Les Cahiers du GERAD ISSN: 0711–2440

A high-order, data-driven framework for joint simulation of categorical variables

I. Minniakhmetov,
R. Dimitrakopoulos

G–2017–88

November 2017

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2017
– Bibliothèque et Archives Canada, 2017

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2017
– Library and Archives Canada, 2017

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

A high-order, data-driven framework for joint simulation of categorical variables

Ilnur Minniakhmetov [a]
Roussos Dimitrakopoulos [b]

[a] *COSMO – Stochastic Mine Planning Laboratory, Montréal (Québec) Canada, H3A 0E8*

[b] *GERAD & Department of Mining and Materials Engineering, McGill University, FDA Building, Montréal (Québec) Canada, H3A 0E8*

ilnur.minniakhmetov@mail.mcgill.ca
roussos.dimitrakopoulos@mcgill.ca

November 2017
Les Cahiers du GERAD
G–2017–88

Abstract: Relatively recent techniques for categorical simulations are based on multi-point statistical approaches where a training image is used to derive complex spatial relationships using patterns. However, simulated geological realizations are driven by the training image utilized, while the spatial statistics of the hard data is ignored. This paper presents a data-driven high-order simulation approach based upon the approximation of high-order spatial indicator moments. The high-order spatial statistics are expressed as functions of spatial distances similar to variogram models for two-point methods. It is shown that the higher-order statistics are connected with lower-orders via boundary conditions. Using an advanced recursive B spline approximation algorithm, the high-order statistics are reconstructed from hard data. Finally, conditional distribution is calculated using Bayes rule and random values are simulated sequentially for all unsampled grid nodes. The main advantages of the proposed technique are its ability to simulate without a training image, which reproduces the high-order statistics of hard data, and to adopt the complexity of the model to the information available in the hard data. The approach is tested with a synthetic dataset and compared to a conventional second-order method, *sisim*, in terms of cross-correlations and high-order spatial statistics.

1 Introduction

Stochastic, or geostatistical simulations, are often required in reservoir modeling, and the quantification of geological uncertainty, pollutants in contaminated areas, and other spatially dependent geologic and environmental phenomena. During the past decades, Gaussian simulation techniques have been used for geostatistical simulations (Matheron 1971; David 1977, 1988; Journel and Huijbregts 1978; Cressie 1993; Kitanidis 1997; Goovaerts 1998; Caers 2005; Webster and Oliver 2007; Remy et al. 2009; Chiles and Delfiner 2012). The choice of the Gaussian distribution is driven by several factors. First of all, the Gaussian variables can be fully described by a small amount of parameters, such as the first-order statistics (i.e. average values), and the second-order statistics (i.e. covariance or variogram). Secondly, the small number of parameters allows one to run simulations on grids with many million nodes.

Natural phenomena are known to exhibit non-Gaussian distributions and have complex non-linear spatial patterns (Guardiano and Srivastava 1993; Tjelmeland 1998; Dimitrakopoulos et al. 2010), which cannot be adequately described by second-order statistics. To overcome these limitations, multiple point spatial simulation (MPS) methods were introduced in the 1990's (Guardiano and Srivastava, 1993; Journel 1993, Strebelle, 2002; Journel 2003; Zhang et al., 2006; Chuginova and Hu, 2008; Straubhaar et al. 2010; Toftaker and Tjelmeland 2013; Strebelle and Cavelius 2014; others). The additional information is taken into account via training images (TI), which are not conditioned on the available data, but contain additional information about complex spatial relations of the attributes to be simulated. To retrieve this information from the training image, the similarity between the local neighborhood of an unsampled location and the training image is calculated in explicit or implicit form. Based on this similarity measure, the value of a node from the training image with the most similar neighborhood is assigned to the unsampled location being simulated. Generally, most of the multi-point simulation techniques are a Monte-Carlo sampling of values from the TI in some form or another. No spatial models are used and, importantly, no spatial information from the hard data is retrieved. As a result, simulations of attributes reflect the TI. In cases where there are relatively large datasets, conflict between the hard data and TI's statistics is clearly observed and the resulting simulations do not reproduce the spatial statistics of the hard data (Dimitrakopoulos et al. 2010; Pyrcz and Deutsch 2014).

Several attempts have been made to incorporate more information from the hard data. Some authors suggest using replicates from the hard data in addition to TI (Mariethoz and Renard 2010), however, in practice, it is hard to find any replicates for three-point relations when data is sparse. Others (Mariethoz and Kelly 2011) apply affine transformations to better condition the hard data; however, TI is still used as the main source of information. Another approach is to construct TI based on the hard data (Yahya 2011), but the resulting simulations may be biased from the method chosen for the TI construction.

Mustapha and Dimitrakopoulos (2010a, 2010b) proposed to use the high-order spatial cumulants as the extension of variogram models to capture complex multi-point relations during the simulation of non-Gaussian random fields. The technique estimates the third- and the fourth-order spatial statistics from hard data and complements them with higher-order statistics from TI. However, this technique is based on the approximation of conditional distribution using Legendre polynomials, which are smooth functions and are not capable of an adequate approximation of the discrete distribution of categorical variables.

