Les Cahiers du GERAD ISSN: 0711–2440

High-order data-driven spatial simulation using Legendre-like orthogonal splines

I. Minniakhmetov,
R. Dimitrakopoulos

G–2016–98

November 2016

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

High-order data-driven spatial simulation using Legendre-like orthogonal splines

Ilnur Minniakhmetov [a]
Roussos Dimitrakopoulos [a, b]

[a] *COSMO Stochastic Mine Planning Laboratory, Department of Mining and Materials Engineering, McGill University, Montréal (Québec) Canada, H3A 0E8*

[b] *GERAD Montréal (Québec) Canada, H3T 2A7*

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

November 2016

Les Cahiers du GERAD
G–2016–98

Abstract: High-order sequential simulation techniques for complex and non-Gaussian spatially distributed variables were developed over the last few years. This paper presents a new extension where high-order statistics are inferred from the available hard data and then missing high-orders are borrowed from a training image. The inferred high-order statistics are then used to estimate spline coefficients that are subsequently employed to approximate conditional probability density functions as needed for the simulation process. The advantage of using orthogonal splines with respect to standard approaches is their ability to approximate a probability density function locally using not only high-order spatial cumulants for the whole range of values in data, but also partial cumulants calculated for particular ranges of values, such as extreme values. The proposed technique provides a general framework for simulation, both of continuous and categorical variables. Developments are tested on a synthetic data set.

Keywords: Stochastic simulation, data-driven, splines, high-order spatial statistics, non-Gaussian distribution, multipoint statistics

1 Introduction

Geostatistical simulations are used to quantify the uncertainty of spatial attributes of interest describing mineral deposits, petroleum reservoirs, hydrogeological horizons, environmental contaminants and others. Since the 1990's, multiple point spatial simulation (MPS) methods and variations (Guardiano and Srivastava, 1993; Strebelle, 2002; Strebelle and Zhang, 2005; Journel, 2005, 2007; Zhang et al., 2006; Chuginova and Hu, 2008; Straubhaar et al., 2010; De Iaco and Maggio, 2011; Honarkhah, 2011; Tjelmeland, 2013; Kolbjørnsen, 2014; Strebelle and Cavelius, 2014; Chatterjee et al., 2015; others) have been developed to advance the simulation technologies beyond the past generation of second-order spatial statistics typically combined with Gaussian processes (e.g., David, 1988; Goovaerts, 1997; Chiles and Delfiner, 2012). A core limitation of MPS approaches is that they are largely algorithmic and do not consistently account for the high-order spatial relations in the available hard data. Patterns and complex spatial relations are derived from the so-termed training images (TI) or geological analogues, rather than from hard data, a topic critical for relatively data-rich type applications (eg. Osterholt and Dimitrakopoulos, 2007; Goodfellow et al., 2012). To address some of these limits, high-order simulation techniques for complex and non-Gaussian spatially distributed variables have also been developed (Mustapha and Dimitrakopoulos, 2010; 2011) based on generating conditional distributions through Legendre polynomials and high-order spatial cumulants introduced by Dimitrakopoulos et al. (2010). To improve these in terms of approximating probability density functions (pdf) during the sequential simulation process, as well as to generalize the proposed framework for both continuous and categorical variables, orthogonal splines are considered herein.

Hereafter, methods that use complex spatial relations derived from TI are called TI-driven, whereas approaches focused on the reproduction of spatial relations of hard data are called data-driven. The topic of data-driven MPS simulations has already been addressed. Attempts to address it may also be found in direct sampling (Mariethoz and Renard, 2010), where the sequential MPS algorithm draws random replicates from a TI and hard data that correspond to spatial configuration of conditional data. The similarity measure of a data event and drawn replicate is calculated and, if a certain threshold is reached, the value in the central node of the replicate is assigned to a grid node. The approach is a random drawing from implicitly approximated conditional probability density function, which does not ensure that high-order statistics of data are reproduced. The high-order simulation framework (Mustapha and Dimitrakopoulos, 2010, 2011; Boogaart et al., 2014) addresses the above and it is based on the Legendre series approximation (Lebedev, 1965) of a conditional pdf at each node to be simulated. The approach is data-driven and uses high-order spatial statistics from hard data, which are complemented by high-order spatial statistics from training image. Here the high-order spatial statistics are shown to capture directional multiple-point periodicity, connectivity (including connectivity of extreme values), and spatial architecture (Dimitrakopoulos et al., 2010).

In an effort to improve upon the estimation of conditional pdfs within the high-order simulation framework, a spline approximation of complex multi-dimensional functions (Piegl, 1989; Hughes, 2005) is considered here. Splines are functions that are piecewise-defined by polynomials, which are connected by some condition of smoothness at the knots. Splines are flexible tools in dealing with discontinuities (Sinha and Schunck, 1992; López de Silanes et al., 2001;) and functions with locally high gradients (Malagù et al., 2014). Additionally, through the proper choosing of knots sequence splines can accurately approximate very complex functions, such as the shapes of three dimensional objects in a computer-aided geometric design (Hoschek and Lasser, 1993; Park and Lee, 2007). That is why splines are chosen herein for approximating of complex multidimensional joint distributions.

However, B-splines do not compose the orthogonal system of functions and therefore cannot be used in the framework proposed by Mustapha and Dimitrakopoulos (2010). In this paper Legendre-like splines (Wei et al., 2013) are used, which are shown to be orthogonal and can be easily integrated in the high-order simulation framework. This spline approach shows improvements in estimating conditional pdfs and resulting simulated realizations in terms of numerical stability and ability to reproduce distribution of extreme values, while also allows the simulation of both continuous and categorical variables.

The paper organized as follows. In Section 2 the high-order simulation framework is outlined. Then, the limitations of the high-order simulation technique using Legendre polynomials are demonstrated and a new approach using Legendre-like orthogonal splines is proposed. In Section 3, a novel data-driven algorithm is introduced within the high-order simulation framework. Further, in Section 4, the proposed approaches are tested on fully-known datasets and conclusions follow.

2 High-order simulation

Let $Z(x_i)$ or Z_i be a random field indexed in R^n, $x_i \in D \subseteq R^n (n = 1,2,3), i = 1 \dots N$, where N is the number of points in a discrete grid $D \subseteq R^n$. The focus of high-order simulation techniques is to simulate the realization of the random field $Z(x_i)$ for all nodes of a grid D with a given set of conditioning data $d_k = \{Z(x_\alpha), \alpha = 1 \dots K\}$ and high-order spatial statistics derived from the training image $Y(x_i)$. The high-order simulation method proposed by Mustapha and Dimitrakopoulos (2010; 2011) uses Legendre polynomials and coefficients in terms of high-order statistics to approximate the conditional probability density function at each node of the simulation grid. This method is presented by the following Algorithm A.1.

