Geostatistics for Spatial Uncertainty Characterization

ZHANG Jingxiong  ZHANG Jinping  YAO Na

Abstract  Most geospatial phenomena can be interpreted probabilistically because we are ignorant of the biophysical processes and mechanisms that have jointly created and observed events. This philosophy is important because we are certain about the phenomenon under study at sampled locations, except for measurement errors, but, in between the sampled, we become uncertain about how the phenomenon behaves. Geostatistical uncertainty characterization is to generate random numbers in such a way that they simulate the outcomes of the random processes that created the existing sample data. This set of existing sample is viewed as a partially sampled realization of that random function model. The random function's spatial variability is described by a variogram or covariance model. The realized surfaces need to honour sample data at their locations, and reflect the spatial structure quantified by the variogram models. They should each reproduce the sample histogram representative of the whole sampling area. This paper will review the fundamentals in stochastic simulation by covering univariate and indicator techniques in the hope that their applications in geospatial information science will be wide-spread and fruitful.

Keywords  uncertainty; stochastic simulation; realization; normal score transform

CLC number  P208

Introduction

Geostatistics provides a coherent framework for spatial prediction and uncertainty quantification. Spatial prediction or interpolation concerns how to estimate the variable under study at an un-sampled location given sample observations at nearby locations. This process of spatial estimation aims at finding a single number at the un-sampled location. Spatial prediction has been the focus for the first two papers of the sequel, whereby univariate, multivariate, and indicator geostatistics is reviewed with key mathematical treatment prescribed. The following text draws much on Rossi et al. and Caers[1, 2].

Finding a single estimate is, however, unsatisfactory in many applications. We are often interested in determining the uncertainty about the unknown values at un-sampled locations. In statistics, uncertainty is quantified through a statistical distribution, describing the frequencies of outcomes expected. In a spatial domain, distribution functions will depend on the number of sample data, their spatial configuration, the data values, and the specific spatial phenomenon under study. With distributions simulated, probability intervals or other measures of uncertainty can be calculated.

In practice, we are seldom interested in determining an estimate or uncertainty model of a single un-sampled location. We are more interested in the joint determination of all the un-sampled values on a
specific regular or irregular grid. For example, a simple way to obtain a grid map of estimates is to perform estimation of un-sampled nodes, based on sample values. However, this single map should and cannot be used as a representation of the unknown spatial phenomenon. Since the true spatial distribution of all values in the grid is unknown, a single map can only provide a smooth representation of the truth. Moreover, to determine that smooth map, estimates at each grid node are determined independent of each other, i.e., the estimate at each node is determined independent of estimates at other nodes.

Instead of providing a single map, modern geostatistics aims at providing a set of alternative maps, termed realizations that quantify the uncertainty about the joint outcomes of the unknown grid node values. This branch of geostatistics concerns stochastic simulation or imaging. Stochastic simulation is a geostatistical technique that allows the reproduction of a given environmental pattern constrained to various types of data. Stochastic simulation generates multiple models, i.e., realizations, each reproducing the same patterns or creating alternative representations of the unknown truth. In the same sense as a set of outcomes drawn from a univariate distribution model represent the uncertainty about a random variable, a set of stochastically generated reservoir models \( \{z(x)^l, x \in A, l=1, 2, \ldots, N\} \) represent the uncertainty about the unknown true reservoir \( \{Z(x), x \in A\} \). These stochastic realizations are equi-probable, meaning that no single “best” realization exists.

In this paper, an overview of the most used geostatistical simulation methods will be provided. Geostatistics has evolved from an emphasis on estimation to stochastic imaging and data integration through Monte Carlo simulations. Recently, multiple-point geostatistics has emerged as a new powerful field for obtaining realistic geostatistical models that can integrate consistently a large variety of different sources of information acting on different scales\(^2\). This paper will describe stochastic simulation and stress its importance over estimation in applications.

1 Stochastic simulation

As discussed previously, kriged maps usually appear smooth. Thus, kriging estimates and variance do not lend themselves to quantification of spatial uncertainty in the mapped fields, since the kriging estimation at an un-sampled location is performed independently from other locations, thus leading to the smoothness in the estimated surfaces. Stochastic simulation may be pursued to correct the smoothing effect of kriging and generates multiple realized maps, each reproducing the same spatial structure as quantified by certain histograms and variograms, thus each representing alternative representations of the unknown truth. The set of equi-probable realizations provide the basis for computing the spatial uncertainty in the mapped surfaces. The underlying rational for stochastic simulation is that, since the properties at un-sampled locations are largely unknown, the single kriging estimates at individual locations should be replaced by sets of alternative models that honor any sample information and reproduce a pattern similar to that of the true reservoir.