The problem of describing complex multi-point relations of categorical variables was addressed in Vargas (2006, 2010). The author uses high-order indicator statistics to characterize spatially distributed rock bodies. In this paper, this idea is developed by introducing the connection between different orders into the mathematical model. For example, consider a third-order spatial indicator moment of a stationary random field, which is a function of two-lags. When one of the lags is equal to zero, the third-order indicator moment becomes the second-order indicator moment. Besides that, instead of exponential functions the B-spline functions are used to estimate high-order spatial indicator moments. It is known (Evans et al. 2009; Babenko 1986), that B-spines provide optimal (in term of accuracy) estimation of equicontinuous functions defined on compacts. Finally, a new recursive algorithm is proposed for better approximation of high-order spatial statistics with nested boundary conditions of lower-level relations. Then, as shown in sub-sequent sections, the conditional distribution for the given neighborhood is calculated from high-order indicator moments and the category is simulated.

The proposed method works without any TI, however, additional information from TI can be incorporated as the secondary condition during the approximation step, for example the conditions on derivatives or the order of continuity. In this case, the high-order spatial indicator moments are fully driven by hard data.

The paper is organized as follows. First, high-order spatial indicator moments are introduced as a function of distances between points for two-point and multi-point cases. Then, a mathematical model for recursive approximation of high-order spatial indicator moments is proposed. Finally, the simulation algorithm using the proposed model is developed and tested on fully-known datasets. Discussion and conclusions follow.

2 High-order spatial indicator moments

Material extracted in a mining complex undergoes various transformations before the concentrated or refined product is shipped. Understanding the effect of these transformations, as well as the associated cash flows, can help develop better planning strategies and provide a more realistic assessment of their impact. Therefore, this section focuses on developing realistic models of material flow in a mining complex, starting by discussing general concepts and then describing a particular implementation for a multi-pit copper mining complex.

Let (Ω, F, P) be a probability space. Consider a stationary ergodic random vector $\mathbf{Z} = (Z_1, Z_2, \dots Z_N)^T$, $\mathbf{Z} : \Omega \to S^N$, defined on a regular grid $D = \{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_N\}$,$\mathbf{x} \in R^n, n = 2, 3$, where Ω is a space of all possible outcomes, F contains all combinations of Ω, S^N is a set of states represented by categories $S = \{s_1, s_2, \dots s_K\}$, and P is the probability measure, or probability. For example, probability of Z_1 being at a state s_k is defined as:

$$P(Z_1 = s_k) \equiv P(\{\omega \in \Omega : \mathbf{Z}(\omega) \in s_k \otimes S^{N-1}\}) \tag{1}$$

Without loss of generality, assume that $s_k = k, k = 1 \dots K$. It can be shown, that the probability is equivalent to spatial indicator moment:

$$P(Z_{i_0} = k_0, Z_{i_1} = k_1, \dots) = E(I_{k_0}(Z_{i_0}), I_{k_1}(Z_{i_1}), \dots) \qquad \forall i_0, i_1 \dots = 1 \dots N, \forall k_0, k_1, \dots = 1 \dots K, \tag{2}$$

where E is the expected value operator and $I_k(Z_i)$ is an indicator function

$$I_k(Z_i) = \begin{cases} 1, & Z_i = k \\ 0, & Z_i \neq k_0 \end{cases} . \tag{3}$$

From now on, indicator moments are denoted as:

$$M_{k_0 k_1, \dots}(Z_{i_0}, Z_{i_1}, \dots) = E(I_{k_0}(Z_{i_0}), I_{k_1}(Z_{i_1}), \dots). \tag{4}$$

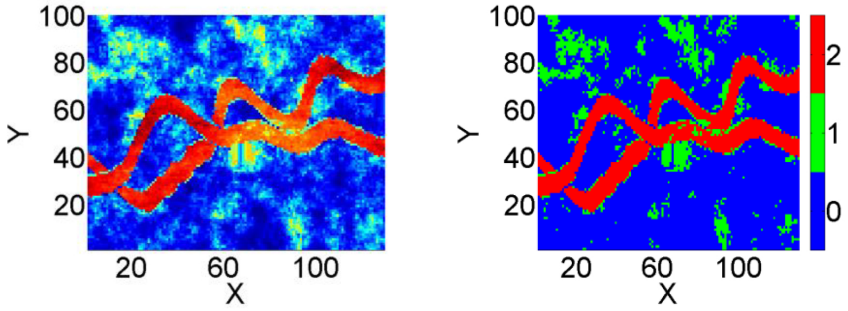
2.1 Second-order spatial indicator moment

Consider two random variables Z_{i_0} and Z_{i_1} separated by the lag $\mathbf{h}_1 = \mathbf{x}_{i_1} - \mathbf{x}_{i_0}$. Due to the stationarity assumption, their second-order spatial indicator moment for categories k_0, k_1 can be expressed as a function of the lag $\mathbf{h}_1$:

$$M_{k_0 k_1}(Z_{i_0}, Z_{i_1}) = M_{k_0 k_1}(\mathbf{h}_1). \tag{5}$$

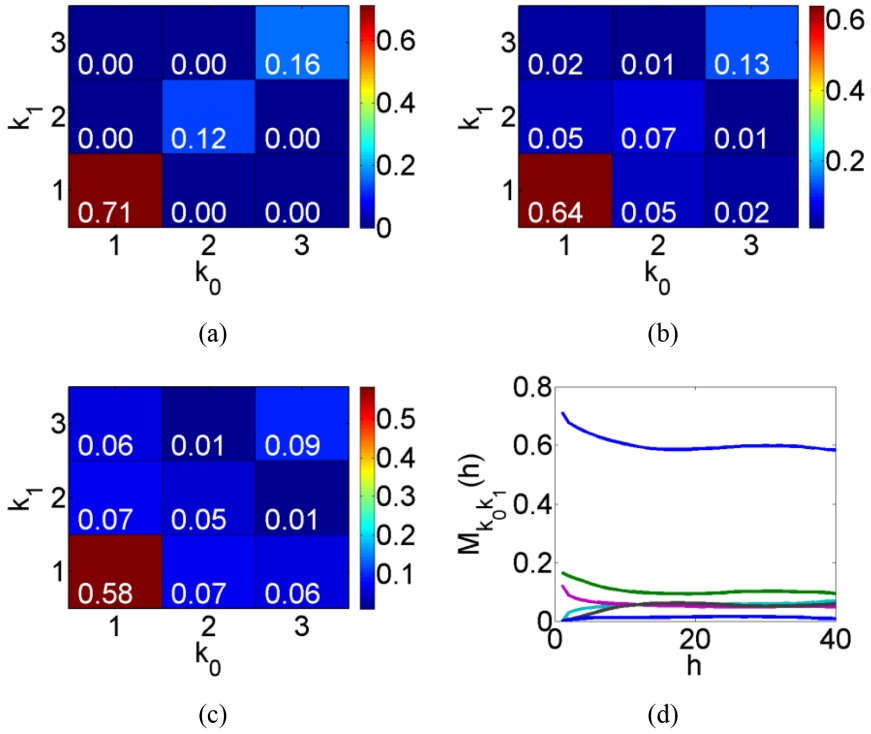
For the sake of demonstration, consider data from the Stanford V reservoir case study (Mao and Journel 1999) on Figure 1a and its categorization on Figure 1b, size of image is $N_x \times N_y$ pixels. Let $W_{i,j}$ be a value at pixel (i, j) of the categorized image, where $i = 1 \dots N_x, j = 1 \dots N_y$. If the image $W_{i,j}$ describes statistical properties of the random vector $\mathbf{Z}$, then the estimation of indicator moment $\hat{M}_{k_0 k_1}(\mathbf{h}_1)$ on the lag $\mathbf{h}_1 = (h, 0)$ can be calculated using pairs $\{W_{i,j}, W_{i+h,j}\}$. From now on, consider that the direction of $\mathbf{h}_1$ is $\mathbf{e}_1 = (1, 0)$ and fixed, then $M_{k_0 k_1}(\mathbf{h}_1)$ is the function of distance h.

The sections of the function $M_{k_0 k_1}(h)$ for fixed h equal to 0, 5, and 40 pixels are shown on Figure 2a-c, respectively. Figure 2d presents the sections of the function $M_{k_0 k_1}(h)$ for fixed values k_0, k_1. Each line correspond to one of the 3x3 possible combinations of k_0 and k_1.



(a) (b)

Figure 1: Image of continuous field and its categorization.



(a) (b)

(c) (d)

Figure 2: The estimation of the second-order spatial indicator moment $\hat{M}_{k_0k_1}(h)$: (a) zero-distance $h = 0$ indicator moments; (b) indicator moments on the lag $h = 5$; (c) indicator moments for far separated points $h = 40$; (d) the sections with different combinations of k_0, k_1 depends on h. Each line on (d) corresponds to value in one of 3x3 cells in (a), and its evolution across different lag separation (b) and (c).

It is not hard to see that for $h = 0$ only diagonal elements are not equal to zero:

$$M_{k_0k_1}(0) = P(Z_{i_0} = k_0, Z_{i_0} = k_1) = P(Z_{i_0} = k_0)\delta_{k_0,k_1} = M_{k_0}\delta_{k_0,k_1}, \tag{6}$$

where δ_{k_0,k_1} is Kronecker delta and $M_{k_0}$is the marginal distribution. Moreover, for two distant locations $h \to \infty$ the values Z_{i_0} and Z_{i_1} can be considered as independent random variables and the indicator moment can be factorized:

$$M_{k_0k_1}(h \to \infty) = M_{k_0}M_{k_1}. \tag{7}$$

That can be traced in the function behavior on Figure 2d. These functions represent two-point spatial cross-relations, similar to indicator covariances, and satisfy boundary conditions (6) and (7).

2.2 High-order spatial indicator moments

In the multi-point case, consider $n+1$ random variables Z_{i_0} and $Z_{i_l}, l = 1 \dots n$, then, the spatial configuration is defined by vectors $\mathbf{h}_l = \mathbf{x}_{i_l} - \mathbf{x}_{i_0}, l = 1 \dots n$. From now on, consider that directions of $\mathbf{h}_l$ are defined and fixed, then the spatial indicator moments are the functions of distances $h_l = \|\mathbf{h}_l\|, l = 1 \dots n$.

Therefore, the high-order spatial indicator moment can be expressed as follows:

$$M_{k_0 k_1 \dots k_n}(Z_{i_0}, Z_{i_1}, \dots Z_{i_n}) = M_{k_0 k_1 \dots k_n}(h_1, h_2, \dots h_n). \tag{8}$$

Hereafter, the following concise notation is used: $\mathbf{k} = k_0 \dots k_n$,$\mathbf{h} = h_1, \dots, h_n$.

Similar to the case of the second-order statistics, boundary conditions can be expressed through lower-order:

$$M_k(h_1, \dots, h_p = 0, \dots, h_n) = M_{k \setminus k_p}(\mathbf{h} \setminus h_p)\delta_{k_0, k_p}, \qquad \forall p \in 1 \dots n, \tag{9}$$

where $\mathbf{h} \setminus h_p$ denotes all the lags $\mathbf{h}$ excluding the lag h_p. Similarly for $\mathbf{k} \setminus k_p$.

