Les Cahiers du GERAD ISSN: 0711–2440

A new non-stationary high-order spatial sequential simulation method

A.A. Haji Abolhassani,
R. Dimitrakopoulos, F.P. Ferrie

G–2018–88

October 2018

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, 2018
– Bibliothèque et Archives Canada, 2018

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, 2018
– Library and Archives Canada, 2018

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 new non-stationary high-order spatial sequential simulation method

Amir Abbas Haji Abolhassani [a,b]

Roussos Dimitrakopoulos [a,b]

Frank P. Ferrie [b]

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

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

[c] *Department of Electrical and Computer Engineering, McGill University, Montréal (Québec) Canada, H3A 0E8*

amir.hajiabolhassani@mail.mcgill.ca
roussos.dimitrakopoulos@mcgill.ca
frank.ferrie@mcgill.ca

October 2018
Les Cahiers du GERAD
G–2018–88

Abstract: A new non-stationary, high-order sequential simulation method is presented herein, aiming to accommodate complex curvilinear patterns and high-order spatial connectivity when modeling non-Gaussian, spatially distributed and variant attributes of natural phenomena. The proposed approach employs spatial templates, training images and a sparse set of sample data similarly to other high-order and multiple point simulation methods. At each step of a multi-grid approach, a template consisting of several data points and a simulation node located in the center of the grid is selected. Sliding the template over the training image generates a set of training patterns, and for each pattern a weight is calculated. The weight value of each training pattern is determined by a similarity measure defined herein, that is calculated between the data event of the training pattern and that of the simulation pattern. This results in a non-stationary spatial distribution of the weight values for the training patterns. Each node is then simulated, conditioned on the training patterns with contribution factors proportional to the calculated non-stationary weights. The proposed new similarity measure is constructed from the high-order statistics of data-events from the available data set, when compared to their corresponding training patterns. In addition, this new high-order statistics measure allows for the effective detection of similar patterns in different orientations, as these high-order statistics conform to the commutativity property. The method is both fast and robust, designed to respect the available data and its spatial statistics. The proposed method is robust against the addition of more training images due to its non-stationary aspect; it only uses replicates from the pattern database with the most similar local high-order statistics to simulate each node. Examples demonstrate the key aspects of the method.

Keywords: High-order spatial statistics, sequential simulation, non-stationary, transformation invariant and multi-point statistics

1 Introduction

The uncertainty of a spatially distributed and variant geological attribute can be quantified by analyzing the variation of this attribute within a set of geostatistical or stochastic simulations. The second-order spatial simulation methods (Journel and Huijbregts 1978; David 1988; Goovaerts 1997; Chiles and Delfiner 2012) have been developed to quantify spatial uncertainty. Since the early 1990's, a new multiple point spatial simulation (MPS) framework has been developed to extend the limitations of previous approaches (Guardiano and Srivastava 1993; Strebelle 2002; Journel 2005; Liu et al. 2006; Arpat and Caers 2007; Hu and Chugunova 2008; de Vries et al. 2009; Mariethoz et al. 2010a,b; Honarkhah and Caers 2010; De Iaco and Maggio 2011; Stien and Kolbjørnsen 2011; Chatterjee et al. 2012; Lochbühler et al. 2013; Rezaee et al. 2013; Toftaker and Tjelmeland 2013; Mustapha et al. 2014; Strebelle and Cavelius 2014; Zhang et al. 2017). By design, MPS methods are able to capture and produce more complex spatial patterns than second-order spatial simulation methods by using multiple point spatial templates that replace the well established two-point spatial statistics. In the MPS framework, sample data are used as the conditioning data, while the multiple point spatial relations of the simulations are captured from patterns extracted from a training image (TI) or analogue. As a result, MPS simulations exhibit the statistics of the TI instead of the data (Osterholt and Dimitrakopoulos 2007; Goodfellow et al. 2012). High-order spatial simulation methods (HOSIM) have been developed to address this issue and to provide a new geostatistical simulation framework that deals with complex spatial patterns. HOSIM does not require distributional assumptions and is mathematically consistent (Dimitrakopoulos et al. 2010; Mustapha and Dimitrakopoulos 2010a,b, 2011a; Mustapha et al. 2011, 2014; Tamayo-Mas et al. 2016; Minniakhmetov and Dimitrakopoulos 2016, Minniakhmetov et al. 2018; Yao et al. 2018). It should be noted that in HOSIM methods, the conditional probability distribution function (CPDF) is modelled by a set of orthogonal polynomials, Legendre polynomials in particular (Abramowitz 1974). Under the stationarity assumption, high-order spatial cumulants (Dimitrakopoulos et al. 2010) are inferred by averaging the same-order cumulants of the samples extracted either from the dataset alone, if sufficient data is provided, or both data and TI in case of sparse data. The high-order statistics enable the CPDF model to capture the complex spatial structures and connectivity of the high values within the data and to later reproduce them via the simulation process (Mustapha and Dimitrakopoulos 2011b). Stationarity is a major assumption made by the commonly employed spatial simulation methods and may not always be easily accommodated in applications. With respect to high-order simulations, the same assumption also results in the use of all patterns extracted from data and the TI for the estimation of spatial cumulants, leading to the slow convergence of the orthogonal polynomials used for modelling CPDFs. In practice, this calculation may result in numerical instabilities (Boyd and Ong 2011). Thus, a non-stationary approach is used herein to address these issues. For each simulation, a set of patterns are chosen from the data and TI based on the similarity of their data-event to the simulation pattern. Then, only the chosen patterns are used for inferring the high-order statistics of the simulation pattern. Similar patterns could be generated from a simple CPDF model and faster convergence using a lower number of polynomial terms.

With respect to numerical stability, in the non-stationary approach detailed herein, a non-stationary spatial filter is first estimated; then, for example, the second-order cumulant is inferred by weighted averaging over samples: $cum(z_0, z_1) \approx \frac{1}{N_x} \sum_{x \in [0,1]} z_0(x) z_1(x)$. The joint PDF $f(z_0, z_1)$, shown in Figure 1 c, d, g and h, and CPDF $f(z_0|z_1 = 0.5)$, illustrated in Figure 1 e, f, i and j, are then estimated using a non-stationary method explained in Section 2 with 10, 20, 30 and 40 Legendre terms, respectively. The example demonstrates that the non-stationary approach selects more similar samples to produce stable and accurate spatial cumulants while avoiding numerical instabilities, a characteristic that extends to both higher dimensions and orders. The high-order sequential simulation method introduced here, first finds a data-event of a fixed size N in the immediate neighbourhood of each simulation node, and a set of N +1 spatial high-order statistics is then calculated from the data-event. A novel similarity measure is then introduced and utilized to compare the extracted statistics from the simulation grid to those extracted from all of the patterns in the training image; the statistics with the highest similarity score are then used for the inference of the high-order statistics at that node. This is the non-stationary aspect of the method, which is the key difference between this method and previous high-order simulation methods (Mustapha and Dimitrakopoulos 2010b, 2011b; Minniakhmetov and Dimitrakopoulos 2018; Yao et al. 2018). In addition to the data-driven property, which is inherited from HOSIM, the proposed method does not become overwhelmed by an excessive number of

patterns extracted from the TI due to the selectivity of the method in terms of choosing the number of related patterns. In addition, the method is able to identify patterns with similar statistics in different orientations efficiently and without reducing the computational efficiency of the algorithm.