Algorithm A.1

1. Assign conditioning data values d_k to the grid's closest nodes of the simulation grid D in term of Euclidian distance. These nodes are called sampled, whereas the rest of the nodes are referred to as unsampled ones.
2. Define a random path visiting all the unsampled nodes.
3. For each node x_0 in the path:
   a. Find the closest sampled grid nodes x_1, x_2, x_n. The conditioning data at these nodes are denoted by $z_1, \dots, z_n$.
   b. Define the template shape $T_{n+1}^{e_1,e_2,\dots e_n}(h_1, h_2, \dots, h_n)$ for the unsampled location x_0 using its neighbors:

$$T_{n+1}^{e_1,e_2,\dots e_n}(h_1, h_2, \dots, h_n) = \begin{Bmatrix} (x, x + h_1, \dots, x + h_n) \\ / \{x, x + h_1, i = 1 \dots n\} \end{Bmatrix}, \quad (1)$$

   where e_i are unit directional vectors from x_0 to x_i , correspondingly, and h_i are distances from x_0 to x_i.
   c. Search all the replicates by scanning a TI with the template $T_{n+1}^{e_1,e_2,\dots e_n}(h_1, h_2, \dots, h_n)$. The set of replicates $\{R_m^i\}_{m=1\dots M}^{i=1\dots n}$ obtained is then given by:

$$\begin{matrix} R_m^0 = Y(x_m) \\ R_m^i = Y(x_m + h_i), i = 1 \dots n \\ \{x_m, x_m + h_1, x_m + h_n\} \in T_{n+1}^{e_1,e_2,\dots e_n}(h_1, h_2, \dots, h_n) \end{matrix}, \quad (2)$$

   where $m = 1 \dots M, M$ is user defined number of replicates, and x_m is the central node of the replicate.
   d. Calculate the coefficients of the Legendre polynomial approximation

$$L_{i_0,i_1,\dots i_n} = \frac{1}{M} \sum_{m=0}^{M} P_{i_0}(R_m^0) P_{i_1}(R_m^1) \dots P_{i_n}(R_m^n), \quad (3)$$

   where P_{i_n} is Legendre polynomial of order i_n whose explicit form will be discussed in section 2.1.
   e. Build the conditional probability density function $f_{z_0}(z_0|z_1, z_2, \dots, z_n)$ of the random variable Z_0 at the unsampled location x_0 given the conditioning data $z_1, \dots z_n$ at the nodes $x_1, x_2, \dots x_n$, correspondingly

$$f_{Z_0}(z_0 \mid z_1, z_2, \dots, z_n) = C \sum_{i_0=0}^{r} \sum_{i_1=0}^{r} \dots \sum_{i_M=0}^{r} L_{i_0,i_1,\dots i_M} P_{i_0}(z_0) P_{i_1}(z_1) \dots P_{i_n}(z_n), \quad (4)$$

   where C is the normalization coefficient defined as $C = 1 / \int f_{z_0}(z_0|z_1, z_2, \dots, z_n)\, dz_0$ and r is the maximal order of approximation defined by user.
   f. Draw a random value from this conditional distribution (4) and assign it to the unsampled location x_0.
   g. Add Z_0 to the set of sample hard data and the previously simulated values.
4. Repeat Steps 3a-g for all the points along the random path defined in Step 2.

2.1 Legendre polynomials

Mustapha and Dimitrakopoulos (2010; 2011) proposed to use Legendre series for conditional pdf approximation (4), where P_m is the Legendre polynomial of order m defined as in Lebedev (1965):

$$P_m = \frac{1}{2^m m!}(z) = \left(\frac{d}{dz}\right)\left[(z^2 - 1)^m\right], -1 \le z \le 1. \quad (5)$$

The set of Legendre polynomials $\{P_m(z)\}_m$ forms a complete basis set on the interval $[-1,1]$ and the function $f(z)$ in the univariate case can be then approximated as follows

$$f(z) \approx \sum_{m=0}^{r} L_m P_m(z), \tag{6}$$

where L_m are coefficients of the Legendre polynomial approximation, r is the maximum order of approximation.

In the multivariate case, expression (6) becomes the equation (4). Legendre series perform well approximating the conditional pdfs and allow explicitly use of high-order spatial statistics. However, there are some practical limitations of using Legendre polynomials for approximation functions in a multidimensional space.

2.1.1 Sensitivity of polynomial approximation

Let us consider an approximation of the conditional pdf for a given unsampled location with 4 neighbors (Figure 1a). Half of the all replicates $M_1 = M / 2$ are randomly chosen from TI and approximations of the conditional pdf using Legendre series with different maximal orders r are calculated.

In Figure 1b, c, d the approximation results for 6 random sets of replicates with maximal order 20, 30, and 40 are represented by solid grey lines. The empirical conditional distribution, depicted by the histogram, is calculated using all replicates from TI.

It is not hard to see that as the maximal order of approximation is increased the approximation becomes unstable.

