Stochastic – advantages and uncertainties for subsurface geological mapping and volumetric or probability calculation

Stohastika – prednosti in negotovosti v podpovršinskem geološkem kartiranju in računanju prostornin ali verjetnosti

Tomislav Malvić

1 University of Zagreb, Faculty of Mining, Geology and Petroleum Engineering, Pierottijeva 6, 10000 Zagreb, Croatia
* tomislav.malvic@rgn.hr

Abstract
Stochastic, especially simulation, occasionally could be found in different geological calculations, mostly as the most advanced mapping method. Its main attribute is description of uncertainties that are inherent not only to any geological mapping dataset but also to any volumetric or probability calculation. Here are presented uncertainties in all three cases – mapping, volume calculation and probability calculation – and reasons why and when to use stochastic in them. The stochastic, and consequently simulation, is a recommended tool in case of a low number of data (<15 inputs) or large dataset (>40 inputs), but in both cases, the descriptive statistics needs to be known and is reliable. Almost the same could be applied in volumetric calculation, but the success of stochastic in probability calculation depends on large datasets, with 15 or more inputs.

Key words: simulations, number of input data, stochastic mapping, volume calculation, probability, Croatia

Izvleček
Stohastiko, zlasti simulacije, lahko občasno najdemo pri različnih računih, zlasti pri najdobrejših metodah kartiranja. Njihova poglavljena značilnost je opisovanje negotovosti, ki so lastne sleherni zbirki podatkov geološkega kartiranja, pa tudi slehernemu volumetrijskemu ali verjetnostnemu računanju. Tu predstavljamo negotovosti v vseh teh treh primerih – pri kartiranju, računu prostornin in računu verjetnosti, ter načine, čemu in kdaj uporabljati pri tem stohastiko. Stohastični pristop in posledično simulacijo je priporočljivo uporabljati v primeru majhega števila podatkov (manj od 15) ali velike datoteke (nad 40 inputov), vendar mora biti v obeh primerih opisna statistika znana in zanesljiva. Skoraj isto velja tudi za volumetrijo, medtem ko lahko uporabljamo stohastiko v računu verjetnosti ugodnega izida samo v primeru večjih datotek s 15 inputi ali več.

Ključne besede: simulacije, število podatkov, stohastika, kartiranje, računanje prostorne, verjetnost, Hrvaška
Introduction

Stochastic simulation or Gaussian simulation (sequential or indicator) is a special geostatistical method based on different algorithms compared to deterministic interpolation methods such as Kriging and Cokriging [1–3]. Differences are a result of extensions introduced in the Kriging algorithm that can have advantages or disadvantages, due to introduction of uncertainties in estimations. Consequently, the selection between Kriging- and Gaussian simulation-based algorithms is very important and asks for experienced professionals [4].

The most common property of simulation is calculation of numerous realisations (values) for each cell in grid (excluded are hard data in conditional ones). The requirement is input dataset characterised using normal distribution. The total set of realisations is characterised with uncertainties, derived from the size of dataset, variogram model and measurement errors. As input dataset, such an error is also characterised using normal distribution.

The simulation obviously calculates an enormous number of new grid values ($10^3$ times larger than the input dataset). Sometimes, it is used for artificially increasing of dataset, combining simulated and input values. As a consequence, descriptive statistics and histogram for analysed data can be easily and clearly calculated. In the grid of $50 \times 50$ cells, in 100 realisations, totally 250000 values are calculated. Hence, the input dataset of usually 10–20 hard data is enlarged in the scale of $10^4$. Moreover, numerous realisations give as outcomes of numerous maps. All of them are equally probable, and some of them can be selected as “representative”, but always at least three. Such a selection is done based on the order of calculation, random sampling, calculation of total map cells’ values, etc., but selection always needs to be unbiased. On contrary, if intention is given only to single map as an outcome, then Kriging or Cokriging methods are chosen as algorithms made just for such a purpose.

Simulations could be conditional and (rarely) unconditional and also Gaussian and (rarely) indicator. The sequential Gaussian simulation (SGS) could be applied to almost all geological variables characterised using Gaussian distribution, naturally or after transformation. Examples of such variables are porosity, depth, thickness or permeability.