In the following sections, the proposed method and corresponding algorithm are first detailed. Then, examples and comparisons are presented, and conclusions follow.

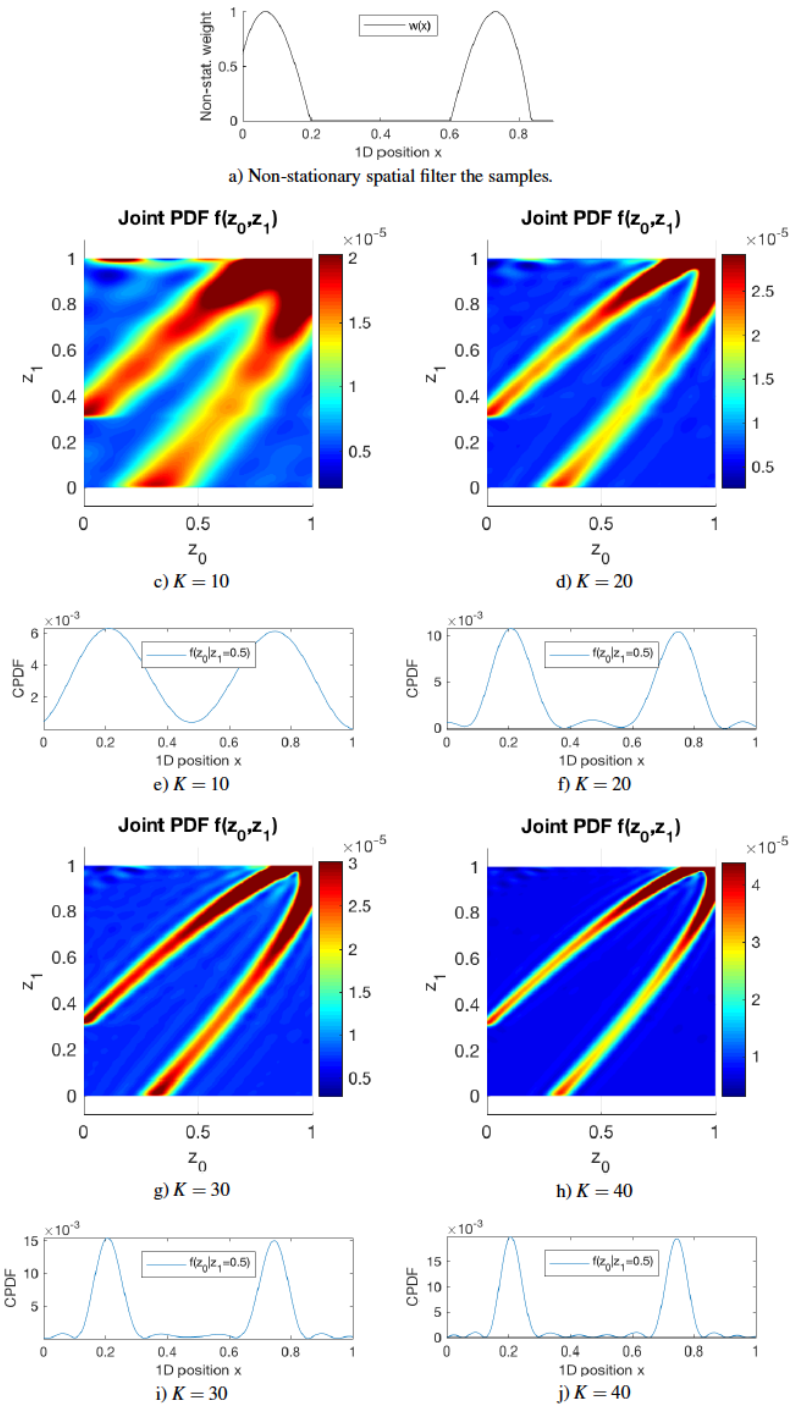

Figure 1: Second order non-stationary HOSIM in one dimension, the non-stationary weight is used to filter the samples to provide a more accurate estimation for the statistics from samples. Results are the conditional probability distribution function, at $z_1 = 0.5$, and the joint probability distribution function with $K \in \{10, 20, 30, 40\}$ number of Legendre terms.

2 The method

2.1 Spatial random field

Given a probability space $(\Omega, \mathcal{F}, P)$, a function $Z : \Omega^{n+1} \rightarrow R^{n+1}$ is a real-valued spatial random field. If, for any $r \in R^{n+1}$, a subset $A_r = \{X \in \Omega^{n+1} | Z(X) \leq r\} \in \mathcal{F}$ exists, where $\{\Omega \subset R^{\mathcal{N}} | \mathcal{N} = 2, 3\}$ is a two- or three-dimensional spatial domain, X is a set of $n+1$ points in space and $X = \{x, x_1, \ldots, x_n\}$ and $Z(X) = \{z(x), z(z_1), \ldots, z(x_n)\}$. Also, the binary operation $\leq$ is true only if all elements on the left-hand side are less than or equal to the elements on the right hand side of the equation.

2.2 Template, pattern, data-event and neighbourhood

Consider a random field $\{Z(X) | X = \{x, x + h_1, \ldots, x + h_n\} \in \Omega^{n+1}\}$ with a fixed structure in space and a template that is characterized by a set of lag vectors (Mustapha and Dimitrakopoulos 2011b), i.e. $T = \{h_1, \ldots, h_n\} \in \Omega^n$. The template connects x to each of the positions in the neighbourhood $N_x = x + T = \{x_1, \ldots, x_n\}$. Without loss of generality, hereafter the real-valued spatial random field is denoted by $Z(x, N_x)$, and its realization is denoted by $Z(x_i, N_{x_i})$. The pattern, at position $x_i \in \Omega$ and with n neighbourhood positions in space, is described by the equation $N_x = x_i + T = \{x_i + h_1, \ldots, x_i + h_n\}$ using a fixed template T. Furthermore, the data-event vector $d_{N_x} = Z(N_x) = [Z(x_1), \ldots, Z(x_n)]^T$ represents the realization of the field at the position of the neighbourhood nodes.