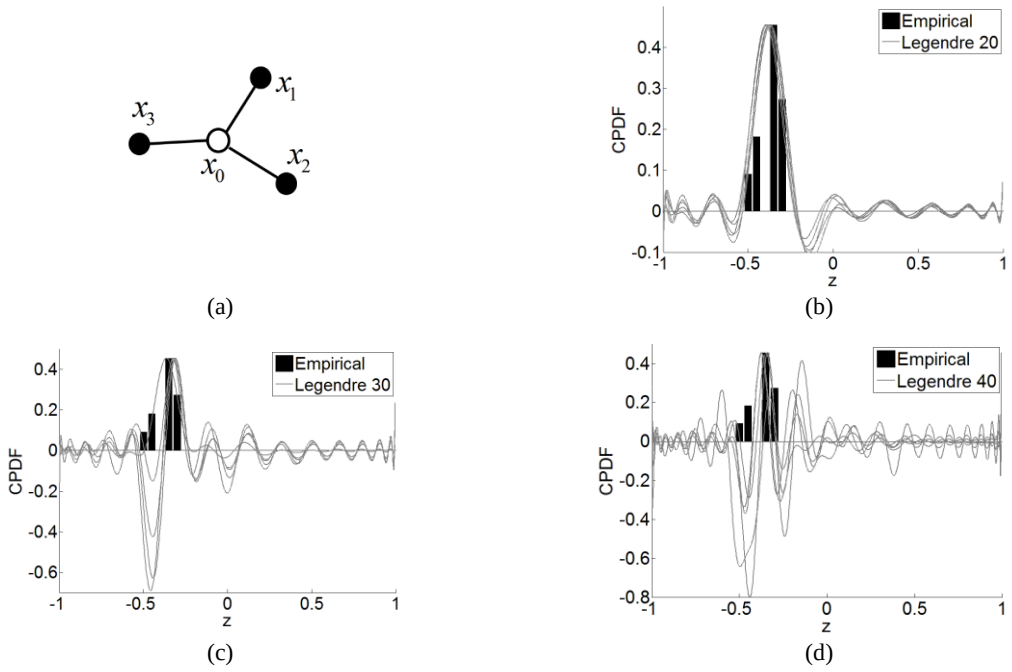


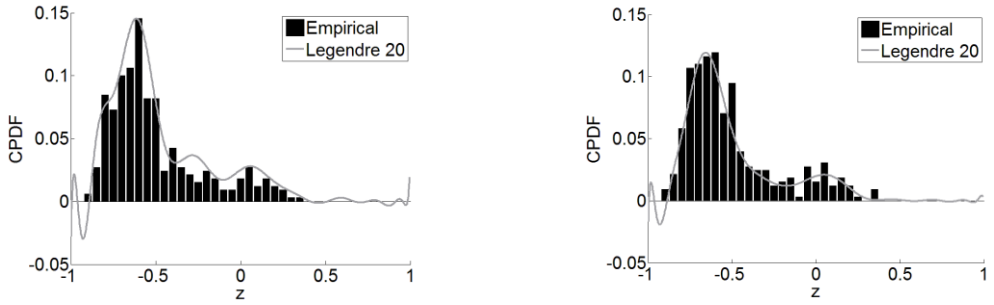
Figure 1. The stability of the polynomial approximation: (a) spatial configuration of the template, (b) the approximation using Legendre series up to order 20, (c) the approximation using Legendre series up to order 30, (d) the approximation using Legendre series up to order 40. Solid grey lines are the approximation results for different sets of replicates from TI. The empirical conditional distribution is depicted by the histogram.

2.1.2 Approximation of extreme values

In the next example the sensitivity of polynomial approximation to changes in the distribution of extreme values is analysed. The unsampled location with one neighbor is considered. Two empirical conditional pdfs with different

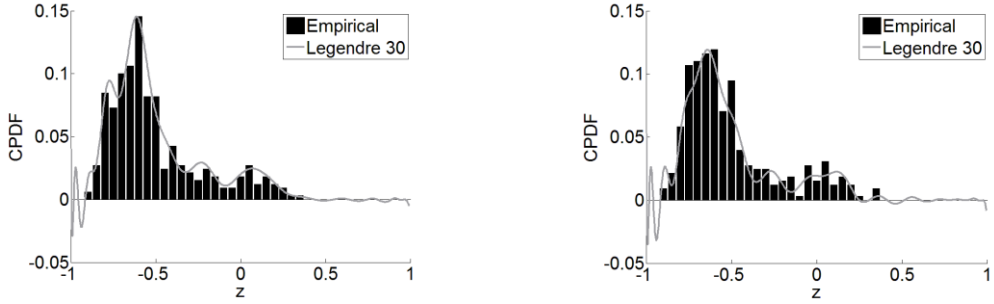
distributions of extreme values are compared with the approximations using Legendre polynomials with maximal orders 20 and 30, Figures 2 and 3, correspondingly.

It is quiet straightforward, that the approximation using Legendre series with maximal order 20 is not sensitive for changes in distribution of extreme values (Figure 2), whereas the approximation using maximal order of 30 takes into account even slight variation in distributions (Figure 3). However, using Legendre polynomial with maximal order 30 is shown to be unstable (Figure 1c).



(a) (b)

Figure 2. The approximation of conditional pdfs using Legendre polynomials with maximal order 20. Cases (a) and (b) differ by distribution of extreme values. The solid grey lines are the approximation results; histograms represent the empirical conditional distribution.



(a) (b)

Figure 3. The approximation of conditional pdfs using Legendre polynomials with maximal order 30. Cases (a) and (b) differ by distribution of extreme values. The solid grey lines are the approximation results; histograms represent the empirical conditional distribution.

2.1.3 Non-physical values of pdf

Legendre series are usually used for approximation of smooth functions. If discrete or piece-wise smooth pdf is considered a polynomial approximation can give high negative values or fluctuations around zero. For example, in Figure 1d the actual pdf is equal to zero on intervals [-1; -0.5], [-0.45; -0.4], and [-0.25; 1] and has several kinks on the edge of these intervals. As the result, the polynomial approximation has negative probability up to -0.7 and oscillations around zero. These numerical issues eventually result in unfounded outliers as in the case study in Section 4.

By all means, there are techniques to deal with all these artifacts (Wilson and Wragg, 1973), but the common algorithm can be sophisticated. Moreover, the actual conditional pdf is not known and the question whether the fluctuations correspond to the real conditional pdf or they are numerical artifacts should be solved providently for each unsampled location.

2.2 Legendre-like orthogonal splines

In addition to Legendre polynomials, any orthogonal system of functions can be used in the high-order simulation described in Algorithm A.1. In this work, Legendre-like splines (Wei et al., 2013) are used to overcome the limitation discussed in Paragraph 2.1.

Let us review the construction of Legendre like orthogonal splines of order r (Wei et al., 2013). Let $[a,b]$ be a domain for function approximation, which is described by the knot sequence:

$$T = \{\underbrace{a,a,\ldots,t_0=a}_{r+1} < t_1 \le t_2 \le \ldots \le t_{k_{\max}} < \underbrace{t_{k_{\max}+1}=b,b,\ldots,b}_{r+1}\}, \tag{7}$$

where knots $t_k, k = 1 \ldots k_{\max}+1$ divide the domain $[a,b]$ into a set of $k_{\max}$ intervals $[a,b] = \cup_{1=0}^{k_{\max}}[t_i, t_{i+1}]$, where $t_0 = \mathrm{a}, t_{k_{\max}+1} = b$.