There are several subsurface structures in the Croatian part of the Pannonian Basin System analysed by stochastic methods. The Kloštar structure (and hydrocarbon field) is the most Croatian geological structure analysed by stochastic geostatistical algorithms till now. An example of how to apply SGS in subsurface mapping of porosity, depth and thickness can be given as a set of maps taken from Zelenika and Malvić [5]. The data were collected from a reservoir of Lower Pontian age in the Kloštar field, located in the western part of the Sava Depression. Simulation used present-day depth, thickness, and locations of areas with higher porosity to successfully reconstruct [5] paleo-depositional environments and the distribution of different lithotypes of turbidites (Figure 1).

However, sequential indicator simulation (SIS) is more specific and used only when mapping is based on cut-off (threshold) values defined for the selected variable. Such an application is most often used in lithofacies mapping, when cut-off selected for porosity or thickness, directly or indirectly, indicates lateral changes in lithofacies. Such lateral changes are one of the main properties of turbidite arenite lithofacies in Neogene, Croatian part of the Pannonian Basin System, developed due to paleotopography, current directions and different lateral densities of turbidite. As results in all Croatian hydrocarbon fields reservoir sandstones laterally gradually are changed into marly sandstones, sandy marls, clayey marls and, eventually, marls.

One of indicator applications in the Croatian Pannonian Basin System (CPBS) Neogene lithostratigraphic units had been published for Lower Pontian, the Kloštar Ivanić Formation in the Kloštar Field (Figures 2 and 3). Porosity and thickness had been mapped using the mapping indicators and interpreted regarding probability to reach at least the selected porosity value. Resulting maps indicated depositional channel of sandstone, transitional lithofacies to marls, transport direction of turbidites and role of regional fault [6,7].
Basics of stochastic simulation algorithm

The basic condition for performing SGS is (approximately) normal or Gaussian distribution of input data. Transformation into normality, if possible, is also allowed. Normal distribution is characterised using statistical properties such as expectation and standard deviation ($\mu$ and $\sigma^2$, respectively), which are basic conditions for calculation of uncertainty range in cells and for estimation of errors.

Figure 1: The first (left) and 100th (right) SGS realisation for porosity. The selection is based on the simple order of calculation. Taken from [5].

Figure 2: Direction of material transport during Lower Pontian, Kloštar Structure. Probability maps for porosity >19% (left) and >20% (right). Red indicates fault, and black dot indicates well. Taken from [7].

Figure 3: Direction of material transport during Lower Pontian, Kloštar Structure. Probability maps for thickness >13 m. Red indicates fault, and black dot indicates well. Taken from [5,7].
**Properties of input data, type of simulation and zero realisation**

All data represented with point values are, so-called, "hard data", which means that they are always constant in space whatever outcome is presented. Consequently, such constants value-in-place in simulation define it as "conditional simulation", where are hard-data un-changeable across the plane or space. It means that inputs will not change whenever simulation algorithm is applied. In addition, there is another type of SGS called "unconditional simulation". "Unconditional" means that input data are not treated as constants. It is a valid view because each grid is defined with numerous cells. Almost always each cell includes one or zero hard data. However, hard data are considered also as point data, i.e. infinity small point, compared with cells that in the 2D area are always much larger than single point. As a consequence, the so-called "hard-data" if moved through the cell area could easily change value. Consequently, it is why they cannot be considered as constants. Eventually, it is allowed to simulate new value into cells with "hard data". Whatever type is chosen, any simulation is based on the variogram model and Kriging interpolation.

The single Kriging map represents deterministic solution for input dataset and is called "zero (or Kriging) realisation". Such a realisation has known mean value (expectation), standard deviation ($\sigma$), Kriging variance ($\sigma_n^2$) and interval margins of simulated values ($\pm 3\sigma$ around $\mu$, i.e. probability of 99% to include all possibilities). Using this realisation, it is possible to perform required number of subsequent realisations for each simulated cell in the grid. As basic Kriging algorithm is very often used, Ordinary Kriging technique (Equation 1, taken from [3]), which is used in the Croatian geological mapping, is the most commonly used Kriging technique until now (e.g. [8]).