Recall that kriging gives us an estimate of both mean and standard deviations of the variable at each grid node, meaning we can represent the variable at each grid node as a random variable following a normal (Gaussian) distribution. Rather than choosing the mean as the estimate at each node, the simulation chooses a random deviation from this normal distribution, according to a uniform random number representing the probability level.

For the example shown in Fig.1 (referring to http://people.ku.edu/~gbohling/cpe940), ordinary kriging provides a mean estimate of 22.9 with an estimate of variance 0.5. In this case, if we happened to generate a uniform random number of \( p=0.55 \) for this grid node,

![Fig.1 The rationale behind Gaussian simulated, where a simulated Z value of 22.99 is determined referring to a Gaussian model of \( N(22.9, 0.5) \)](image)
then the assigned $Z$ value would be 22.99, the corresponding value of the cumulative normal probability function (CDF).

We use a random path to avoid artifacts induced by walking through the grid in a regular fashion. We include previously simulated grid nodes as “data” in order to preserve the proper covariance structure between the simulated values. For the simulation, it is important that the data actually follow a Gaussian distribution. If they do not, we can use a normal score transform, as shown in Fig.2.

![Fig.2 Transformation of an empirical cumulative distribution function (CDF) (left) to normal CDF](image)

Journel and Huijbregts recalled that the conditional mean and variance of a stationary multivariate normal random function’s $(Y(x))$’s univariate conditional distribution are exactly equal to the simple kriging estimate and the corresponding simple kriging variance, respectively:  

$$E\{Y(x_i)\} = \sum_{i=1}^{N} \lambda_i \{Y(x_i) - m\} + (1 - \sum_{i=1}^{N} \lambda_i) E\{Y(x_i)\}$$ \hspace{1cm} (1)

$$Variance\{Y(x_i)\} = Variance\{Y(x_i)\} - \sum_{i=1}^{N} \lambda_i \text{cov}(x_i)$$ \hspace{1cm} (2)

In other words, the conditional distributions required to implement Gaussian sequential simulation can be acquired easily through simple kriging for a multivariate normal random function.

Gaussian sequential simulation capitalizes on the "nice" properties of the normal distribution, but ecological data, typically, are not normally distributed. Therefore, a normal score transform (Eq.3) of the sample data is required prior to the simulation and a back-transform (Eq.4) is required after analysis.

$$y(x) = \phi\{z(x)\}$$ \hspace{1cm} (3)

$$z(x) = \phi^{-1}\{y(x)\}$$ \hspace{1cm} (4)

This normal score transform $\phi$ needs to be monotonic or one-to-one and invertible so that the simulation results expressed as $y$ values can be restored in the original $z$ sample data units, as discussed in Reference[4].

One common problem with a normal transform is how to treat tied sample values. Tied sample values create “spikes” in a probability density function, and a normal score transform of tied values would no longer be one-to-one. One solution is to rank the ties according to the mean of the surrounding or neighborhood data values[4]. Another is to assign each tie the average rank score[5]. Because it preserves duplicate values, Isaaks’s transformation is not strictly a normal score transform; however, the back transformation reproduces the original histogram precisely because the correct proportion of tied values is preserved. This can be a distinct advantage when the simulation takes place in areas where tied values (especially large and small sample values) are clustered.