Then, the set of $r+k_{\max}$ orthogonal splines can be constructed. The first $r+1$ splines are defined as the Legendre polynomials up to order r:

$$S_m(t) = P_m(t), m = 0 \ldots r. \tag{8}$$

The next splines are constructed on subsets $T_k = \{t_{i,k}\}_{i=-r}^{r+k+1}, k = 1 \ldots + k_{\max} - 1$ of the knot sequence T, where $t_{i,k}$ are defined as follows

$$t_{i,k} = \begin{cases} a, & -r \le i \le 0 \\ t_i, & 1 \le i \le k \\ b, & k+1 \le i \le k+r+1 \end{cases}. \tag{9}$$

For example, first and second subsets are $T_1 = \left\{\underbrace{a,a,\ldots,a}_{r+1} < t_1 < \underbrace{b,b,\ldots,b}_{r+1}\right\}$ and $T_2 = \left\{\underbrace{a,a,\ldots,a}_{r+1} < t_1 \le t_2 < \underbrace{b,b,\ldots,b}_{r+1}\right\}$, respectively.

Let $B_{i,r,k}(t)$ be a B-spline of order r on the knot sequence T_k

$$B_{i,r,k}(t) = \frac{t - t_{i,k}}{t_{i+r-1,k} - t_{i,k}} B_{i,r-1,k}(t) + \frac{t_{i+r,k} - t}{t_{i+r,k} - t_{i+1,k}} B_{i+1,r-1,k}(t), \tag{10}$$

where the zero-order B-spline

$$B_{i,0,k} = \begin{cases} 1, & t_{i,k} \le t \le t_{i+1,k} \\ 0, & otherwise \end{cases}. \tag{11}$$

Then, the remaining Legendre-like splines $S_m(t), m = r+2 \ldots r + k_{\max}$ are determined by

$$S_{r+k}(t) = \frac{d^{r+1}}{dt^{r+1}} f_k(t), k = 1 \ldots k_{\max}, \tag{12}$$

where $f_k(t)$ is the determinant of the matrix

$$f_k(t) = \det \begin{pmatrix} B_{-r,2r+1,k}(t) & B_{-r+1,2r+1,k}(t) & \cdots & B_{-r+k-1,2r+1,k}(t) \\ B_{-r,2r+1,k}(t_1) & B_{-r+1,2r+1,k}(t_1) & \vdots & B_{-r+k-1,2r+1,k}(t_1) \\ \vdots & \vdots & \ddots & \vdots \\ B_{-r,2r+1,k}(t_{k-1}) & B_{-r+1,2r+1,k}(t_{k-1}) & \cdots & B_{-r+k-1,2r+1,k}(t_{k-1}) \end{pmatrix}. \quad (13)$$

Wei et al. (2013) demonstrated the construction of orthogonal splines on a simple example with $r = 3$ and the knot sequence $T = [0,0,0,0,1,2,3,3.5,4,4,4,4]$. The full set of splines $S_m(t), m = 0 \ldots r + k_{\max}$ is presented in Figures 4 and 5.

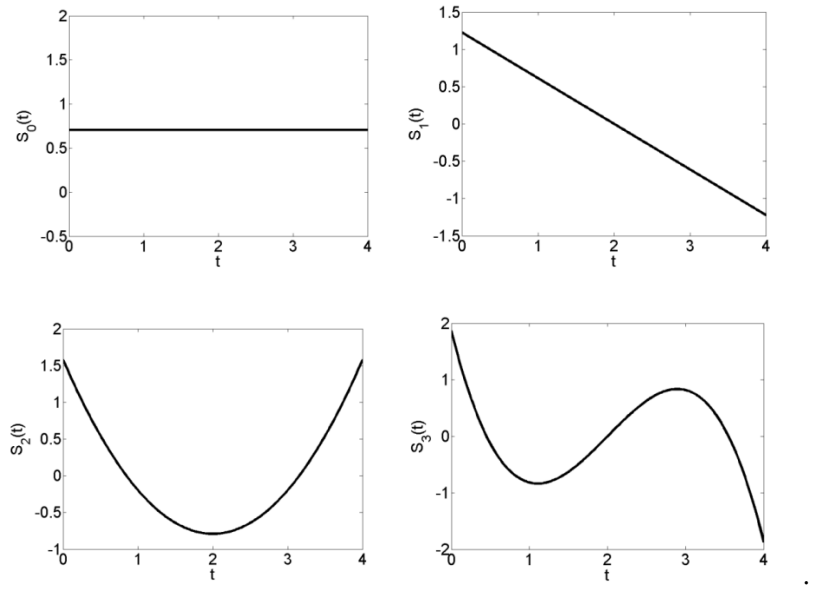


Figure 4. First four Legendre-like splines over the knot sequence T.

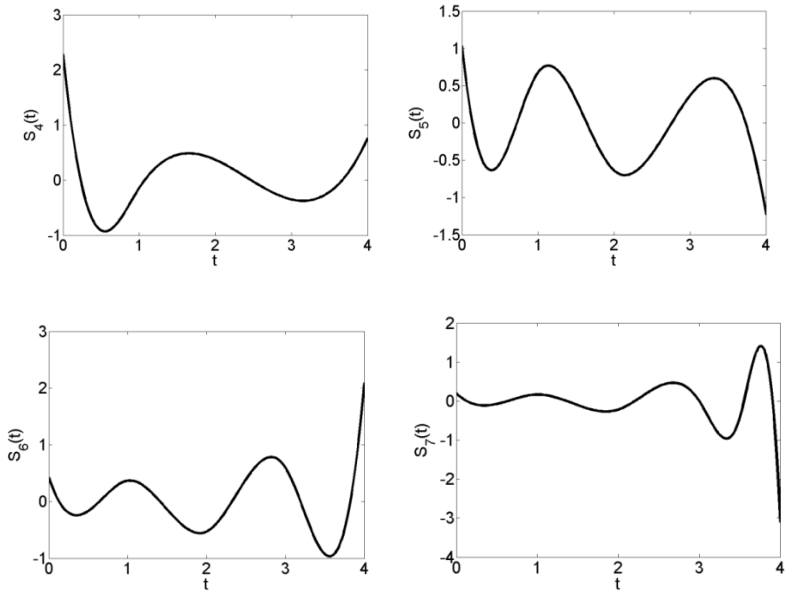


Figure 5. Last four Legendre-like splines over the knot sequence T.

3 Data-driven simulation

Up until this point, only statistics from TI are used for high-order simulation and no information about the spatial relationship from hard data has been taken into account. However, in practice, it is difficult to obtain a reliable TI, which is consistent with spatial statistics of the hard data. Furthermore, the TI is a subjective matter and two geologists can provide training images with fairly different statistics. Therefore, hard data should be the main source of spatial statistics and TI should only be the complimentary one. Ideally, all information available in the hard data should be taken into account and then, if required, completed from a training image.