If the directions are quite different, then additional boundary conditions are valid:

$$M_k(h_1, \dots, h_p \to \infty, \dots, h_n) = M_{\mathbf{k} \setminus k_p}(\mathbf{h} \setminus h_p) M_{k_p}, \qquad \forall p \in 1 \dots n. \tag{10}$$

Thus, the high-order spatial indicator moments are bounded with lower-order moments and this information should be taken into account during simulation.

For example, in case of three-point relations, for the image $W_{i,j}$ the sampling third-order spatial indicator moment $\hat{M}_{k_0 k_1 k_2}(h_1, h_2)$ of random variables separated by the vectors $\mathbf{h}_1$, $\mathbf{h}_2$ with directions $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$ can be calculated using triplets $\{W_{i,j}, W_{i+h_1,j}, W_{i,j+h_2}\}$. The indicator moment $\hat{M}_{111}(h_1, h_2)$ is shown on Figure 3. The values of the function $\hat{M}_{111}(h_1, h_2)$ on boundaries$(h_1, 0)$, $(h_1, 50)$, $(0, h_2)$, and $(50, h_2)$ correspond to two-point statistics shown on Figure 2d.

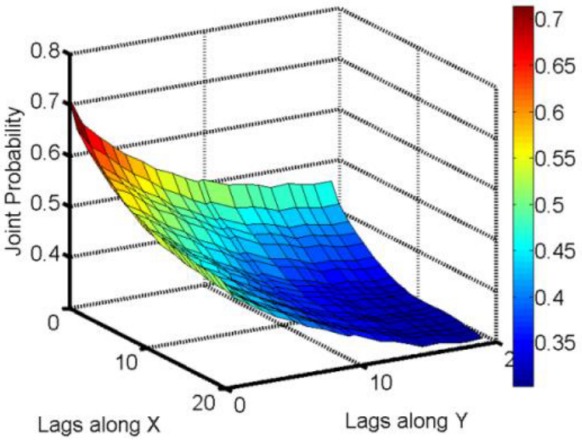


Figure 3: The third-order spatial indicator moment $\hat{M}_{111}(h_1, h_2)$.

3 Mathematical model

In this paper, multi-dimensional B-spline approximation under constrains (8) and (9) is used to model the high-order spatial indicator moments $M_{\mathbf{k}}(\mathbf{h})$. Consider the fixed categories $k_0, \dots k_n$, then, the function $M_{\mathbf{k}}(\mathbf{h})$ is a multi-dimensional function which values are known at the limited number of points $\mathbf{h}^d = (h_1^d, \dots, h_n^d), d = 1 \dots m$ estimated from the hard data. The calculation of sampling indicator moments $\hat{M}_{\mathbf{k}}(\mathbf{h}^d)$ is presented in the subsequent section. Then, the function $M_{\mathbf{k}}(\mathbf{h})$ can be approximated using the following recursive model:

$$M_{\mathbf{k}}(\mathbf{h}) = M_{\mathbf{k}}^0(\mathbf{h}) + \delta M_{\mathbf{k}}(\mathbf{h}), \tag{11}$$

$$M^0_{\mathbf{k}}(\mathbf{h}) = \frac{1}{\sum_{p=1}^{n} e^{-ah_p} + e^{-a(1-h_p)}} \left[\sum_{p=1}^{n} M_{k_0 \dots k_n}(h_1, \dots, h_p = 0 \dots, h_n) e^{-ah_p} + \right.$$
$$\left. + \sum_{p=1}^{n} M_{k_0 \dots k_n}(h_1, \dots, h_p \to \infty \dots, h_n) e^{-a(1-h_p)} \right], \qquad \forall p = 1 \dots n, \quad (12)$$

$$\delta M_{\mathbf{k}}(\mathbf{h}) = \sum\nolimits_{i_1=1}^{\omega} \cdots \sum\nolimits_{i_n=1}^{\omega} \alpha_{i_1,\dots,i_n} B_{i_1,r}(h_1) \dots B_{i_n,r}(h_n), \quad (13)$$

where user-defined parameter a determines the influence of the boundary conditions, $\alpha_{i_1,\dots,i_n}$ are coefficients of B-splines approximation, and $B_{i,r}(t)$ is i-th B-splines of order r on uniformly divided knot sequence $\{t_0, t_1, t_2, \dots, t_p\}$, where knots are separated by step $dt = (t_p - t_0)/p$, $t_0 = 0$, and t_p are equal to the minimal lag distance at which the variables can be considered as independent.

The coefficients $\alpha_{i_1,\dots,i_n}$ are found using least-square algorithm to fit points:

$$\delta M_{\mathbf{k}}(\mathbf{h}^d) = \hat{M}_{\mathbf{k}}(\mathbf{h}^d) - M^0_{\mathbf{k}}(\mathbf{h}^d), \qquad d = 1 \dots m, \quad (14)$$

under zero boundaries constrains:

$$\begin{array}{l} \delta M_{\mathbf{k}}(h_1, \dots, h_p = 0, \dots, h_n) = 0, \\ \delta M_{\mathbf{k}}(h_1, \dots, h_p \to \infty, \dots, h_n) = 0, \qquad \forall p = 1 \dots n \end{array} . \quad (15)$$

In this paper, the additional regularization condition of minimum curvature (Wang et al., 2006) is used to avoid overfitting.