To ensure that the correct spatial structure is reproduced, one needs to define the joint probability model of properties at all grid locations taken together, not one-by-one as done in kriging. A joint distribution is defined as:

$$F(z_1, z_2, \ldots, z_N) = Pr(Z(x_1) \leq z_1, \ldots, Z(x_N) \leq z_N)$$ \hspace{1cm} (5)

where $N$ is the number of grid nodes. Since it is mathematically too difficult to define and draw from a joint distribution model, one relies on the fact that any joint distribution can be decomposed in a product of $N$ conditional distributions:

$$Pr(Z(x_1) \leq z_1, \ldots, Z(x_N) \leq z_N) = Pr(Z(x_N) \leq z_N | Z(x_1) \leq z_1, \ldots, Z(x_{N-1}) \leq z_{N-1})$$

$$\times \cdots \times Pr(Z(x_2) \leq z_2 | Z(x_1) \leq z_1) \times Pr(Z(x_1) \leq z_1)$$ \hspace{1cm} (6)

which implies that drawing from a joint distribution is equivalent to drawing from $N$ univariate conditional distributions. This lays down the basis for sequential simulation whereby each node is simulated based on the previously simulated nodes. The conditional sequential simulation refers to the technique whereby some data, denoted as $(n)$, are available.
Each conditional distribution in decomposition (6) is then based on previously simulated nodes and conditional data in the grid:

\[
\Pr(Z(x_i) \leq z_i, \ldots, Z(x_N) \leq z_N | (n)) = \Pr(Z(x_i) \leq z_i | (n+N-I)) \times \cdots 
\]

(7)

\[
\times \Pr(Z(x_N) \leq z_N | (n+1)) \times \Pr(Z(x_1) \leq z_1 | (n))
\]

The sequential simulation algorithms for simulating a single realization are outlined as follows:

1) Assign any conditioning data \((n_i)\) to the grid;
2) Define a random path visiting all nodes \(\{x_i\}\) in the grid;
3) Construct a conditional distribution \(F(x_i, z | (n+i-1)) = \Pr(Z(x_i) \leq z | (n+i-1))\);
4) Draw a simulated value \(z(x_i)\) from the conditional distribution \(F(x_i, z | (n+i-1))\);
5) Add the simulated value to data-set \((n+i-1)\);
6) Continue the next node along the random path until all nodes are exhausted.

To implement sequential Gaussian simulation, it is often assumed that the joint distribution is multivariate normal. Any conditional distribution under the multi-Gaussian model is also Gaussian, hence fully determined by the mean and variance. Therefore, the conditional Gaussian distribution at step 3 can be simplified as

\[
G(x_i, z | (n+i-1)) = G(z - z^* G(x_i) / \sigma(x_i))
\]

where \(z^* G(x_i)\) and \(\sigma(x_i)\) refer to the kriging mean and standard error at grid node \(x_i\) respectively. It is required to transform the sample data to standard normal scores, and to transform the entire simulation back to the original data histogram, as a multi-Gaussian assumption implies that the histogram of the data is standard Gaussian.

2 Indicator stochastic simulation

Although the sequential Gaussian simulation technique is mathematically sound and often used in practical applications, it has some severe shortcomings attributed to the multi-Gaussian distribution model. Geologists often criticize reservoir models constructed by sgsim in GSLIB\(^6\) as noisy or unrealistic, even though the models honor variogram information extracted from wells or outcrops. The multi-Gaussian model enforces a so-called maximum entropy property on the resulting simulations, meaning that, for a given variogram model, sgsim will construct reservoir models containing maximally disconnected extremes. A closer inspection of sgsim realizations shows that the extreme values are maximally separated; hence the texture appears disorganized. Often, nature is more organized than assumed by the multi-Gaussian model.

To correct this default of sgsim, one needs to model the connectivity of extremes from any sample or outcrop data and to enforce the connectivity to exist in the resulting realizations. In order to quantify the connectivity better, one uses so-called indicator random functions introduced in previous papers in the sequel, i.e., for various thresholds \(z_k (k=1, \ldots, K)\), one defines a variable

\[
I(x, z_k) = \begin{cases} 
1, & \text{if } Z(x) \leq z_k \\
0, & \text{else} 
\end{cases} 
\]

(8)

Each of \(K\) indicator variables allows us then to focus on various ranges of the variable. Taking \(z_k\) equal to the median allows quantifying the connectivity of average \(z\)-values, while taking \(z_k\) equal to some extreme quantile allows quantifying the connectivity of extremes. A spatial measure of connectivity between any two points in space separated by a lag-distance \(h\) is now the indicator variogram,

\[
\gamma_I(h, z_k) = \text{var}(I(x, z_k) - I(x+h, z_k))
\]

(9)

which can be modeled from the sample data by transforming each sample into a vector of size \(K\) of indicator data (zeros and ones).