3.1 Data-driven algorithm

Mustapha and Dimitrakopoulos (2010; 2011) proposed to calculate high-order spatial statistics based on founded replicates in hard data and TI. For a given cumulant, the calculation is performed using replicates found in the hard data. If the number of replicates is less than a user defined threshold, the search continues through the TI. Nevertheless, in practice it is difficult to find replicates in hard data even for the third-order cumulant. Moreover, all the orders of spatial statistics are connected by condition of non-negative range of pdf. For example, the value of the second-order statistics imposes restrictions on values of higher-orders. Therefore, generally speaking, combining low-order cumulants calculated from data with high-order cumulants derived from the TI can result in pdf with negative values.

In this paper, a straightforward data-driven algorithm is proposed (Algorithm A.2).

Algorithm A.2.

1. Run steps 1,2,3a-b of Algorithm A.1.
2. For a given template $T_{n+1}^{e_1,e_2,\dots e_n}(h_1,h_2,\dots,h_n)$ of unsampled location x_0 for each hard data location x_m
   a. Define a replicate at location x_m

$$\begin{aligned} &R_m^0 = Z(x_m) \\ &R_m^i = Z(x_m + h_i), i = 1 \dots n \\ &\{x_m, x_m + h_1, x_m + h_n\} \in T_{n+1}^{e_1,e_2,\dots e_n}(h_1,h_2,\dots,h_n) \end{aligned} , \tag{14}$$

   If $Z(x_m + h_i)$ is unsampled consider R_m^i is undefined, otherwise R_m^i is defined.
   b. Define a random path visiting all undefined values of replicate R_m^i
   c. Run step 3 of Algorithm A.1 to complete the replicate R_m^i.
   d. Add the completed replicate to the set of replicates $\{R_m^i\}_{m=1\dots M}^{i=1\dots n}$
3. Use the set of completed replicates $\{R_m^i\}_{m=1\dots M}^{i=1\dots n}$ to calculate the coefficients $L_{i_0,i_1,\dots i_n}$ at the unsampled location x_0 using expression (3).
4. Run steps 3e-g, 4 of Algortihm A.1.

In this way, spatial statistics are implicitly calculated from hard data and the TI, therefore the higher-order statistics are consistent with each other.

To better analyse the influence of hard data and the TI, let us consider the 2-neighbors spatial relationships, i.e. $n = 2$. Let $\{R_m^i\}_{m=1\dots M}^{i=0\dots 2}$ be a set of replicates after step 2a of Algortihm A.2 for an arbitrary unsampled location x_0. Let M_0, M_1, M_2 be numbers of replicates R_m^i with defined first value R_m^0, defined first R_m^0 and second values R_m^1, and with all values R_m^0, R_m^1, R_m^2 defined, respectively. It is not hard to see, that $M_0 \geq M_1 \geq M_2$ and $M_0 = M$ because $R_m^0 = Z(x_m)$, where x_m is sampled location. Then, the amount of information is taken from data and TI can be expressed in terms of M_i. For example, to estimate r-order moments $Mom_r\big(Z(x_0)\big) = E[Z(x_0)^r]$ of random variable at location x_0 only information from hard data is used, because $R_m^0 = Z(x_m)$ belong to hard data by the construction. In the estimation of the two-points relationships, i.e. r-order moment $Mom_r\big(Z(x_0), Z(x_1)\big) = E[Z(x_0)^p Z(x_1)^{r-p}], p = 1 \dots r$ of the random variables at locations x_0 and $x_1 = x_0 + h_1, M_1$ replicates are associated with hard data, and the rest $M - M_1$ are completed using the TI. For the three-points relationships, i.e. r-order moment $Mom_r\big(Z(x_0), Z(x_1), Z(x_2)\big) = E[Z(x_0)^p Z(x_1)^q Z(x_1)^{r-p-q}], p = 1 \dots r, q = 1 \dots r - p$ of the random variables at locations x_0, x_1 and $x_2 = x_0 + h_2$, only M_2 replicates use information from hard data and the rest is taken from the TI.

Thus, the influence of the TI is greater for multi-point relationships, whereas hard data's contribution is dominated for one- and two-points correlations.

It should be noted, that more hard data is available, more complex multi-point relationships are taken from hard data during the simulation.

4 Case study

4.1 The image of fractures

Simulation of continuous variables is tested on the case study, which is based on the real image of a fracture network (Figure 6). Grey-scale values of image in Figure 6 are transformed to the $[0,1]$ domain. The left half of Figure 6 is used as a reference image, see Figure 7a, and the right one is used as a training image, Figure 7b. The simulation grid consists of 100×100 nodes and 5% of nodes (500 points in Figure 7c) are randomly sampled from the reference image and used as hard data.

Figure 6. Image of real fracture network (public dataset).

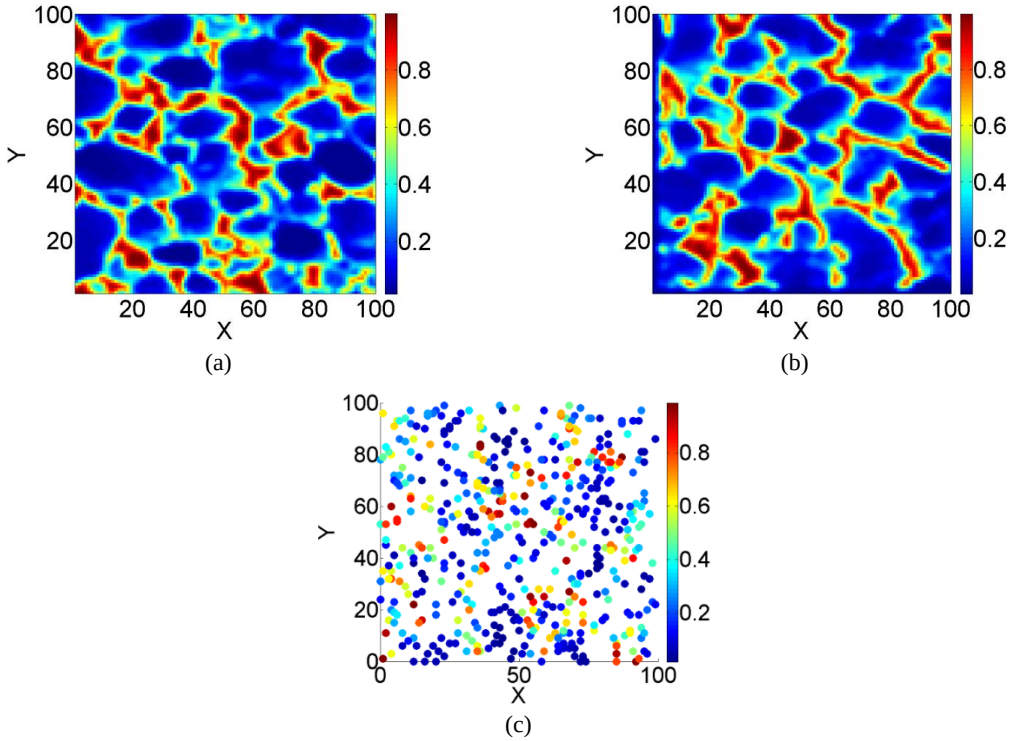


Figure 7. Fracture network case study: (a) reference image; (b) training image; (c) 500 samples from reference image.