The high-order moments are recursively constructed by starting from the second-order indicator moments. First, second-order indicator moments $M_{k_0 k_p}(h_p), p = 1 \dots n$ are calculated from the basic variogram model (David 1977). Then, the trend $M^0_{k_0,k_1,k_2}(h_1, h_2, h_3)$ is calculated using equation (12) and relations to the lower-orders (9) and (10). Further, the residuals $\delta M_{k_0,k_1,k_2}(h_1^d, h_2^d, h_3^d)$ can be estimated from sampling indicator moments $\hat{M}_{k_0,k_1,k_2}(h_1^d, h_2^d, h_3^d)$ and equation (14). These residuals are used as points in the B-spline approximation (13) of the multi-dimensional function $\delta M_{k_0,k_1,k_2}(h_1, h_2, h_3)$ under zero-boundary constrains (15). Finally, the third-order spatial indicator moments are retrieved using equation (11). The same procedure is recursively repeated for fourth, fifth, and higher-orders until the desired order is achieved.

3.1 Calculation of sampling statistics

The octant model is used (Figure 4) to estimate the sampling indicator moments $\hat{M}_{\mathbf{k}}(\mathbf{h}^d)$ from the hard data. The neighborhood area of each hard data sample is divided into $N_\phi = 8$ sectors representing N_ϕ directions. Then, each sector is divided into N_r lags and forms an $N_r \times N_\phi$ bin template. Only one point within each bin is randomly chosen to construct a replicate. Finally, the values $\hat{M}_{\mathbf{k}}(\mathbf{h}^d)$ are estimated from replicates using law of large numbers:

$$\hat{M}_{\mathbf{k}}(\mathbf{h}^d) = \frac{1}{N_{\mathbf{h^d}}} \sum_{j=1}^{N_{\mathbf{h^d}}} I_{k_0}(z^j_{i_0}) \dots I_{k_n}(z^j_{i_n}), \qquad d = 1 \dots m, \quad (16)$$

where the sum is taken over all $N_{\mathbf{h^d}}$ replicates with the spatial configuration $\mathbf{h}^d$, data samples $z^j_{i_0} \dots z^j_{i_n}$ in the replicate j are separated by lags $\mathbf{h}^d$, and d is the index of different spatial configurations $\mathbf{h}^d$.

It should be noted, that replicates separated by at least half the variogram range should be used for the law of the large numbers to be applicable.

The amount of information about high-order statistics that can be retrieved from data crucially depends on the number of categories K, the total number of data samples N and the level of correlation between values. It is not hard to see that the higher order of statistics considered are, the larger the number of samples available should be.

In order to have an adequate number of replicates for a particular order m of statistics, the minimum number of replicates $N_{repl}(m) \leq N_{\mathbf{h^d}}, \forall \mathbf{h^d}$ is set up by the user. However, more advanced techniques based on an entropy or information theory should be considered (Arndt 2004). Having a minimum number of replicates $N_{repl}(m)$ for the given order of statistics m, the optimal number of lags N_r^{opt} is calculated.

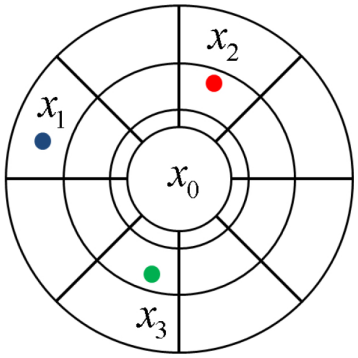


Figure 4: Octant model for calculation of the sampling joint distribution $\hat{M}_{\mathbf{k}}(\mathbf{h}^d)$.

4 Simulation algorithm

Combining all of the above, the new data-driven algorithm is formulated as follows:

Algorithm 1

1. Starting from two-point statistics $m = 2$ until stopping criteria is reached
   (a) Define the minimum number of replicates $N_{repl}(m)$ for given order m.
   (b) Scan the hard data using the octant model (Figure 4) with different N_r starting with the number of the radial divisions $N_r = 2$ and find the higher N_r for which the average number of replicates is bigger than $N_{repl}(m)$.
   (c) If $N_r = 2$ and the average number of replicates is less than $N_{repl}(m)$, then exit the loop.
   (d) Save all the replicates for obtained N_r and the order of statistics m.
   (e) Increase the order of statistics $m = m + 1$.
2. Define a random path visiting all the unsampled nodes.
3. For each node $\mathbf{x}_{i_0}$ in the path:
   (a) Find the closest data samples $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots \mathbf{x}_{i_n}$. The categories at these nodes are denoted by $k_1, \ldots k_n$.
   (b) For all $k_0 = 1 \ldots K$ calculate the high-order spatial indicator moments $\hat{M}_{\mathbf{k}}(\mathbf{h}^d)$ using formula (15) from the replicates found in step 1 and recursive model (11)–(13). Note that $k_1, \ldots k_n$ are fixed. For the orders higher than maximum order m consider $\delta M_{\mathbf{k}}(\mathbf{h}) \equiv 0$.
   (c) Calculate the conditional distribution from joint distribution:
   $$P\left[Z_{i_0} = k_0 | Z_{i_1} = k_1 \ldots, Z_{i_n} = k_n\right] = A M_{\mathbf{k}}(\mathbf{h}), \tag{17}$$
   where coefficient A is the normalization coefficient:
   $$A = 1 / \sum_{k_0=1}^{K} M_{\mathbf{k}}(\mathbf{h}). \tag{18}$$
   (d) Draw a random value z_{i_0} from this conditional distribution (17) and assign it to the unsampled location $\mathbf{x}_{i_0}$.
   (e) Add z_{i_0} to the set of sample hard data and the previously simulated values.
4. Repeat Steps 3a-e for all the points along the random path defined in Step 2.

