Applied Geostatistics 2
Part 3: Simulations

Georg Kosakowski
Waste Management Laboratory, Paul Scherrer Institut
5200 Villigen PSI, Switzerland
email: georg.kosakowski@psi.ch

Problem: Kriging is smoothing the profiles! There is not guarantee that the kriged profile has the same variance and variogram as the original data.
Estimation
Get the best possible estimate at a unsampled location (minimize error variance via kriging - BLUE).
This does not mean that the kriged data-set has the same variogram and variance than the measured data.
Estimation strives for accuracy.

Simulation
Reproduce the statistical properties of the original data (e.g. variance, variogram).
Explore the range of possible realizations with or without honouring the measured data.
Simulation strives for realism.

There exist a lot of stochastic simulation techniques such as

- Sequential Gaussian Simulation
- Sequential Indicator Simulation
- Matrix decomposition
- Turning bands
- Simulated annealing
- ...see much more in e.g. Chilès and Delfiner (1999)

The generated realizations are used as input to complex transfer functions, such as flow and transport models for example.
To make the notation easier we write the equations in matrix notation and express the weights by the covariances:

$$\lambda C = C_x$$

$$Z^*_SK (x_0) = m + (Z - m) C^{-1} C_x$$

$$\sigma^2_{SK} = C(0) - C_x^T C^{-1} C_x$$

$C$: coefficient matrix for the covariances
$C_x$: vector on the right hand side
$\lambda$: column vector of the weights
$Z-m$ the row vector of the $Z(x)-m$

Sequential Gaussian Simulation

First recall simple kriging:

simple kriging estimator:

$$Z^*_SK (x_0) = m + \sum_{i=1}^{n} \omega_i (Z(x_i) - m)$$

simple kriging equations:

$$\sum_{j=1}^{n} \omega_j C(x_i - x_j) = C(x_i - x_0)$$

for $i = 1, ..., n$

simple kriging variance:

$$\sigma^2_{SK} = C(0) - \sum_{i=1}^{n} \omega_i C(x_i - x_0)$$

sequentiel simulation

Now we assume that $Z(x)$, $Z(x_1)$, $Z(x_2)$, $Z(x_3)$,... $Z(x_n)$ are multivariate gaussian.

If we try to estimate $Z(x)$ based on $Z(x_1)$, $Z(x_2)$,... $Z(x_n)$ by simple kriging we find that $Z(x)$ is univariate gaussian whose mean and variance is given by

$$Z^*_SK (x_0) = m + (Z - m) C^{-1} C_x$$

$$\sigma^2_{SK} = C(0) - C_x^T C^{-1} C_x$$

This gives a way to simulate a value for $Z(x)$, conditioned on the data $Z(x_1)$, $Z(x_2)$,... $Z(x_n)$. 
Simplified procedure:

- Choose (at random) a point where we would like to simulate a data value.
- Obtain a simple kriging estimate and the simple kriging variance based on the data.
- Sample from that univariate gaussian distribution and apply the value to the point.
- Add the point to the data set of already known points.
- Proceed to the next point.

In a more general sense it can be shown, that a N-point multivariate distribution can be decomposed in a set of N one-point conditional cdf's as:

\[
F(x_1, \ldots, x_N; z_1, \ldots, z_N|n) = \\
F(x_N; z_N|n+N-1) \\
F(x_{N-1}; z_{N-1}|n+N-2) \\
F(x_2; z_2|n+1) \\
F(x_1; z_1|n)
\]

\(F(x_1, z_1|n+N-1)\) is the conditional cdf of \(Z(x_1)\) given the set of \(n\) original data values and the previous \((N-1)\) realizations \(z^{(j)}(x_i), j=1,\ldots,N-1\).

General algorithm for gaussian sequential simulation

- Perform a transformation of the data into Gaussian space -> score transformation into a standard normal cdf
- define a random path visiting all nodes
- For each node \(x_0, i=1,\ldots,N\) do
  - model the conditional distribution \(F(x_i; z(n+i-1))\) of \(Z(x_i)\), given the \(n\) original data values and all \(i-1\) previously drawn values \(z^{(0)}(x_i), x_j, j=1,\ldots,i-1\)
  - draw the simulator value \(z^{(0)}(x_i), F(x_i; z(n+i-1))\)
  - end loop
  - Perform a backtransform to identify the target histogram (if needed).

Example for transformation into gaussian space:

Simple methods for estimating a point distribution

From Chiles & Definer (1999)
Another example for a transformation into gaussian space:

**Conditions:**
- The distributions are independent.
- The variances of the single distributions are larger than zero but finite.
- The third moments are bounded (so that no single component will dominate the distribution of the summation as \( n \to \infty \)).

Example for distributions which will not fulfill these conditions: power law distributions (variance and higher moments are not bounded)

**The central limit theorem**
Normal distributions are important in many fields of science and engineering because many random variables are empirically well approximated by normal distributions.

The reason for this arises from the **central limit theorem**:

The distribution of an average tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal.