The simulations using three different algorithms are compared in terms of spatial statistics: (a) the TI-driven Algorithm A.1 with Legendre polynomials of order 20, (b) the TI-driven Algorithm A.1 with Legendre-like splines of order 3, and (c) the data-driven Algorithm A.2 with Legendre-like splines of order 3.

4.1.1 Testing

The simulations of the case study are shown in Figure 8. The simulation using Legendre polynomials (Figure 8a) has numerical noise due to limitations discussed in paragraph 2.1, whereas simulations using the spline approximation (Figure 8b and 8c) show a stable reproduction of complex geometrical features.

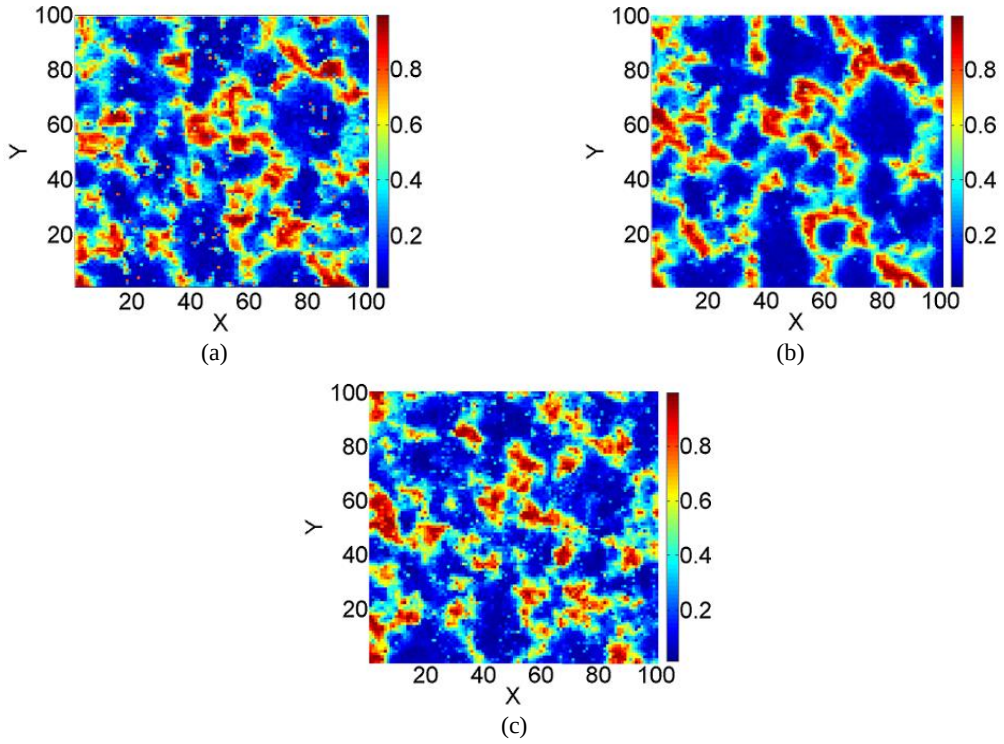


Figure 8. Simulation results with (a) the TI-driven simulation using polynomials; (b) the TI-driven simulation using splines; (c) the data-driven simulation using splines.

Variograms of the hard data, the TI, and three different simulations are similar (Figure 9). Note that the variogram of the simulation using Legendre polynomials has the biggest nugget effect (dashed line in Figure 9).

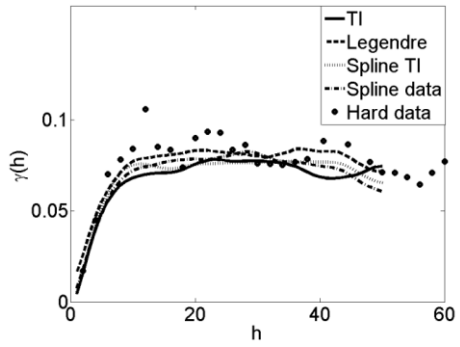


Figure 9. Variograms of hard data (dots), the TI (solid line), the TI-driven simulation using Legendre series (dashed line), the TI-driven simulation using splines (dotted line), the data-driven simulation using splines (dash-dotted line).

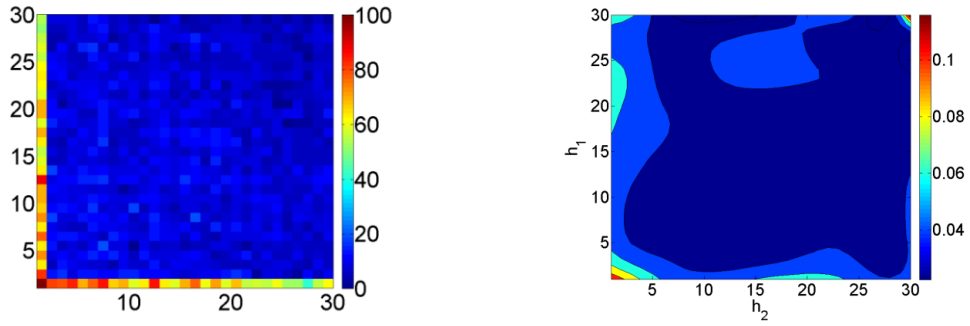


Figure 10. Estimation of third-order spatial moments of hard data: (a) number of replicates used for estimation, (b) estimation result.

The high-order spatial statistics are estimated based on a L-shape template with directional vectors $e_1 = \{1,0,0\}$ and $e_2 = \{0,1,0\}$. It is hard to analyse higher-order spatial cumulants of data samples due to the low number of replicates found in neighbourhoods (Figure 10a). However, along the X and Y axes, where the number of replicates is big enough and the third-order spatial cumulants of hard data (Figure 11a) are better reproduced in simulation using data-driven Algorithm A.2 (Figure 10d), than in the simulation using the TI-driven Algorithm A.1 (Figure 10 b and c).