2.3 Modelling the spatial non-stationary joint cumulative probability distribution function (CPDF) as a finite sum of disjoint CPDFs

Given a measurable space $(R^{n+1}, \beta(R^{n+1}))$, the real-valued spatial random field $Z(x, N_x)$ is fully characterized by its non-stationary joint cumulative probability distribution function $F(x, Z(x, N_x))$. In this work, this function is modelled by the statistical ensemble (Landau and Lifshitz 1980) given by

$$F(x_0, Z(x, N_x)) = \sum_{i=1}^{m} \phi_i(x_0) F_i(Z(x, N_x)), \tag{1}$$

which is a sum over m mutually exclusive stationary CPDFs, F_i's, with a set of non-stationary mixing coefficients, known as a state set with binary variables $\phi_i(x_0) \in \{0, 1\}$. At any position $x_0 \in \Omega$, the value of only one of the coefficients is one, and the rest are zero, i.e. $\phi_k = 1$ and $\phi_i = 0$ for $i \neq k$.

2.4 Mutually exclusive CPDFs

On the grid of an input image with a fixed template T, (1) is considered to have a mutually exclusive set of stationary CPDFs only if it satisfies the condition

$$\begin{aligned} &\exists \epsilon \in R | \forall x_1, x_2 \in \Omega \\ &\text{if } |D(d_{N_{x_1}}, d_{N_{x_2}})| < \epsilon \\ &\Rightarrow F(x_1, Z(x, N_x)) = F(x_2, Z(x, N_x)) = F_k(Z(x, N_x)) \\ &\Rightarrow \text{The high} - \text{order statistics of two patterns match} \\ &(U_{Z(x_1, N_{x_1})} \approx U_{Z(x_2, N_{x_2})}), \end{aligned} \tag{2}$$

where D(.) is the high-order-statistics distance vector calculated from (6), introduced in Section 2.6, to compare two sets of high-order statistics vectors of two data-events $d_{N_{x_1}}$ and $d_{N_{x_2}}$. In addition, $U_{Z(x_1, N_{x_1})}$ and $U_{Z(x_2, N_{x_2})}$ are the high-order statistics of two patterns, introduced in Section 2.5, with elements calculated from relation (4). Equation (2) states that, if the statistics of the data-event of two patterns are close, then, their CPDFs are equal to an identical stationary CPDF defined as $F_k(Z(x, N_x))$, and, further, that their high-order statistics are the same. This also implies that $\phi_k(x_1) = \phi_k(x_2) = 1$ and $\phi_k(x_1) = \phi_k(x_2) = 1$.

A set of mutually exclusive CPDFs are used as the basis functions to span a more complex non-stationary CPDF using (1). At each position on the grid, only one basis function is responsible for the simulation of

the node, i.e. each basis function $F_k(Z(x, N_x))$ simulates a set of nodes on the grid Ω_k in Equation (1). The Ω_ks are mutually disjoint, and their union is the entire grid $\Omega = \cup_{k=1}^{m}\Omega_k$. As a result, for each simulation node, the entire TI is searched for nodes with similar statistics. Considering (2), the resulting set of nodes are generated from one of the basis CPDFs and used for estimating the high-order statistics of that basis CPDF, and the simulation node value is then sampled from the CPDF.

2.5 Invariant high-order statistics

Vieta's formula (Funkhouser 1930) presents a relation between the coefficients and the roots of a polynomial. For a polynomial of degree n,

$$p(X) = X^n + c_1u_1X^{n-1} + \cdots + c_nu_n, \tag{3}$$

with the roots $[x_1, \ldots, x_n]^T$ the coefficients are $U = [u_1, \ldots, u_n]^T$ where,

$$u_m = \frac{1}{c_m} \sum_{(1\leq i_1\leq n-m+1)} \sum_{(i_i<i_2\leq n-m+2)} \cdots \sum_{(i_{m-1}<i_m\leq n)} x_{i_1} \ldots x_{i_m} \tag{4}$$

with $m \in \{1, \ldots, n\}$ and the normalization factor $c_m = \binom{n}{m}\sigma^m$, where σ is the standard deviation of the x in the TI. Furthermore, a recurrence relation could present the coefficients of a polynomial $q(X) = (X - x_0)p(X) = X^{n+1} + c_1\overline{u}_1X^n + \cdots + c_{n+1}\overline{u}_{n+1}$, which has one extra root x_0, i.e. $\{x_0, x_1, \ldots, x_n\}$. Then, it follows that

$$\overline{u}_m = \frac{c_{m-1}}{c_m}x_0 + u_m \qquad\qquad m \in 1, \ldots, n+1\ , c_0 = 0, u_{n+1} = 0. \tag{5}$$

Vieta's formula transforms the input vector of the roots of the polynomial $[x_1, \ldots, x_n]^T$ into the vector of the coefficients $[u_1, \ldots, u_n]^T$, with the advantage of being invariant to the ordering of the domain, given that the polynomial is invariant under a re-ordering of its roots. Herein, (4) is used to transform either the patterns or the data-events of the simulation and TI, such that

$$\begin{aligned} Z(x_s, N_{x_s}) &\rightarrow U_{Z(x_s,N_{x_s})}, (\text{pattern simulation grid}) \\ Z(x_t, N_{x_t}) &\rightarrow U_{Z(x_s,N_{x_s})}, (\text{pattern TI}) \\ d_{N_{x_s}} &\rightarrow U_{d_{N_{x_s}}}, (\text{data} - \text{event simulation grid}) \\ d_{N_{x_t}} &\rightarrow U_{d_{N_{x_t}}}.(\text{data} - \text{event TI}). \end{aligned} \tag{6}$$

2.6 High-order statistics distance vector

Given two sets of data events $d_{N_{x_s}}$ and $d_{N_{x_t}}$ of order n, in order to develop an L2-norm distance measure, one must compare all possible ordering of these two data events, which results in $(n \times n)!$ number of operations. Hence, for a pattern of order $n = 5$, the number of operations become 600 per pattern, which is computationally expensive and can only operate on small size TIs. In a new approach (Abolhassani et al. 2017), the L2-norm distance is calculated for the high-order statistics vectors of the data-events

$$\begin{aligned} D\left(d_{N_{x_s}}, d_{N_{x_t}}\right) &= U_{d_{N_{x_s}}} - U_{d_{N_{x_t}}} \\ D(Z(x_s, N_{x_s}), Z(x_t, N_{x_t})) &= U_{Z(x_s,N_{x_s})} - U_{Z(x_t,N_{x_t})}, \end{aligned} \tag{7}$$

where the high-order statistics of the data-event Us are calculated from (4). It is worth mentioning that the number of operations for this new distance measure is reduced dramatically to $n\times2^{n-1}$ per simulation node versus $(n \times n)!$ in L2-norm. For $n = 5$, $\#op(L2 - norm) = 600$ *vs.* $\#op(high - ord) = 80$.

