs. averaged process variability
- two examples from mining and from aquifer testing

2. Mining example: ore grade analyses obtained from diamond-cores are used to estimate spatial variability of ore grade in stope block-sized samples during mining

- what kind of problem / bias might be expected? how might the variograms of ore grade from the two different sampling schemes differ? (variance, λ)

3. Aquifer test example: K estimates derived from permeameter tests on cores and from single-well slug tests are pooled with multiple-well, long-duration pumping tests to predict spatial variability of K in the aquifer

- what sorts of problem / bias might be expected? how might the variograms of K from the three different testing / sampling schemes differ?

4. The approach: two commonly used alternatives

- numerical scaling (Gelhar example): stationarity, well-behaved scaling
 - population variance of the averaged process will always be smaller than that of the corresponding point process (measurements made at a small-scale)
 - correlation length of the averaged process will be larger
 - use to scale the point process-derived correlation parameters to be applicable for estimation at the averaged scale or vice versa
- block averaging of small-scale "point" simulations: always the safer approach
 - use the point process correlation parameters to simulate variability within larger blocks representing the averaged process sampling volume
 - compute block-averages from the simulated results and from multiple blocks and estimate the population variance; estimate correlation length separately

Week 15 Lecture Material Evaluation of Uncertainty, Risk, Probability of Failure/Success

1. The stochastic method is naturally suited to quantifying uncertainty
 - typically, stochastic estimation chosen *because* uncertainty must be evaluated
 - probabilistic estimates can be coupled with various deterministic transfer (or transformation) functions to predict probability of outcomes dependent on the stochastically-generated estimates (eg: ground water transport time, mining cost, risk of remediation failure)

2. Binary indicator kriging is a simple form of probability estimation
 - the estimation of spatial distribution of a categorical variable (indicator = 1, if present) produces a range of values between 0 and 1 which are equivalent to probability of occurrence of the categorical variable

3. Multiple stochastic simulations provide equivalent information
 - usually conceptually easier to "sell" or convey the concept of variability and uncertainty
 - required if simulation results must be transformed via a process model into predicted outcomes, which are the goal of the uncertainty analysis (eg: ground water flow)

4. Examine use of POSTSIM options for summarizing different kinds of information about uncertainty
 - Probability Maps - $P_{\text{exceedance}}$ (for a given threshold)
 - Quantile Maps - z value exceeded by $1 - P_{\text{quantile}}$ of realizations at a location
 - Maps of Spread -
 - Conditional Variance Maps
 - Inter-Quartile Range Maps

5. Decisions on additional sampling - most information return and reduction of uncertainty will be obtained from additional samples where Probability Map values are near 0.5

6. GIS Applications -
 - natural analysis environment for spatial data, spatial statistics
 - limited built-in capabilities currently available, but changing (e.g. IDRISI's statistical package)
 - dominance of vector-based over raster-based GIS software (spatial statistics easy to implement in raster format)
 - spatial statistics and spatial modeling mostly done outside the GIS, then imported and displayed in the GIS

Geostatistics on the Internet:

- Excellent summary overview of just about the whole field of geostatistics:
<http://curie.ei.jrc.it/faq/index.html>

- Excellent source for geostatistical analysis of hydrologic data:
<http://earth.agu.org/revgeophys/kitan01/kitan01.html>

- Home page of Clayton Deutsch with some additional and very useful GSLIB add-on programs:
<http://www.ualberta.ca/~smpe/people/dcvd.htm>