5 Simulation results

The proposed approach is tested on the data set from the Stanford V reservoir case study (Figure 1a). This image is discretized on categories 0, 1 and 2, and is used as a reference image (Figure 5a). Hard data is randomly selected from the image and shown in Figure 5b. This represents 520 points (5% of the image points). The results are compared with sequential indicator simulation algorithm (sisim; Journel and Alabert, 1990; Deutsch and Journel, 1998).

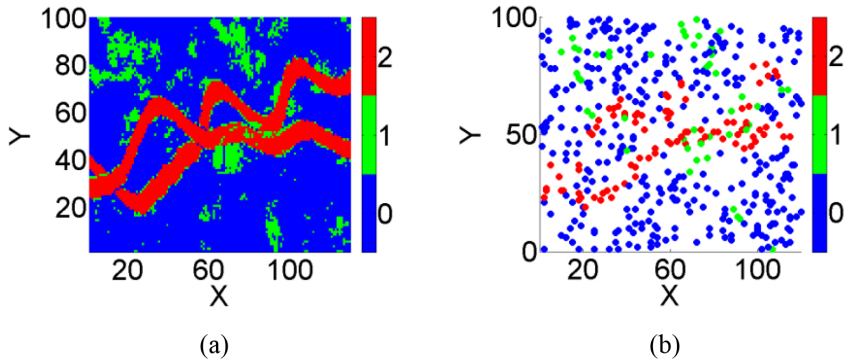


Figure 5: Case study with 520 data samples: (a) the reference image, (b) data samples.

The simulation results for the case with 520 data samples are shown in Figure 6. Neither the training image nor the reference image (Figure 5a) are used during the simulation and are presented herein only for the sake of comparison. Simulations are done in two modes: using only boundary conditions (Figure 6b), and using both conditions (14) and (15) (Figure 6c).

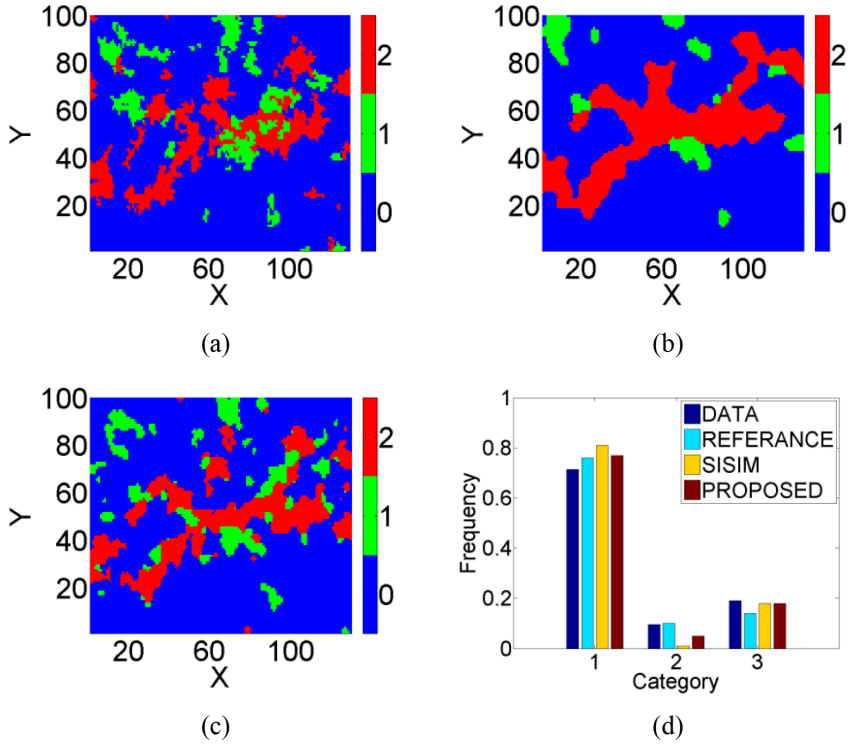


Figure 6: Case study with 520 data samples: (a) sisim simulation result, (b) the simulation using proposed algorithm with only boundary conditions, (c) the simulation result using both boundary conditions and high-order statistics from data. Subfigure (d) shows histograms for data samples, the reference image, sisim simulation, and the simulation using the proposed technique presented by blue, light blue, yellow and red bars, respectively.

In the case of using just boundary conditions, the result is smooth and the width of channels is overestimated because all high-order statistics are derived from the second-order statistics. However, sisim simulation results (Figure 6a) are less connected and the channels can be hardly detected. The result obtained with the account of higher-order statistics from data (Figure 6c) reproduces the channels quite well with adequate dimensions of geometrical bodies.

On Figure 6c, the histograms for data samples, the reference image, sisim simulation, and the simulation using the proposed technique are shown by blue, light blue, yellow, and red bars, respectively. The deviations from the distribution in hard data are small for both sisim and proposed algorithm simulations.

The second-order statistics are compared in Figure 7a. The direction $\mathbf{e}_1 = (1, 0)$ is used. The indicator moments $M_{01}(\_h_1)$ of simulations using sisim and the proposed algorithm are reproduced well. Nevertheless, the third-order statistics $M_{012}(\_h_1, h_2)$ of the reference image (Figure 7b), simulation using sisim algorithm (Figure 7c), and the result of the proposed technique (Figure 7d) are quite different.

The third-order spatial indicator moments are calculated using directions $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$. Some similarities of patterns can be traced in the bottom part of Figure 7b and d, which correspond to statistics of the reference image and the simulation using the proposed algorithm. However, the point of interest is the reproduction of the high-order spatial statistics of the hard data.