2.7 High-order transformation invariant sequential simulation

The goal is to simulate the non-stationary, real-valued random field $\{Z(x_s, N_{x_s})|x_s \in \Omega_s\}$ at the position of all simulation nodes Ω_s with a fixed template T. A sequential multi-grid process is used when provided with

a sparse set of data $\{Z(x_h, N_{x_h})|x_h \in \Omega_h\}$ on a regular grid Ω_h and a fine resolution training image (TI) $\{Z(x_t, N_{x_t})|x_t \in \Omega_t\}$. The order of the sequential simulation is shown in Figure 2. In the hierarchy of the sequential multi-grid simulation, the connecting lines represent the conditioning dependencies: the sample data nodes are shown in blue; red nodes are simulated conditioned on their 4 nearest neighbours; black nodes are simulated conditioned on both blue and red nodes from the previous sequence; orange nodes are simulated conditioned on previous simulated nodes and available data, and green nodes are simulated in the last sequence. The resolution of the grid doubles at each sequence. Unicolour lines in each figure represent the spatial templates, connecting the conditioning data and a single node to be simulated. The physical size of the templates reduces after each sequence to maintain the same number of conditioning nodes. This is the natural representation of the multi-grid approach for the simulation.

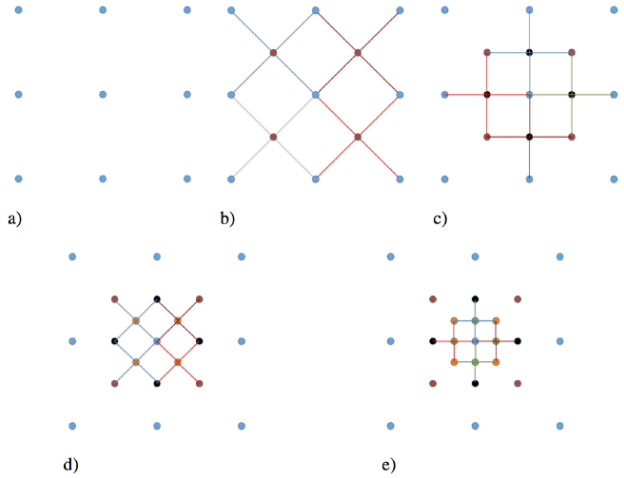

Figure 2: The hierarchy of the sequential-multi-grid simulation. The order of the simulation starting from the blue nodes (sample data) and the red nodes (simulated first), followed by black nodes, orange nodes and eventually the green nodes. The simulation of each stage is conditioned on the previous simulated nodes.

At each sequence, the path is chosen randomly and saved into a vector containing the indices of the visiting nodes I_s. Each successive random variable $Z(x_s, N_{x_s})$ at node $\{x_s|s \in I_s\}$ is conditioned to a data event $d_{N_{x_s}}$ and the neighbourhood nodes that are selected from the set of previously simulated nodes and the sample data (Goovaerts 1997). A template $T = \{h_1, \ldots, h_1\}$ is formed spatially by connecting each data-event node to the simulation node. The first goal is to simulate the high-order statistics of the pattern, i.e. $U_{Z(x_s,N_{x_s})}$, and then sample the simulation node $Z(x_s, N_{x_s})$, given the conditional probability distribution function $P(U_{Z(x_s,N_{x_s})}|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})})$. In this case, the $U_{(.)}$ function refers to the high-order statistics vector with the elements calculated from (4). This probability is complex and cannot be simulated directly unless it is represented by a model with a set of parameters $\theta \in \Theta$ that are independent from the data event and TI. The parameter θ is optimized to express the data event and TI. Thus, this probability can be decomposed using the Bayes rule into

$$P(U_{Z(x_s,N_{x_s})}|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})}) = \int_{\theta\in\Theta} P(U_{Z(x_s,N_{x_s})}|\theta)P(\theta|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})})d\theta. \tag{8}$$

An estimation of the argument of the integral further simplifies (8). Figure 3 represents the term $P(\theta|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})})$ as a function of θ. The contribution of this function is negligible except for a narrow band near an optimal value for the parameters maximum likelihood estimate θ_{MLE}. Consequently, (8) can be estimated as

$$\begin{aligned} P(U_{Z(x_s,N_{x_s})}|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})}) &\approx P(U_{Z(x_s,N_{x_s})}|\theta_{MLE}) \overbrace{P(\theta_{MLE}|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})})\Delta\theta}^{constant} \\ &= \frac{1}{C_{MLE}} P(U_{Z(x_s,N_{x_s})}|\theta_{MLE}). \end{aligned} \tag{9}$$

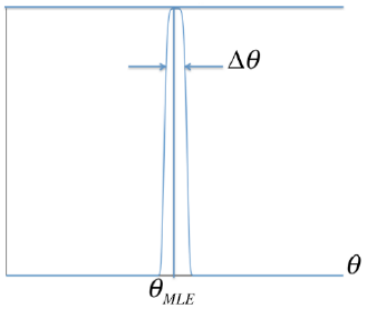

Figure 3: The representation of $P(\theta|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})})$ as a function of θ. In practice, this probability is negligible except at a narrow $\Delta\theta$ band near θ_{MLE}.

In Equation (9), C_{MLE} is the normalization factor to assure the validity of the probability $C_{MLE} = \sum_{u\in\mathcal{U}} P(u|\theta_{MLE})du$. Equation (9) implies that the probability distribution function of the node x_s can be calculated if θ_{MLE} is known. To estimate θ_{MLE} based on (Figure 3) $\theta = \theta_{MLE}$ when $P(\theta|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})})$ is maximized for $\theta \in \Theta$, i.e.

$$\theta_{MLE} = \underset{\theta\in\Theta}{argmax}\, P\left(\theta \,\middle|\, U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})}\right). \tag{10}$$

Using Bayes rule, the right-hand side of the equation is extended into

$$\theta_{MLE} = \underset{\theta\in\Theta}{argmax} \frac{P(U_{Z(\Omega_t,N_{\Omega_t})}|\theta, U_{d_{N_{x_s}}})\overbrace{P(\theta)}^{\text{uniform prior}}}{\underbrace{P(U_{Z(\Omega_t,N_{\Omega_t})})}_{\text{independent from }\theta}}. \tag{11}$$