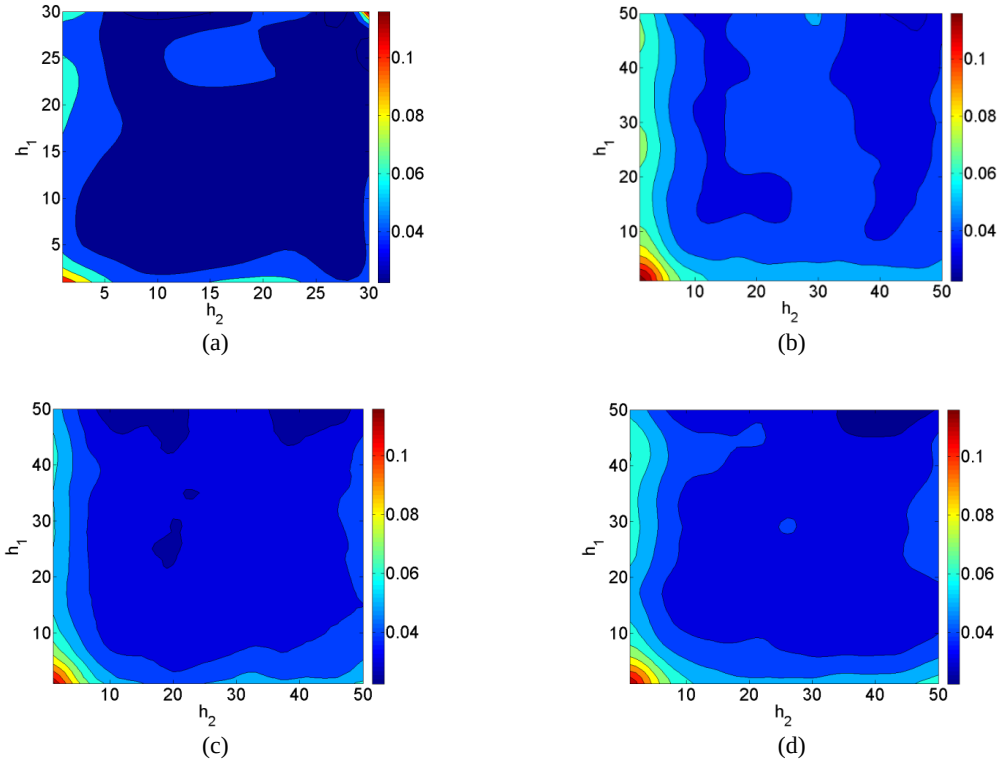
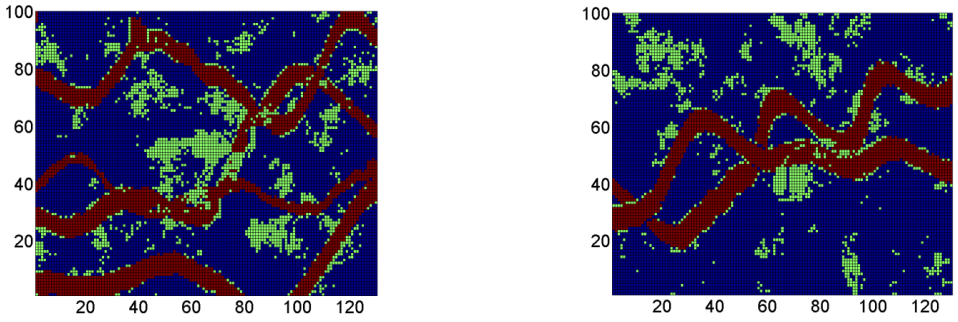


Figure 11. Third-order spatial moments of (a) hard data samples; (b) the TI-driven simulation using Legendre polynomials; (c) the TI-driven simulation using splines; (d) the data-driven simulation using splines.

4.2 Categorical variables

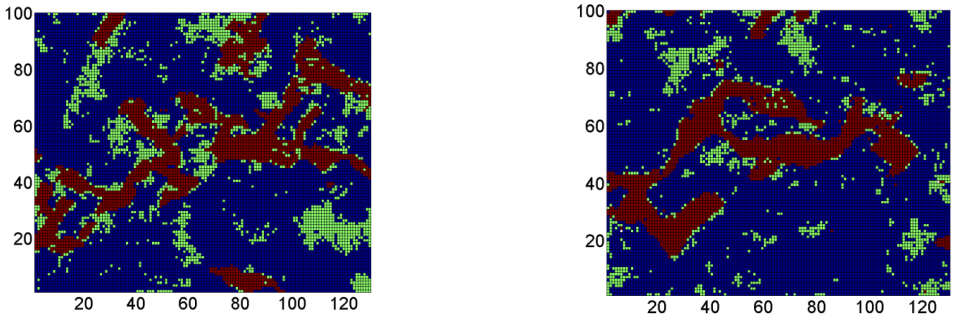
The proposed high-order simulation technique using Legendre-like splines allows also simulate categorical variables in the same framework by changing only the order of the splines to zero. For the sake of demonstration, data from the Stanford V reservoir case study (Mao and Journel, 1999) is used here. The training image and reference image are two different 2D-sections of the 3-D training image (Figure 12). 150 points are randomly sampled from the reference image and used as hard data.

The simulations using *snesim* (Strebelle, 2002) and the proposed are shown in Figure 12.



(a) (b)

Figure 12. Case study with categorical variables: (a) the reference image, (b) the training image, and (c) hard data. Different colours represent different categories.



(a) (b)

Figure 13. Simulation result for case study with categorical variables: (a) the simulation using snesim; (b) the simulation using the proposed approach. Different colours represent different categories.

5 Conclusions

This paper presents a new data-driven approach for high-order simulation of continuous and categorical variables based on Legendre-like orthogonal splines. Splines are flexible tools for the approximation of complex pdf. Using different knot sequences, orders of splines, and smoothness of piece-wise polynomials, it is possible to obtain a stable approximation with a good reproduction of spatial connectivity of the extreme values. The simulations are completely consistent with spatial statistics of hard data and share high-order spatial statistics of hard data and the TI. It is important to stress that the more information about high-order spatial statistics is available in the hard data, the less TI's statistics are used.

Additionally, the proposed approach provides a general framework for high-order simulation techniques. For example, by using just one interval for spline construction, the technique becomes the one proposed by Mustapha and Dimitrakopoulos (2010; 2011). Moreover, using splines of order 0 is obtained an implementation that is comparable to snesim (Strebelle, 2002). Furthermore, the technique can also be used for the simulation of multiple correlated continuous and discrete variables within a general framework.