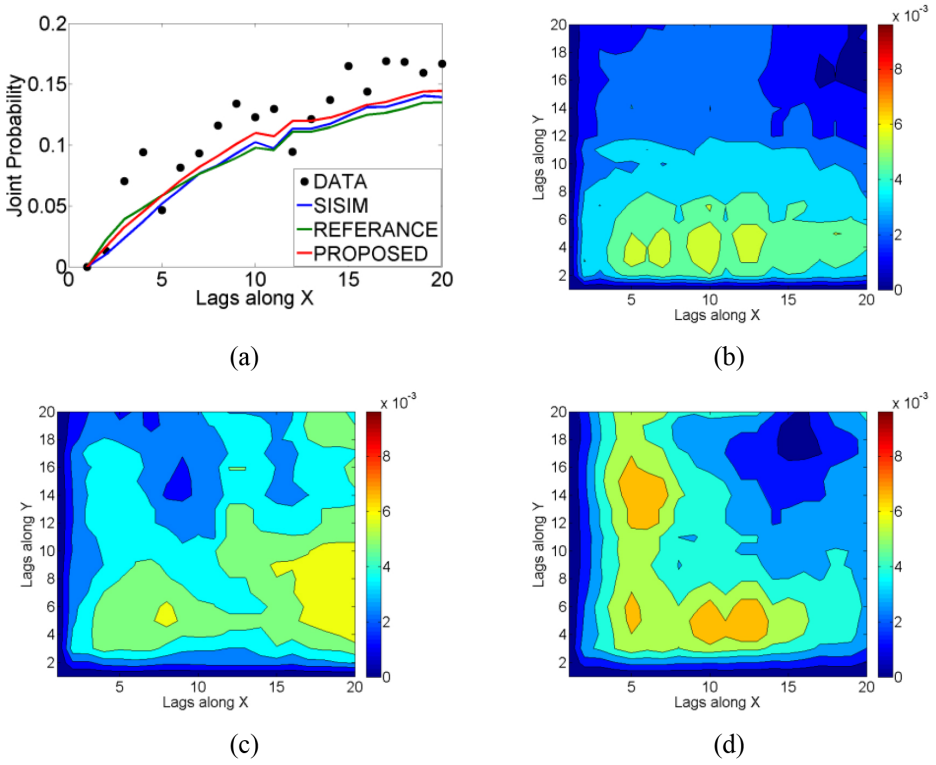


Figure 7: Case study with 520 data samples. (a) Second-order spatial indicator moment $M_{01}(\_h_1)$ using directions $\mathbf{e}_1 = (1, 0)$ for data samples, the reference image, sisim simulation, and the simulation using proposed technique are presented by black dots, and green, blue, red lines, respectively. The third-order spatial indicator moment $M_{012}(\_h_1, h_2)$ using directions $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$ for: (b) the reference image, (c) sisim simulation result, (d) the simulation using both boundary conditions and high-order statistics from data.

Colors show the number of triplets found in data using template on Figure 4.

The surface on Figure 8 is a 3D view of Figure 7d. Dots represent the statistics calculated from the hard data. Colors show the number of triplets used for the calculation of the third-order spatial indicator moment $M_{012}(\_h_1, h_2)$. The higher the number of triplets, the more reliable the value of the point is. The spatial indicator moment $M_{012}(\_h_1, h_2)$ of the simulation using the proposed algorithm tends to fit more reliable points and is consistent with the boundary conditions (Figure 8).

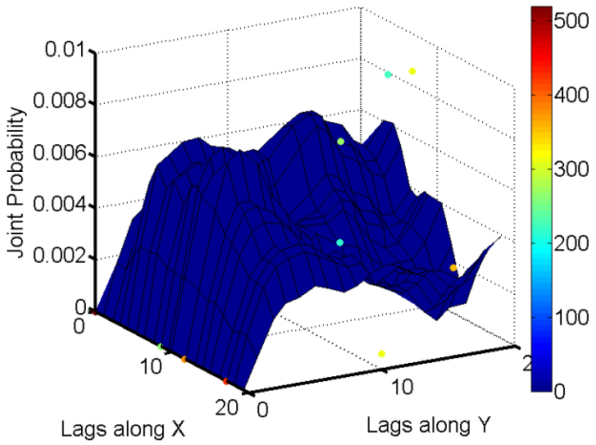


Figure 8: The third-order spatial indicator moment $M_{012}(\_h_1, h_2)$ using directions $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$ for the simulation using both boundary conditions and high-order statistics from data.

6 Conclusions and future work

This paper presented a new data-driven high-order sequential method for the simulation of categorical random fields. The sequential algorithm is based on the B-spline approximation of high-order spatial indicator moments that are consistent with each other. The main distinction from commonly used MPS methods is that, in the proposed technique, conditional distributions are constructed using high-order spatial indicator moments as the functions of distances based on hard data. The simulations herein are generated without a TI. Note that in applications with relatively large numbers of data, as in the simulation of mineral deposits, the higher-order statistics are deduced from hard data. The option of adding a TI to a dataset is available for sparse datasets.

The basic concept of the algorithm is to use recursive approximation models with enclosed boundary conditions, which are derived from the nested nature of high-order spatial indicator moments presented herein. To provide robust estimation the regularized B-splines are used.

Another important aspect is the different amount of information that can be retrieved for different levels of relations. In the proposed method, each order of spatial statistics is approximated using the appropriate number of B-splines to provide robustness to the algorithm and to avoid overfitting. Thus, lower-order statistics are estimated with higher resolution than the higher-order statistics.