In this case, one assumes a uniform prior in the parameter space Θ, with the marginal probability in TI $P(U_{Z(\Omega_t,N_{\Omega_t})})$ remaining independent from θ. Hence, (11) simplifies into

$$\theta_{MLE} = \underset{\theta\in\Theta}{argmax}\, P(U_{Z(\Omega_t,N_{\Omega_t})}|\theta, U_{d_{N_{x_s}}}). \tag{12}$$

Each node in TI is only conditioned on its neighbours N_{x_t} and the parameter set θ. Hence, the joint distribution in (12) can be decomposed further into

$$\theta_{MLE} = \underset{\theta\in\Theta}{argmax} \prod_{x_t\in\Omega_t} P(U_{Z(x_t,N_{x_t})}|\theta, U_{d_{N_{x_s}}}, U_{d_{N_{x_t}}}). \tag{13}$$

A probability is maximized if the logarithm of that probability is maximized:

$$\theta_{MLE} = \underset{\theta\in\Theta}{argmax} \sum_{x_t\in\Omega_t} \log P(U_{Z(x_t,N_{x_t})}|\theta, U_{d_{N_{x_s}}}, U_{d_{N_{x_t}}}). \tag{14}$$

Furthermore, at the maximum, the derivative with respect to θ should be zero, i.e.

$$\frac{\partial}{\partial\theta}\Big(\sum_{x_t\in\Omega_t} \log P(U_{Z(x_t,N_{x_t})}|\theta, U_{d_{N_{x_s}}}, U_{d_{N_{x_t}}})\Big) = 0. \tag{15}$$

Equation (15) is solved for an optimum solution θ_{MLE} by modeling the $P(U_{Z(x_t,N_{x_t})}|\theta, U_{d_{N_{x_s}}}, U_{d_{N_{x_t}}})$ with an exponential family in Section 2.8.

2.8 High-order simulation model

The exponential family is used here to model the likelihood function in (15) with the parameter set $\theta = [\theta_1, \ldots, \theta_{n+1}]^T$. In particular,

$$P(U_{Z(x_t,N_{x_t})}|\theta, U_{d_{N_{x_s}}}, U_{d_{N_{x_t}}}) = \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\omega(d_{N_{x_s}}, d_{N_{x_t}}, {\sigma_0}^2)(U_{Z(x_t,N_{x_t})} - \theta)^T(U_{Z(x_t,N_{x_t})} - \theta)\Big) \tag{16}$$

is a probability since

$$\int_0^1 \int_{-\infty}^{\infty} P(U_{Z(x_t,N_{x_t})}|\theta, U_{d_{N_{x_s}}}, U_{d_{N_{x_t}}})dud\omega = 1. \tag{17}$$

In (16), $\omega(d_{N_{x_s}}, d_{N_{x_t}}, \sigma_0^2)$ is introduced as the similarity measure of the data-event $d_{N_{x_s}}$ and $d_{N_{x_t}}$. It ensures that the TI patterns with more similar data-events contribute more to the likelihood functionn. The similarity measure is defined by

$$\omega(d_{N_{x_s}}, d_{N_{x_t}}, {\sigma_0}^2) = \exp(-\frac{D(d_{N_{x_s}}, d_{N_{x_t}})^T D(d_{N_{x_s}}, d_{N_{x_t}})}{2\Sigma\left(d_{N_{x_s}}, {\sigma_0}^2\right)}), \tag{18}$$

where $D(.,.)$ is given by (7) as the high-order statistics distance vector, a distance measure between two data-events. $\Sigma(.,.)$ is the covariance matrix. Substituting (16) into (15), the solution of the θ_{MLE} is

$$\theta_{MLE} = \frac{\sum_{x\in\Omega_t} \omega(d_{N_{x_s}}, d_{N_{x_t}}, {\sigma_0}^2) U_{Z(x_t,N_{x_t})}}{\sum_{x\in\Omega_t} \omega(d_{N_{x_s}}, d_{N_{x_t}}, {\sigma_0}^2)}, \tag{19}$$

the non-stationary expected value of the vector of the high-order statistics of the simulation pattern. Having calculated the $\theta_{MLE} = [\theta_1, \ldots, \theta_{n+1}]$ in (19) and considering (9), one gets

$$P(U_{Z(x_s,N_{x_s})}|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})}) \approx \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(U_{Z(x_s,N_{x_s})} - \theta_{MLE})^T(U_{Z(x_s,N_{x_s})} - \theta_{MLE})). \tag{20}$$

Applying the recurrence relation (5), (20) is rewritten in terms of the simulation data-event $U_{d_{N_{x_s}}}$ and the simulation node $Z(x_s)$ (the extra root in Vieta's formula (5)), i.e.

$$\begin{aligned} P(U_{Z(x_s,N_{x_s})}|U_{d_{N_{x_s}}}, U_{Z(\Omega_t,N_{\Omega_t})}) &\approx \frac{1}{\sqrt{2\pi{\delta_{n+1}}^2}}\exp(-\frac{\left(Z(x_s) - \mu_{n+1}(\theta_{MLE})\right)^2}{2{\delta_{n+1}}^2}) \\ \frac{1}{{\delta_{n+1}}^2} &= \sum_{m=1}^{n+1}(\frac{u_{m-1}c_{m-1}}{c_m})^2 \\ \mu_{n+1}(\theta_{MLE}) &= {\delta_{n+1}}^2 \sum_{m=1}^{n+1} \frac{(u_m - \theta_m)u_{m-1}c_{m-1}}{c_m}, \end{aligned} \tag{21}$$

where u_m could be calculated with (4) and $c_m = \binom{n}{m}\sigma^m$, in which σ is the standard deviation of the x in the TI. Eventually, the simulation node $Z(x_s)$ could be sampled from the distribution in (21).

2.9 Algorithm

1. Define a random path $\overline{\Omega}_s$ to visit all positions on the simulation grid Ω_s.
2. Select the next un-sampled position on the simulation grid $x_s \in \overline{\Omega}_s$.
3. Search the conditioning data Ω_h for n immediate neighbourhood positions $N_{x_s} = \{x_1, \ldots, x_n\} \subset \Omega_h$ of the simulation node x_s.
4. Generate the data-event vector $d_{N_{x_s}} = Z(N_{x_s}) = [Z(x_1), \ldots, Z(x_n)]^T$ with the node values at the neighbourhood positions N_{x_s}.
5. Compute the high-order-statistics vector of the data-event $U_{d_{N_{x_s}}} = [u_1, \ldots, u_n]^T$ with elements u_m calculated from (4).
6. Generate template $T = \{h_1, \ldots, h_n\} = \{x_1 - x_s, \ldots, x_n - x_s\}$.
7. Search all positions in the TI and pick the positions of the patterns (Ω_t, N_{Ω_t}) matching the template T.
8. Select the next unvisited TI pattern $(x_t, N_{x_t}) \in (\Omega_t, N_{\Omega_t})$.
9. Generate the TI data-event $d_{N_{x_t}} = Z(N_{x_t}) = [Z(y_1), \ldots, Z(y_n)]^T$ according to the node values in the TI.