Further research will address the adaptive knot sequence for better approximation of conditional pdf and the simulation of multiple correlated variables.

References

Boogaart, K.G., Tolosana-Delgado, R., Lehmann, M., Mueller, U. (2014). On the joint multi point simulation of discrete and continuous geometallurgical parameters. Orebody Modelling and Strategic Mine Planning 2014, Conference Proceedings, AusIMM.

Chatterjee, S., Mustapha, H., Dimitrakopoulos, R. (2015). Fast wavelet-based stochastic simulation using training images. Computational Geosciences.

Chuginova, T. and Hu, L. Y. (2008). Multiple-point simulations constrained by continuous auxiliary data. Mathematical Geosciences, 40 : 133–146.

Chiles, J. P. and Delfiner, P. (2012). Geostatistics, modelling spatial uncertainty (2nd ed). New York: Wiley.

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

De Iaco, S. and Maggio, S. (2011). Validation techniques for geological patterns simulations based on variogram and multiple-point statistics. Mathematical Geosciences, 43(4) : 483–500.

Dimitrakopoulos, R., Mustapha, H., Gloaguen, E. (2010). High-order statistics of spatial random fields: Exploring spatial cumulants for modelling complex, non-Gaussian and non-linear phenomena. Mathematical Geosciences, 42(1) : 65–97.

Goodfellow, R, Albor Consuegra, F, Dimitrakopoulos, R, Lloyd, T. (2012). Quantifying multi-element and volumetric uncertainty, Coleman McCreedy deposit, Ontario, Canada, Computers & Geosciences.

Goovaerts, P. (1997). Geostatistics for natural resources evaluation. Oxford University Press, Oxford.

Guardiano F.B., and Srivastava R.M. (1993). Multivariate geostatistics: Beyond bivariate moments. In Geostatistics Troia 1992, Proceedings of the Fourth International Geostatistics Congress vol. 1, (ed: A. Soares), Kluwer: Dordrecht, Netherlands, 133–144.

Journel, A.G. (2005). Beyond covariance: the advent of multiple-point geostatistics. In, Geostatistics Banff 2004. Springer, Netherlands, 225–233.

Journel, A.G. (2007). Roadblocks to the evaluation of ore reserves - the simulation overpass and putting more geology into numerical models of deposits. Orebody Modelling and Strategic Mine Planning - Uncertainty and Risk Management, The AusIMM, Spectrum Series 14, 2nd Ed., 29–32.

Honarkhah, M. (2011). Stochastic simulation of patterns using distance-based pattern modeling. Ph.D. thesis, Stanford University, Stanford, Ca.

Hoschek, J. and Lasser, D. (1993). Fundamentals of computer aided geometric design. AK Peters, London (UK).

Hughes, T. J. R., Cottrell, J.A., Bazilevs, Y. (2005). Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement computer methods. Applied Mechanics and Engineering, 194(39-41) : 4135–4195.

Kolbjørnsen, O., Stien, M., Kjønsberg, H., Fjellvoll, B., Abrahamsen, P. (2014). Using multiple grids in Markov mesh facies modeling. Mathematical Geosciences, 46(2) : 205–225.

Lebedev, N. N. (1965). Special Functions and their Applications. Prentice-Hall Inc., New York.

López de Silanes, M.C., Parra, M.C., Pasadas, M., Torrens, J.J (2001). Spline approximation of discontinuous multivariate functions from scattered data. Journal of Computational and Applied Mathematics, 131(1-2) : 281–298.

Malagù, M., E. Benvenuti, C.A., Duarte, C.A., Simone, A. (2014). One-dimensional nonlocal and gradient elasticity: Assessment of high order approximation schemes. Computer Methods in Applied Mechanics and Engineering, 275 (15) : 138–158.

Mao, S. and Journel A. (1999). Generation of a reference petrophysical and seismic 3D data set, The Stanford V reservoir, s.l. Stanford Center for Reservoir Forecasting Annual Meeting, SCRF Report, Stanford University.

Mariethoz, G. and Renard, P. (2010). Reconstruction of incomplete data sets or images using direct sampling. Mathematical Geosciences, 42 (3) : 245–268.

Mustapha, H. and Dimitrakopoulos, R. (2010). High-order stochastic simulations for complex non-Gaussian and non-linear geological patterns. Mathematical Geosciences, 42(5) : 455–473.

Mustapha, H. and Dimitrakopoulos, R. (2011). HOSIM: A high-order stochastic simulation algorithm for generating three-dimensional complex geological patterns. Comp. & Geosc., 37(9) : 1242–1253.

Osterholt, V. and Dimitrakopoulos, R. (2007). Simulation of wireframes and geometric features with multiple-point techniques: application at Yandi iron ore deposit. Orebody modelling and strategic mine planning, AusIMM, Spectrum Series 14, 2nd Ed., 95–124.

Park, H. and Lee, J.-H. (2007). B-spline curve fitting based on adaptive curve refinement using dominant points. Computer-Aided Design, 39(6) : 439–451.

Piegl, L. (1989). Modifying the shape of rational B-splines. Part 1: curves. Computer-Aided Design 21(8) : 509–518.

Sinha, S. and Schunck, B.G. (1992). A two-stage algorithm for discontinuity-preserving surface reconstruction. IEEE Transactions, Pattern Analysis and Machine Intelligence, 14(1) : 36–55.

Straubhaar, J., Renard, P., Mariethoz, G., Froidevaux R., Besson, O. (2011) An improved parallel multiple-point algorithm using a list approach. Mathematical Geosciences, 43(3) : 305–328.

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiplepoint statistics. Mathematical Geology, 34 : 1–22.

Strebelle, S. and Zhang T. (2005). Non-stationary multiple-point geostatistical models. Geostatistics Banff 2004, Springer, Part 1, 235–244.

Strebelle, S. and Cavelius, C. (2014). Solving speed and memory issues in multiple-point statistics simulation program SNESIM. Mathematical Geosciences, 46(2) : 171–186.

Tjelmeland, H. (2013). Construction of binary multi-grid Markov random field prior models from training images. Mathematical Geosciences, 45(4) : 383–409.

Wilson, GA. and Wragg, A. (1973). Numerical methods for approximating continuous probability density functions over $[0,\infty)$ using moments. IMA J. Appl. Math., 12 : 165–173.

Wei, Y., Wang, G., Yang, P. (2013). Legendre-like orthogonal basis for spline space. Computer-Aided Design, 45(2) : 85–92.

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