The simulation algorithm is tested on the categorized data from the Stanford V reservoir case study and compared with results of the *sisim* algorithm. No TI is used during simulations. According to the results, the proposed method reproduces the complex spatial patterns, such as channels, and preserves high-order statistics.

The proposed technique is fully data-driven; however, the information from the TI can be incorporated with the suggested model as a trend to capture high-frequency features of the TI. Further research is concerned with the application to 3D models, improving the efficiency, testing for unbiasedness of the proposed approximation model, and generalization to the continuous random fields.

References

Arndt, C. (2004). Information Measures: Information and its Description in Science and Engineering. Amsterdam: Springer.

Caers, J. (2005). Petroleum geostatistics. Houston: SPE–Pennwell Books.

Chilès, J. P., and Delfiner, P. (2012). Geostatistics: modeling spatial uncertainty. 2nd Edition. New York: Wiley.

Chugunova, T. L., and Hu, L. Y. (2008). Multiple-point simulations constrained by continuous auxiliary data. Math Geosci, 40, 133–146.

Cressie, N. A. (1993). Statistics for spatial data. New York: Wiley.

David, M. (1977). Geostatistical ore reserve estimation. Amsterdam: Elsevier.

David, M. (1988). Handbook of applied advanced geostatistical ore reserve estimation. Amsterdam: Elsevier.

Deutsch, C., and Journel, A. (1998). GSLIB: Geostatistical Software Library and User's Guide (Second ed.). New York: Oxford University Press.

Dimitrakopoulos, R., Mustapha, H., and Gloaguen, E. (2010). High-order statistics of spatial random fields: exploring spatial cumulants for modeling complex non-Gaussian and non-linear phenomena. Math Geosci, 42, 65–99.

Goovaerts, P. (1998). Geostatistics for natural resources evaluation. Cambridge University Press: Cambridge.

Guardiano, J., and Srivastava, R. M. (1993). Multivariate geostatistics: Beyond bivariate moments. Geosatistics Tróia '92, 133–144.

Journel, A. G. (1993). Geostatistics: Roadblocks and Challenges. Stanford Center for Reservoir Forecasting.

Journel, A. G. (2003). Multiple-point Geostatistics: A State of the Art. Stanford Center for Reservoir Forecasting.

Journel, A. G., and Alabert, F. (1990). New Method for reservoir mapping. Petroleum Technology, 42(2), 212–218.

Journel, A. G., and Huijbregts, C. J. (1978). Mining geostatistics. San Diego: Academic Press.

Kitanidis, P. K. (1997). Introduction to geostatistics—Applications in hydrogeology. Cambridge : Cambridge Univ Press.

Mao, S. a. (1999). Generation of a reference petrophysical/seismic data set: the Stanford V reservoir. 12th Annual Report. Stanford: Stanford Center for Reservoir Forecasting.

Mariethoz G, R. P. (2010). Reconstruction of incomplete data sets or images using direct sampling. Math Geosci, 42, 245–268.

Mariethoz, G., and Kelly, B. (2011). Modeling complex geological structures withelementary training images and transform-invariant distances. Water Resour Res, 47, 1-2.

Mariethoz, G., and Renard, P. (2010). Reconstruction of incomplete data sets or images using direct sampling. Math Geosci, 42, 245–268.

Matheron, G. (1971). The theory of regionalized variables and its applications. Cahier du Centre de Morphologie Mathematique, No 5.

Mustapha, H., and Dimitrakopoulos, R. (2010a). A new approach for geological pattern recognition using high-order spatial cumulants. Computers & Geosciences, 36(3), 313-334.

Mustapha, H., and Dimitrakopoulos, R. (2010b). High-order stochastic simulations for complex non-Gaussian and non-linear geological patterns. Math Geosci, 42, 455-473.

Pyrcz, M., and Deutsch, C. (2014). Geostatistical Reservoir Modeling (Second ed.). New York: Oxford University Press.

Remy, N., Boucher, A., and Wu, J. (2009). Applied geostatistics with SGeMS: A user's guide. Cambridge: Cambridge University Press.

Straubhaar J, R. P. (2011). An improved parallel multiple-point algorithm using a list approach. Math Geosci, 43, 305–328.

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple point statistics. Math Geosci, 34, 1–22.

Strebelle, S., and Cavelius, C. (2013). Solving speed and memory issues in multiple-point statistics simulation program SNESIM. Math Geosci, 46, 171–186.

Tikhonov, A. N., and Arsenin, V. Y. (1977). Solution of Ill-posed Problems. Washington: Winston & Sons.

Toftaker H, T. H. (2013). Construction of binary multi-grid Markov random field prior models from training images. Math Geosci, 45, 383–409.

Vargas-Guzmán, J. (2011). The Kappa model of probability and higher-order rock sequences. Comput. Geosci., 15, 661–671.

Vargas-Guzmán, J., and Qassab, H. (2006). Spatial conditional simulation of facies objects for modeling complex clastic reservoirs. J. Petrol. Sci. Eng., 54, 1–9.

Wang, W., Pottmann, H., and Liu, Y. (2006). Fitting B-spline curves to point clouds by curvature-based squared distance minimization. ACM Transactions on Graphics, 25(2), 214–238.

Webster, R., and Oliver M, A. (2007). Geostatistics for environmental scientists. New York: Wiley.

Yahya, W. J. (2011). Image reconstruction from a limited number of samples: A matrix-completion-based approach. Montreal.

Zhang, T., Switzer, P., and Journel, A. (2006). Filter-based classification of training image patterns for spatial simulation. Math Geosci, 38(1), 63–80.