Once the \(K\) indicator variograms are determined, one can perform the indicator simulation. The goal of the indicator simulation is not to simulate zeros and ones (indicators) but to determine local probability models that do not rely on any Gaussian assumption. Indeed, in the indicator methods one relies on the following simple rule that

\[
E[I(x, z_k)] = \Pr(I(x, z_k) = 1) = \Pr(Z(x) \leq z_k)
\]

(10)

Hence, any estimation of \(I(x, z_k)\) is also an estimation of \(\Pr(Z(x) \leq z_k)\), which means that kriging of an indicator variable is nothing more than determining the local uncertainty about the original \(Z\)-variable

\[
E[I(x, z_k) | (n)] = \Pr(Z(x) \leq z | (n))
\]

(11)
For example, to determine the probability of porosity at un-sampled location \( x \) to be above \( z_k \) given the local well data, one only needs to krig (estimate) the indicator variable by indicator kriging. By estimating \( \Pr(Z(x) \leq z_k \mid (n)) \) for various thresholds \( z_k \), the indicator kriging provides the uncertainty about the property at an un-sampled location \( x \) given any sample data \( (n) \). Note that no Gaussian or any parametric distribution model is assumed, the uncertainty is quantified through a series of cumulative probabilities at given thresholds \( z_k \).

The indicator coding of data enables the specification of the conditional probability distribution for any unsampled location. Say, \( \alpha = 1, \ldots, n \) samples or conditioning data are coded as indicators for \( k = 1, \ldots, K \) thresholds or cutoffs:

\[
i(x_\alpha; z_k) = \begin{cases} 1, & \text{if } z(x_\alpha) \leq z_k \\ 0, & \text{if } z(x_\alpha) > z_k, \end{cases}
\]

(12)

where \( i(x_\alpha; z_k) \) represents the indicator conditioning data at location \( x_\alpha \).

The conditional probability distribution at any unsampled location is found by computing the expected value of the indicators, \( i(x_\alpha; z_k) \), across all cutoffs, \( k = 1, \ldots, K \), using all data, \( \alpha = 1, \ldots, n \):

\[
\Pr[Z(x) \leq z_k \mid Z(x_\alpha) = z_{\alpha}, k = 1, \ldots, K; \alpha = 1, \ldots, n] = E[I(x; z_k) \mid i(x_\alpha; z_k) = i(x_\alpha; z_k), k = 1, \ldots, K; \alpha = 1, \ldots, n]
\]

(13)

In other words, the estimation of the conditional probability distribution amounts to estimating the conditional expectation of an indicator random variable. This is useful because the computation of expected values is simple and easy.

Notice that Eq.13 states that each indicator simulation requires \( K \cdot N \) conditioning data. This can be a formidable computational necessity. Alternatively, only the indicators of the same threshold are considered in practice, thus, Eq.13 reduces to the approximation:

\[
\Pr[Z(x) \leq z_k \mid Z(x_\alpha) = z_{\alpha}, \alpha = 1, \ldots, n] = E[I(x; z_k) \mid i(x_\alpha; z_k) = i(x_\alpha; z_k), \alpha = 1, \ldots, n]
\]

(14)

This new expression calls for only \( n \) conditioning data and is tantamount to ignoring any cross-correlation between the indicators\(^5\).

Sequential indicator simulation (SIS) is very similar to sequential Gaussian simulation, except that the indicator kriging is used to build up a discrete cumulative density function for the individual categories at each case and the node is assigned a category selected at random from this discrete CDF.

Very briefly, an indicator representation for a categorical variable would be formulated as:

\[
i(x_\alpha; k) = \begin{cases} 1, & \text{if class is present at } x_\alpha \\ 0, & \text{otherwise} \end{cases}
\]

(15)

where you would have one indicator variable for each of the \( K \) different classes. We can then use kriging (based on indicator semivariograms) to produce a set of class membership probabilities at each grid point, build up a CDF from the probabilities, and select a class at random from the CDF.