\[
\begin{bmatrix}
\gamma(Z_1 - Z_1) & \gamma(Z_1 - Z_2) & \ldots & \gamma(Z_1 - Z_n) & 1 \\
\gamma(Z_2 - Z_1) & \gamma(Z_2 - Z_2) & \ldots & \gamma(Z_2 - Z_n) & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\gamma(Z_n - Z_1) & \gamma(Z_n - Z_2) & \ldots & \gamma(Z_n - Z_n) & 1 \\
1 & 1 & \ldots & 1 & 0
\end{bmatrix}
\]

\[
\begin{bmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_n \\
\mu
\end{bmatrix}
= 
\begin{bmatrix}
\gamma(x_1 - x) \\
\gamma(x_2 - x) \\
\vdots \\
\gamma(x_n - x) \\
1
\end{bmatrix}
\]

**Simulation**

Normalised data, sometimes after transformation, with known N($\mu$, $\sigma$) are used to simulate values into grid cells. Simulation is based on, so called, two methods of introducing "randomness", which makes the entire process unbiased and repeatable (i.e. sequential). The first randomness method is selection of estimated (simulated) cells inside the grid. When cells are selected, the value is calculated using Kriging (or Cokriging) on hard data. Previously estimated cells in each consequent selection are considered as new "hard data", using the same variogram model as in the first simulation, but repeat the “zero realisation". Hence, each estimated cell is characterised with a new value and also with an uncertainty interval, wide $\pm 3\sigma$ around expectation in that cell. Eventually, entire realisation has its own variance calculation from "zero realization". It is also known as Kriging variance.

The second randomness method, i.e. introduction of stochastic into simulation for the second time, includes cell value estimation and is dependent on interval $\pm 3\sigma$, different for each cell. Random selection of any value from this interval represents the final cell value in this realisation. Each cell is also characterised using its own probability distribution function (PDF). This is why it is possible to calculate almost infinite number of equally probable realisations in one simulation.

**Calculation of numerous realisations**

Obviously, it is possible to calculate numerous realisations, which all have the same cell val-
ues only in hard data locations. Of course, it is valid only if conditional simulation has been performed, which is the most common case in geological mapping. However, the main task is always selection of the most appropriate realisations, regarding researching goal, graphical looks and, sometimes, cross-validation. The main purpose of simulation is presentation of numerous possible solutions, but obviously dozens or hundreds of maps are impossible to present as a single outcome. Hence, there are several quick steps to check how reliable simulation results are and which number of realisation could be appropriate. It is summarised in Table 1.

There is a general opinion that 100 realisations in simulation represents a large enough number of solutions that stochastic is representative for quality input dataset. Such a large set of realisations need to be post-processed to reach several realisations that are the most common characteristic for analytical purpose and simulated space. Such selection techniques are called ranking. The ranked variable is usually cumulative summation of all cells in the observed realisation. For example, if porosity had been mapped in some layer or unit, all cell values in one realisation can be summed and give total porosity for the entire realisation. If there are 100 realisations, such summation can be repeated 100 times, one for each realisation. Eventually, an entire set of 100 realisations can be ranked from the lowest total porosity per realisation (P0), through the median realisation (P50), to the highest ranked realisation (P100). "P" value defines how many realisations have lesser total score than the observed one, i.e. for the P0, there is 0% of lesser solutions, and for the P100, there is 100% realisations with a low total value.

However, such a method of ranking is not always possible to perform, because summation of cells across realisation would not result in a meaningful value. In such cases, especially when only a small number of realisations are performed (e.g. Table 1, the right-end column), some pure statistical techniques could be used. Selection of only the first and last realisations in sequence or each nth realisation or any number of realisations by the "random seed number" algorithm is allowed. Such a selection is usually applied when simulation includes only a small number of realisations (5 or 10), made from small datasets with the purpose of getting quick visual insight into areas with the largest uncertainties.

To summarise, each new realisation needs to fulfil two conditions: (a) order or simulated cells is defined completely randomly, i.e. "random seed number generator" is applied and (b) consequently, number of "hard data" values inside variogram ellipsoid does not need be equal in the same cell and different realisations.