Let’s \( Z \) be \( 1,2,\ldots,n \) independent distributions with mean \( m_i \) and variance \( \sigma_i^2 \) then \( Z = Z_1 + Z_2 + \ldots + Z_n \) approaches a normal distribution with

\[
\begin{align*}
\mu &= \sum_{i=1}^{n} m_i \\
\sigma^2 &= \sum_{i=1}^{n} \sigma_i^2
\end{align*}
\]

**Example:**
Convolution of several (independent) uniform distributions gives a normal distribution.

For the uniform distribution is:

\[
\begin{align*}
m_i &= \frac{T}{2} \\
\sigma_i^2 &= \frac{T^2}{12}
\end{align*}
\]

\( n = 2 \)

\[
\begin{align*}
m &= \frac{T}{2} \\
\sigma^2 &= \frac{T^2}{6}
\end{align*}
\]

\( n = 2 \)

\[
\begin{align*}
m_i &= \frac{3T}{2} \\
\sigma_i &= \frac{T^2}{4}
\end{align*}
\]

Papoulis (1984)

Other examples on the web (search other examples with e.g. google):

http://www.math.cas.cz/faculty/stanton/m262/central_limit_theorem/CLT.html
http://www.stat.sc.edu/~west/javahtml/CLT.html
the central limit theorem can be viewed as a property of convolutions of positive functions: 

\[ f(x) = f_1(x) * f_2(x) * \ldots * f_n(x) \]

Convolution of two functions \( f(x) \) and \( g(x) \) over a finite range \([0, t]\) is given by

\[ f \ast g = \int_{0}^{t} f(\tau) \cdot g(t - \tau) \, d\tau \]

http://mathworld.wolfram.com/Convolution.htm

Conditioning of transformed data does generally not preserve the histogram when the studied random function and its nonconditional simulation are not Gaussian functions.

1. It is advisable to check that the bivariate distributions of the transformed data \( Y(x) \) can be considered Gaussian:
   - The scatterplot of \( (Y(x), Y(x+h)) \) for fixed \( h \), ought to be elliptical.
   - The variograms of order 1 and 2 (the normal variogram) of \( Y(x) \) ought to satisfy
     \[ \gamma_1(h) \gamma_{sub1}(\infty) = \sqrt{\gamma(h) \gamma(\infty)} \]

The variograms of \( Z(x) \) and \( Y(x) \) ought to be in a relationship that depends on the “transformation” (the one we used to transform the non-gaussian \( Z(x) \) to the gaussian \( Y(x) \) function.

![Diagram](http://mathworld.wolfram.com/Convolution.htm)
meuse example

## ex06.cmd ##

Unconditional Gaussian simulation on a mask

(local neighbourhoods, simple kriging)

```r
x <- krig(log(zinc) ~ 1, ~ x + y, data = NULL, newdata = meuse.grid,
model = v, nmax = 20, beta = c(5, 9), nsim = 5, dummy = TRUE)
levelplot(z ~ x + y | name, map.to.lev(x, z=c(3:7)), aspect = mapasp(x),
main = "five unconditional realisations of a correlated Gaussian field")
```

## ex07.cmd ##

Gaussian simulation, conditional upon data

(local neighbourhoods, simple and ordinary kriging)

```r
x <- krig(log(zinc) ~ 1, ~ x + y, meuse, meuse.grid,
model = v, nmax = 20, beta = c(5, 9), nsim = 5)
levelplot(z ~ x + y | name, map.to.lev(x, z=c(3:7)), aspect = mapasp(x),
main = "five conditional realisations of a correlated Gaussian field")
```

unconditional vs. conditional simulations

sequential Indikator simulation

The steps of the SIS procedure are as follows:

1.) Discretize the given data set into K+1 classes using indicator transformation with K threshold values \( z_k \):

\[
h(k, z_i) = \begin{cases} 1 & \text{if } z_i \leq z_k \\ 0 & \text{otherwise} \end{cases} \quad k = 1, \ldots, K
\]

2.) Perform a structural analysis by computing experimental indicator variograms \( \gamma^*(l_h, z_k) \):

\[
\gamma^*(l_h, z_k) = \frac{1}{n(l_h)} \sum_{i=1}^{n(l_h)} \left[ h(z_i + l_h, z_k) - h(z_i, z_k) \right]^2
\]

\( n(l_h) \): Number of pairs of indicators, separated by a distance \( l_h \)

and fitting indicator variogram models for each threshold \( z_k \).
3.) Perform sequential simulation.

- Define a random path to visit each node of the simulation grid only once.
- At each simulation node \( y_0 \), determine the \( K \) cdf values:
  \[
  F(y_0; z_i[n]) = \Pr\{Z(Z_{i}) \leq z_i[n]\}
  \]
  using indicator kriging for each threshold \( z_i \):
  \[
  \sum_{i=1}^{K} \nu_i(y_0; z_i) = C_{z}(z_i) + \lambda(y_0; z_i) = C_{e}(z_i)
  \]
  \[
  \nu_i(y_0; z_i) = 1
  \]
  The conditioning information consists both of indicator transforms of the originally measured data and the simulated indicator transforms from the nodes visited previously.
  Employ kriging neighborhood if necessary.

4.) Assign parameter values to the simulated class values.

At each grid node, the SIS technique first simulates a \textit{class} value. The classes are defined by the thresholds \( z_i \) (see above). A simulated \textit{parameter} value is finally obtained by e.g. drawing a parameter value, using uniformly distributed random numbers, from the class of the original data cdf corresponding to the simulated class value.
Sequential indicator simulation

meuse example: indicator cosimulation