Fig.3  In the same way simple kriging naturally extends into sequential Gaussian simulation, indicator kriging extends into sequential indicator simulation (sisim). Sisim does not require any distribution assumption, as sgsim does

1) Assign the data \((n)\) to the grid.
2) Define a random path visiting all nodes \( x \).
3) Loop over all nodes \( x_i \).
   (1) Construct a conditional distribution \( F(x_i; z \mid (n + i - 1)) \) by estimating \( \Pr(Z(x) \leq z_k \mid (n + i - 1)) \) for various threshold using indicator kriging;
   (2). Draw a simulated value \( z(x_i) \) from the conditional distribution \( F(x_i; z \mid (n + i - 1)) \);
   (3) Add the simulated value to data-set \((n + i - 1)\);
4) End simulation.

For a continuous variable such as permeability, indicator variables are built by comparing data values to a set of thresholds, \( z_k \), as shown in Eq.12. We might define thresholds, for example, at the 10th, 25th, 50th, 75th, and 90th percentiles of the data distribution. In this case, kriging the indicator values for the \( k \)th
threshold, \( z_{k+1} \), gives the estimates of \( P(Z(x) \leq z_{k+1}) \) at each estimation point. Since this is a cumulative probability, we do not need to go through the process of summing to get a CDF, although we will need to correct any violations of the expected order relationships, \( P[Z(x) \leq z_k] \leq P[Z(x) \leq z_{k+1}] \), that happens to occur.

3 Conclusion

The information about a spatially varying phenomenon is usually incomplete. Most often, only a few samples of the variable under study are available, although there may be an abundance of indirect information gathered with remote sensors, which may suffer from unknown accuracy. This implies that one cannot determine with full confidence the exact unknown true outcome of that variable at every location.

Geostatistics allows quantifying the uncertainty about the unknown spatial phenomenon in terms of a set of alternative representations, i.e., realizations, of the unknown truth\(^\text{[7-9]}\). The aim of stochastic simulations is to enforce various properties on these realizations, so that they should have a similar pattern of variability as the unknown truth, and be constrained to any local sample data. Once these realizations are simulated, they can be post-processed using any type of transfer function (flow simulation, mining operation, environmental clean-up) to determine the local distributions of the variable and their associated risks\(^\text{[10-12]}\).

This paper has discussed univariate and indicator stochastic simulation. As hinted above, ground samples and remote sensing measurements are often available. It is thus interesting to consider integrating multi-source data via co-kriging techniques and to assess spatial uncertainty in multivariate analysis. Co-simulation can play an important role therein.

Geospatial patterns can be represented by a simple variogram, a 2-point statistical model, as discussed in this paper. There are applications where multi-point statistics must be enforced in stochastic simulation, as explained by Caers\(^\text{[2]}\). In such occasions, geospatial patterns are contained in a training image, which are deemed relevant for the reservoir under study. Stochastic simulation needs to be extended to be able to accommodate multiple-point statistics in the simulated surfaces. This will lead to another avenue of further investigation.

References

[1] Rossi R E, Borth P W, Tollefson J J (1993) Stochastic simulation for characterizing ecological spatial patterns and appraising risk[J]. Ecological Applications, 3(4): 719-735
[2] Caers J (2001) Geostatistical reservoir modelling using statistical pattern recognition[J]. Journal of Petroleum Science and Engineering, 29(3-4): 177-188
[3] Journel A G, Huijbregts C (1978) Mining geostatistics [M]. New York and London: Academic Press
[4] Verly G W (1984) Estimation of spatial point and block distributions: the multi-Gaussian model[D]. California: Stanford University
[5] Isaaks E H (1990) The application of Monte Carlo methods to the analysis of spatially correlated data[D]. California: Stanford University
[6] Deutsch C V, Journel A G (1998) GSLIB: geostatistical software library and user’s guide[M]. Oxford: Oxford University Press
[7] Chiles J-P, Delfiner P (1999). Geostatistics: modeling spatial uncertainty[M]. New York: John Wiley & Sons
[8] Goovaerts P (1997) Geostatistics for natural resources estimation[M]. Oxford: Oxford University Press
[9] Isaaks E H, Srivastava S (1989) An introduction to applied geostatistics[M]. Oxford: Oxford University Press
[10] Journel A G (1989) Fundamentals of geostatistics in five lessons[M]. Washington D C: American Geophysical Union
[11] Kyriakidis P C (2001) Geostatistical models of uncertainty for spatial data[M]/ Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS applications. New York: Springer-Verlag
[12] Rossi R E, Mulla D J, Journel A G, et al. (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence[J]. Ecological Monographs, 62(2): 277-314