10. Calculate the high-order-statistics vector $U_{d_{N_{x_t}}} = [u_1, \ldots, u_n]^T$ of the data event $d_{N_{x_t}}$ with elements u_m calculated from (4).
11. Calculate the high-order distance $D(d_{N_{x_s}}, d_{N_{x_t}})$ in (6).
12. Calculate the similarity measure $\omega(d_{N_{x_s}}, d_{N_{x_t}}, {\sigma_0}^2)$ from (18).
13. Continue to Step 8 until all of the TI patterns are visited.
14. Calculate $\theta_{MLE} = [\theta_1, \ldots, \theta_{n+1}]^T$ using (19).
15. Draw a sample $\overline{Z}(x_s)$ randomly from the distribution function in (21).
16. Continue to Step 2 until all of the grid is simulated.

3 Examples and comparisons

Applied aspects of the simulation method presented in the previous section are examined in this section using two-dimensional horizontal layers from the Stanford V fluvial reservoir (Mao and Journel 1999). Selected layers represent exhaustive images that are then sampled to generate datasets subsequently used to generate simulations with the proposed new non-statationary high-order spatial transformation invariant simulation method, HOSTSIM. The statistics of the simulated realizations, namely, histogram, variogram map and third-order L-shape cumulant map are then compared to the ones calculated from the initial data. In addition, the simulations generated by the HOSTSIM method are compared to the ones generated by the well-known FILTERSIM multiple point simulation method (Zhang et al. 2006) through its implementation available in the public domain (Remy et al 2009).

Figure 4 shows two exhaustive images used for the simulations E_1 and E_2; each one is a horizontal section of the complete data with size of $100 \times 100m^2$. Two sets of data are then selected from each section E_1 and E_2, one sampled at 625 and the other at 156 locations, respectively, resulting in 4 different data sets: DS_1 (625 points from E_1), DS_2 (156 points from E_1), DS_3 (625 points from E_2) and DS_4 (156 points from E_2), as shown in Figure 5. The TI used for all simulations is a different section selected from the Stanford V fluvial reservoir and is shown in Figure 6.

In each example, one of the data sets in Figure 5 and the TI in Figure 6 are used as input for simulations and two realizations are generated by each method noted above. The exhaustive images, the input data and the realizations generated by HOSTSIM and FILTERSIM are illustrated in Figures 7 to 10 for data samples DS_1 through DS_4, along with their respective validation graphs. The validation graphs are generated for both the input data and the HOSTSIM simulation and include the histogram, variogram map and third-order L-shape spatial cumulant map. Note that the second-order variogram map of an image represents the variogram for each lag direction, i.e. $r \in [0, 65]$ and $\theta \in \{0, \frac{\pi}{30}, \ldots, \pi\}$, in a polar coordinate system. The third-order L-shape spatial cumulant of an image is generated by using an L-shape template, which represents two orthogonal lag vectors from a common point. For all possible pairs of horizontal and vertical lags, the template is shifted over the image to extract the third-order cumulant by averaging the values calculated for each pattern within the image (Mustapha and Dimitrakopoulos 2010b, 2011a).

3.1 Example 1

Figures 7 and 8 show the results generated from DS_1 and DS_3 with 625 data samples. There are low-contrast regions in the HOSTSIM realizations (light blue/green area seen toward the bottom of E_1 in Figure 7, and in the bottom-right and top-center of the E_2 realization in Figure 8) in both realizations. This behaviour is expected, as a denser set of data is available in these regions, and HOSTSIM could match similar patterns within the TI and simulate the corresponding nodes better. This is also the case in the high value areas on the simulated realization as well (red color regions, or channels). Although there are some cases where the connectivity of the channels are not well generated, e.g. in the top-right part of both realizations in Figure 8, the narrow channel is not well represented in the data, and the HOSTSIM could not estimate the high-order statistic of the data-event accurately to match appropriate patterns from either the data or the TI. The histogram and variogram graphs of the HOSTSIM simulations are similar to the ones calculated for

DS_1 and DS_3 in both Figures 7 and 8. The third-order L-shape spatial cumulants are well generated for the lags <50.

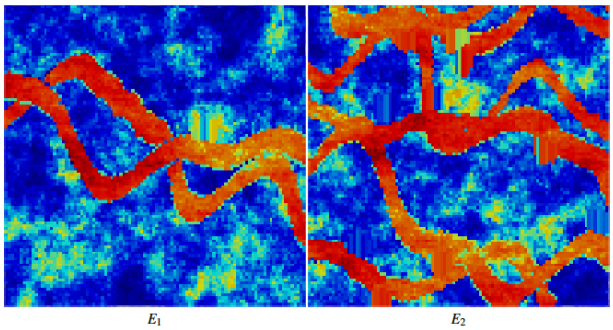

Figure 4: From left to right, the exhaustive images E_1 and E_2 (ground truth).