### Table 1: Qualitative estimation of simulation reliability and recommended number of realisations.

<table>
<thead>
<tr>
<th></th>
<th>Completely fulfilled</th>
<th>Partially fulfilled</th>
<th>Rarely (irregularly) or not at all fulfilled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Could variable be characterised with normal distribution?</td>
<td>Mostly (like porosity)</td>
<td>Sometimes (after transformation, like permeability)</td>
<td>Not sure; test is needed (variables like depth or thickness)</td>
</tr>
<tr>
<td>Could variable be ranked by map values?</td>
<td>Always (any realisation can be ranked according to their cumulative value)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Which number of realisations represent real spatial uncertainties' characteristic to data?</td>
<td>100 or more realisations could be calculated</td>
<td>10–99 realisations could be calculated</td>
<td>Less than 10 realisations could be calculated (only quick quick insight)</td>
</tr>
</tbody>
</table>
Stochastic derived from deterministic models

Stochastic and deterministic models are closely connected. In fact, any deterministic model, map or equitation, is an accepted approximation of input dataset, characterised with “artificial certainty”. Geological deterministic models could be different, but mathematically, they are often purely numerical, such as calculation of probability of success (POS; e.g. [9–11]), volumetric calculation of structure (e.g. [12]) and graphical outcome like different deterministic maps (e.g. [13]). The last one is described in the previous chapter, and in the following are given models of probability and volumetric calculations.

Stochastic in geological risk or probability of success calculation

Deterministic models are just approximation of stochastic systems, where the variable properties are geologically and statistically known. However, increase in data always partially changes deterministic solution, like maps or statistics. It was proven, e.g. in the calculation of POS that a well-known deterministic model can be used for hydrocarbon discovery in plays and prospects in the Croatian offshore [11] or the CPBS [9]. Even such numerically simple multiplication could be partially described and modified with inclusion of stochastic (e.g. [14,15]). This is why because calculation of such values can be a significant uncertainty process, even if it is applied in geological areas covered with published probability tables for geological categories such as existence of trap, reservoir, source rocks, effective migration of fluids and preservation of hydrocarbons in reservoir. Consequently, Equation 1 is used:

\[
\text{POS} = p(t) \times p(r) \times p(s) \times p(m) \times p(p). \quad (1)
\]

Here, POS is the geological probability, \( p(t) \) the probability for trap existence, \( p(r) \) the reservoir existence, \( p(s) \) the source rock existence, \( p(m) \) the probability for effective migration and \( p(p) \) the probability that hydrocarbons are preserved.

Although raised in hydrocarbon geology, this methodology is easily modified in other geological disciplines, like storage of \( \text{CO}_2 \) in the subsurface [16]. However, it was proven [15] that at least three (sub)categories could be represented by stochastic simulation. Porosity subcategory maps make possible to directly calculate values such as minimum, median and maximum realisations. Hydrocarbon shows and quality of cap rocks are descriptive variables, but often they can be expressed as percentage. Hence, descriptive variables can be transferred into indicators or real number. For example, detection of new gas in the layer can be observed if concentration is >10% (indicator 0 or 1). The measured value of 15% can indicate that seal rocks are not completely impermeable, i.e. probability of sealing is 0.75. Such a probability can be lowered for any critical value of new gas (like 40% is 0.50, 60% is 0.25, 80% is 0.05), which is determined experimentally in-site or in laboratory. This showed that several (sub) categories in POS calculation can be described with several possible values, which is stochastic definition of an event.

Stochastic in numerical integration of structures

Stochastic is a property of numerical calculation of volumes of geological structures. This is the most often used method for calculation of (sub) surface structures. It is based on 2D approximation of 3D space. An object with volume \( V \) whose boundaries extend from \( x = a \) to \( x = b \) is defined by the definite integral (Equation 2) as follows:

\[
V = \int_a^b A(x) \, dx. \quad (2)
\]

The area \( A(x) \) of the section made by a plane parallel to the YZ plane has to be known at every point. In practice, the integrand \( A \) is defined by the table of values of the definite integral. It can be obtained by several formulas, including the areas (isopachs) measured by the mechanical device called planimeter. Two such formulas are the most often used in geology – trapezoidal and Simpson’s rule (e.g. [17–19,12]).
Equation 4) follows:

\[ \frac{h}{2} (a_0 + 2a_1 + 2a_2 + 2a_3 + a_4) \]  

Simpson’s rule is based on integral \( \int_a^b f(x)dx \) using approximation of \( y = f(x) \) by a parabola (Figure 5a), i.e. polynomial of second degree that passes through the points \( (a, f(a)), (b,f(b)), (c,f(c)) \), where \( c = \frac{1}{2}(a + b) \). The final equation (Figure 5b) is the following:

\[
\int_a^b f(x)dx \approx \frac{h}{3} \left[ f(a) + f(b) + 2 \sum_{i=1}^{n-1} f(x_{2i}) + 4 \sum_{i=1}^{n-1} f(x_{2i-1}) \right]
\]  