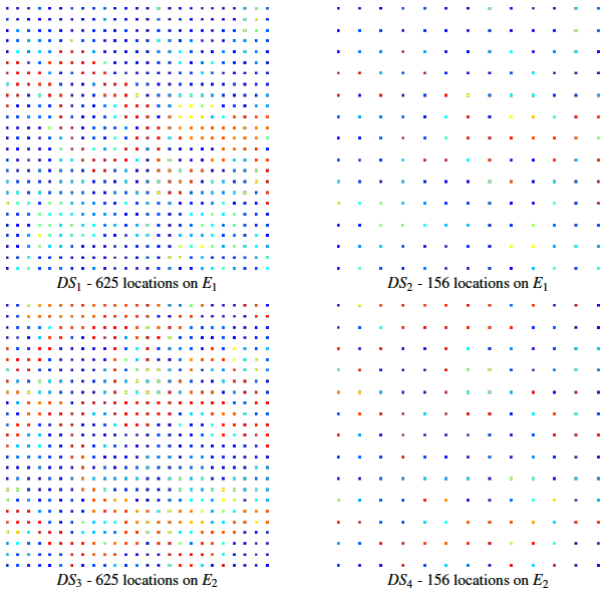

Figure 5: Data samples used for the simulations

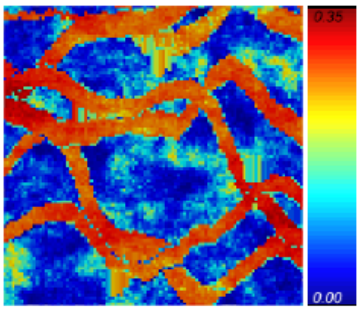

Figure 6: TI used in all four simulation cases

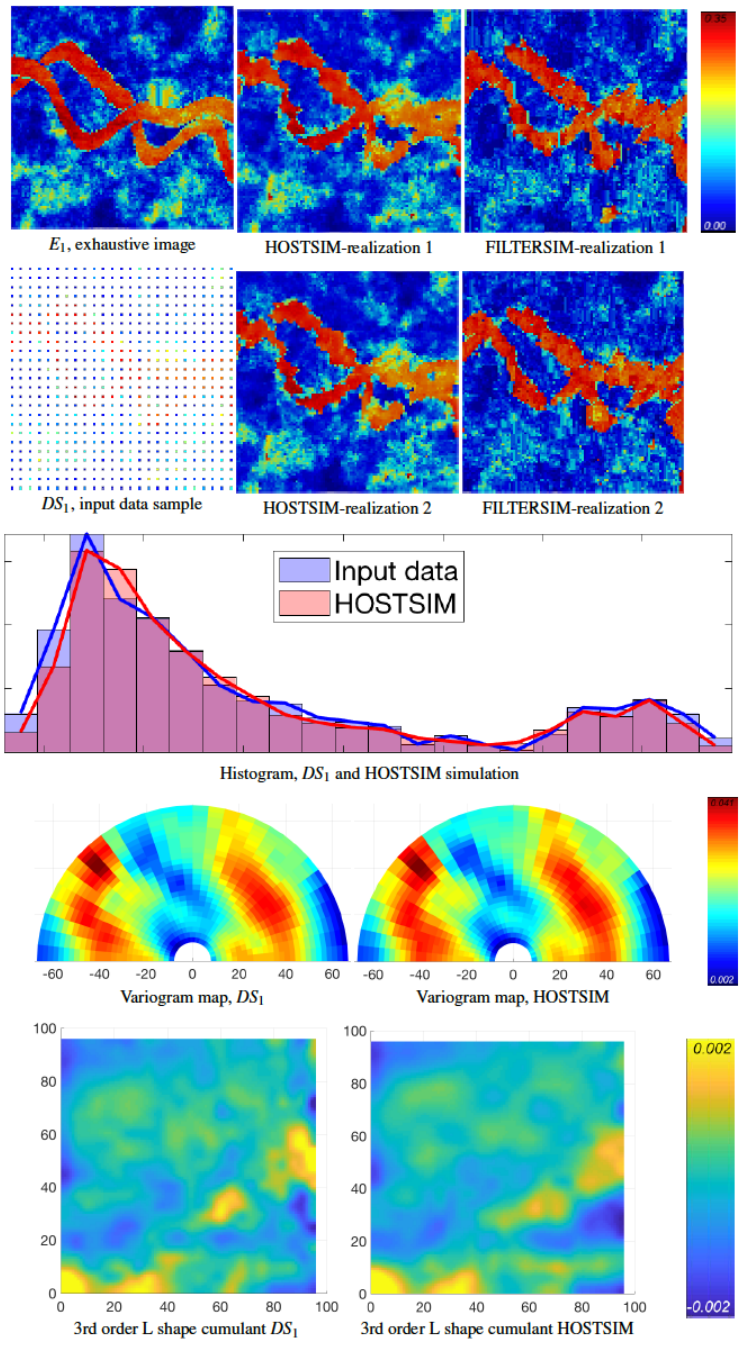

Figure 7: Example 1. The E_1 exhaustive image is provided on the top row (left), and the input image DS_1 is on the second row. Two HOSTSIM and FILTERSIM simulations are in the second and third columns (top section), followed by the histogram, variogram map and third-order spatial cumulant map below.

3.2 Example 2

The HOSTSIM realizations are presented in the Figures 9 and 10. As the number of input data are reduced to 156 samples, the quality of the results is reduced (relative to Example 1). The low-contrast regions are, however, simulated in these examples (light blue/green areas). Some parts of the channels are not generated in both realizations (the center-left part of the channel in Figure 9 and the center-left and bottom-left part of the channels in Figure 10) due to a lack of high-value samples in those regions, which are necessary in order to represent narrow channels. On the other hand, some channels are merged in the HOSTSIM