In practice, again for 5 sections (or 4 subintervals), such a formula becomes as given in the following:

\[ V_s = \frac{h}{3} (a_0 + 4a_1 + 2a_2 + 4a_3 + a_4) \]  

Application of these rules is not unique, but some strong recommendations are raised from experience. The Simpson’s rule starts approximation with a higher polynomial (2nd order vs. 1st order in trapezium), and the approximation, with numerous sections, is always better. However, there is not strong definition of “nu-
merous sections”. For 5 or less sections, almost always simpler trapezoidal rule will lead to better results. For more sections, the Simpson’s rule has obvious advantages. Moreover, the Simpson’s rule has two versions, resulting from practice. One version, for paired number of sub-intervals, has been proven mathematically and is given here. However also, there is version for even number of subintervals used in Croatian reservoir geology practice for decades ([20]). If in Equation 7 one more subinterval is added, then there will be even number of subintervals and the equation will be as follows:

\[ V_S = \frac{h}{3} (a_0 + 2a_1 + 4a_2 + 2a_3 + 4a_4 + 2a_5) \]  

(7)

**Discussion, recommendations and conclusions**

Stochastic and deterministic models are closely entangled. In fact, any deterministic model, map or equitation, is only an accepted approximation of natural input dataset or “artificial certainty”. Geological deterministic models could be different, but mathematically, they could be divided into (a) purely numerical, like calculation of probability of success or volumetric calculation of geological structure or (b) graphical outcome like different deterministic maps.

Regarding mapping, SIS is probably the most advanced simulation technique that uses original and indicator data for variogram calculation and mapping. Moreover, if Gaussian and indicator simulations are compared, indicator maps sometimes represent more uniform distribution, i.e. differences among realisations are not so large as in Gaussian ones (e.g. [5]). It is a result of variance of indicator variables, and consequently, indicator simulation gives more uniform distribution of cell values. Generally, if indicators are used, the larger number of cut-offs results in larger reduction of in-class “noise” [21]. Eventually, the main purpose of Gaussian simulation is mapping of real values, but the main intention of indicator simulations is probability mapping, i.e. mapping assuming that some cells will have values larger than cut-off. In both cases, it would partially remove the so called “bull’s-eye” effect, a very strong feature sometimes observed in deterministic maps. Removal could be even stronger if indicators are used.

Stochastic is an inherent property that is also used for other calculations in numerical geology, such as probability of success and volumetric calculation, which are previously described. Introduction of stochastic in such calculations gives some degree of freedom in selection of categories or fine-tuning of geological models. Consequently, if stochastic is applied, the algorithms for POS calculation or volumetric calculation need to be theoretically well known. Any decision to introduce stochastic (or not)
is based on the type of analyses and number of data (Figure 6).

As general recommendation, when to apply stochastic could be outlined (Figure 6) here:

1. The strongest criterion is number of input data. Although each dataset includes uncertainties, they are the highest in smaller datasets. On contrary, in large datasets, such uncertainties could be easily and precisely calculated.

2. Consequently, it means that in “medium-sized” datasets, stochastic could be described, but it does not play an important role in the analytical procedure. Such “medium-sized” datasets are still too small that uncertainties cannot be precisely numerically calculated (almost as constant) and too large that the representative statistics cannot be calculated.

3. This is why stochastic is recommended for “small” datasets with <15 inputs or for “large” ones with >40 points.

4. The calculation of probability of success for any geological category deviates from such recommendations, because it is purely a numerical method, where for <15 points, the porosity cannot be stochastically mapped, as well as other subcategories cannot be reliably estimated with several solutions.

Figure 6: Decision tree for introducing stochastic in geological mapping and numerical calculations, based on the type of method and number of input data.
General recommendation for any kind of input dataset, regarding each of three analysed approaches, is clearly summarised in Figure 6, which represents “all-purpose” table that could be applied in all research that include stochastic in geology.

Acknowledgement

This study is financially supported by “Mathematical Methods in Geology II” given by the University of Zagreb, Faculty of Mining, Geology and Petroleum Engineering in 2017 (no. M048xSP17).

References