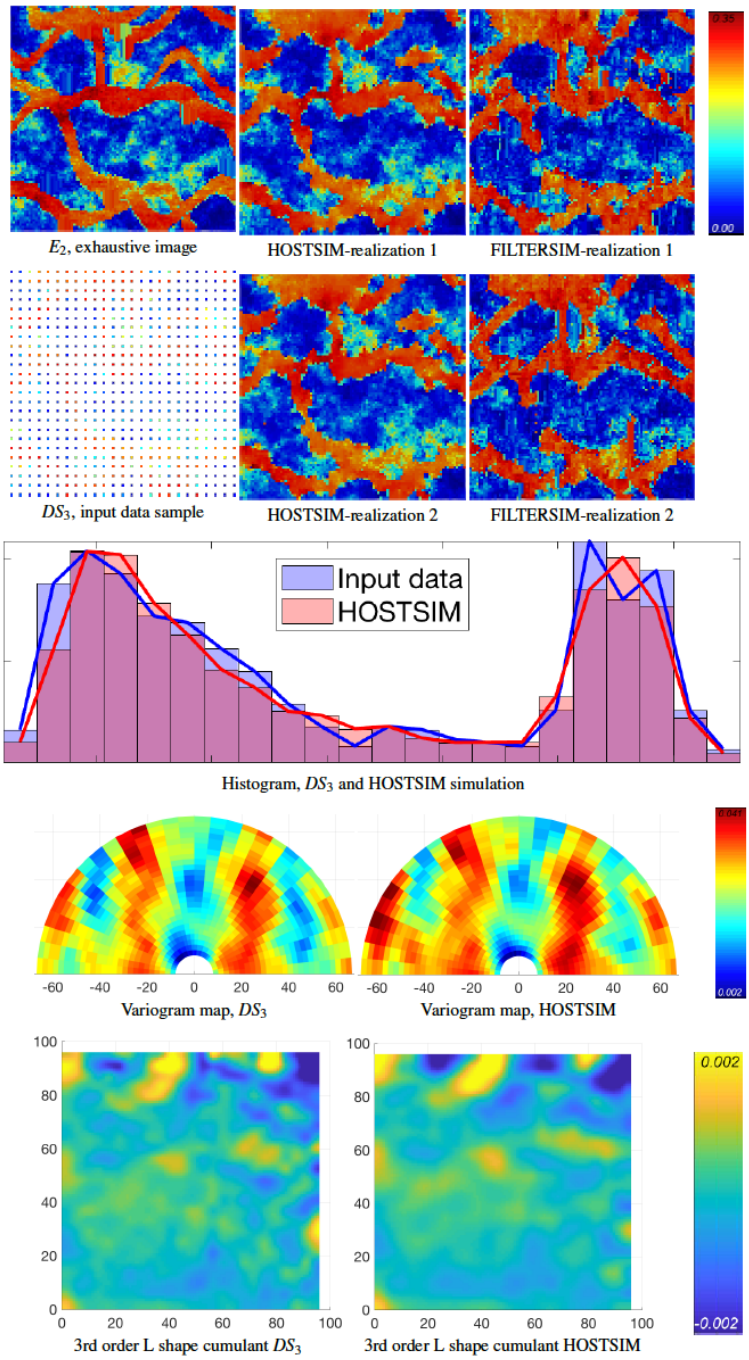

Figure 8: Example 1. The E_2 exhaustive image is shown on top row (left), and the input image DS_3 is on the second row. The two HOSTSIM and FILTERSIM simulations are in the second and third columns (above), followed by the histogram, variogram map and third-order spatial cumulant map below.

realizations (center-right part of both Figures 9 and 10), which is due to a lack of low-value samples between two channels in the input data samples. The validation graphs of the HOSTSIM simulations are similar to the ones generated for the input data samples in Figures 9 and 10, although they are somewhat degraded due to the sparsity of the input data. While the HOSTSIM realizations generated for this example are somewhat degraded compared to the previous example, there is still better connectivity in the data compared to results from FILTERSIM; for example, the channel in the center-right part of Figure 9 is partially generated in

HOSTSIM realizations but not well presented in the FILTERSIM realizations. Low-contrast regions are better represented in the HOSTSIM realizations as well (see the bottom-left part of the Figure 9).

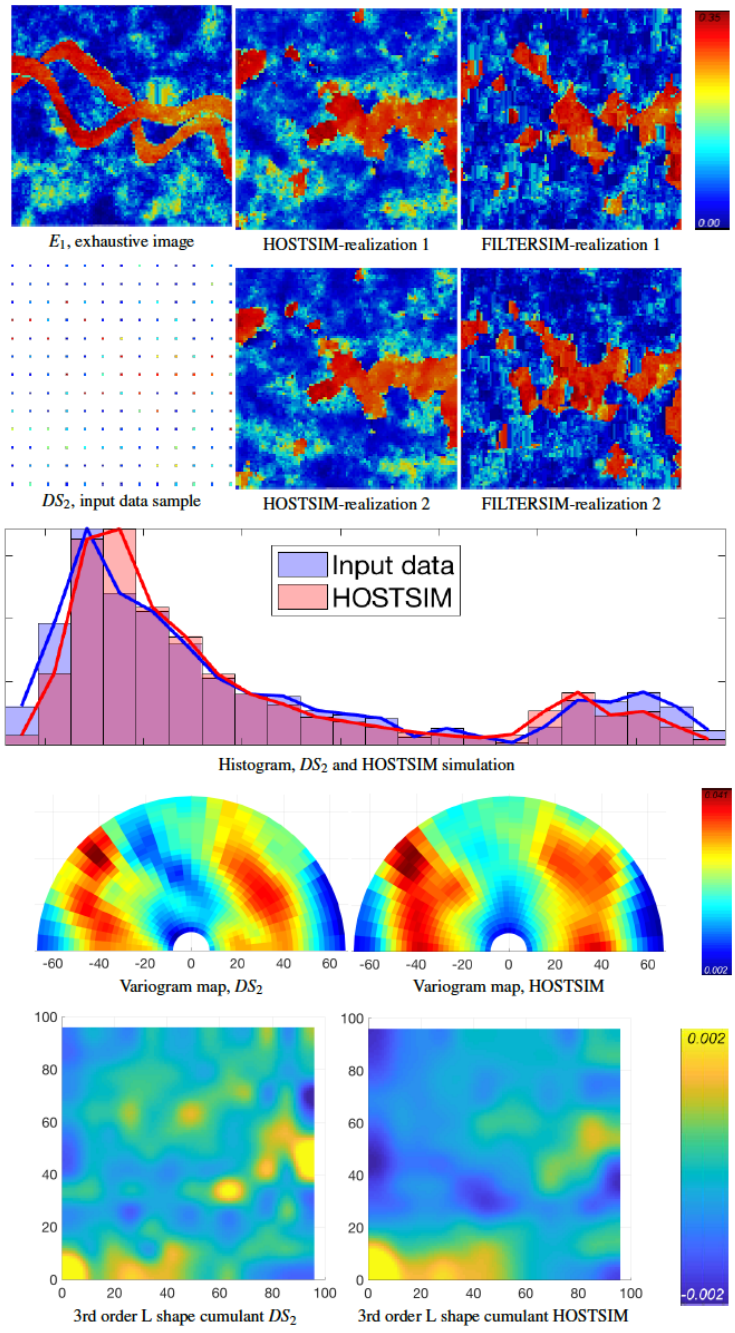

Figure 9: Example 2: The E_1 exhaustive image is on top row (left), and the input image DS_2 is on the second row. The two HOSTSIM and FILTERSIM simulations are in the second and third columns (above) followed by the histogram, variogram map and third-order spatial cumulant map (below).

To complete our analysis, two commonly used second-order similarity measures in computer vision are applied here to compare the generated realizations by HOSTSIM and FILTERSIM for all cases. These measures are PSNR and SSIM (Salomon 2007). The PSNR is a point-to-point difference measure between two images in the logarithmic scale; in particular, the realization image is compared to the exhaustive image from point to point. Realizations that are more similar to the exhaustive image result in greater values, and where PSNR values approach infinity, the images are identical. Hence, this measure is useful for qualitatively

comparing two realizations generated for an exhaustive image, as the PSNR values do not have an absolute interpretation. The SSIM, on the other hand, provides an absolute quantity comparing the similarity between a realization and an exhaustive image with a value between 0 and 1, where SSIM=1 indicates identical images. Table 1 presents the PSNR and SSIM results, which were calculated for HOSTSIM and FILTERSIM realizations generated for all four datasets (DS_1, DS_2, DS_3 and DS_4). The reference exhaustive image for each dataset is defined as the reference image in the calculation of PSNR and SSIM. HOSTSIM realizations are more similar to the exhaustive images in all of the cases. The SSIM values of the HOSTSIM realizations are $\approx 10\%$ more accurate than FILTERSIM realizations.

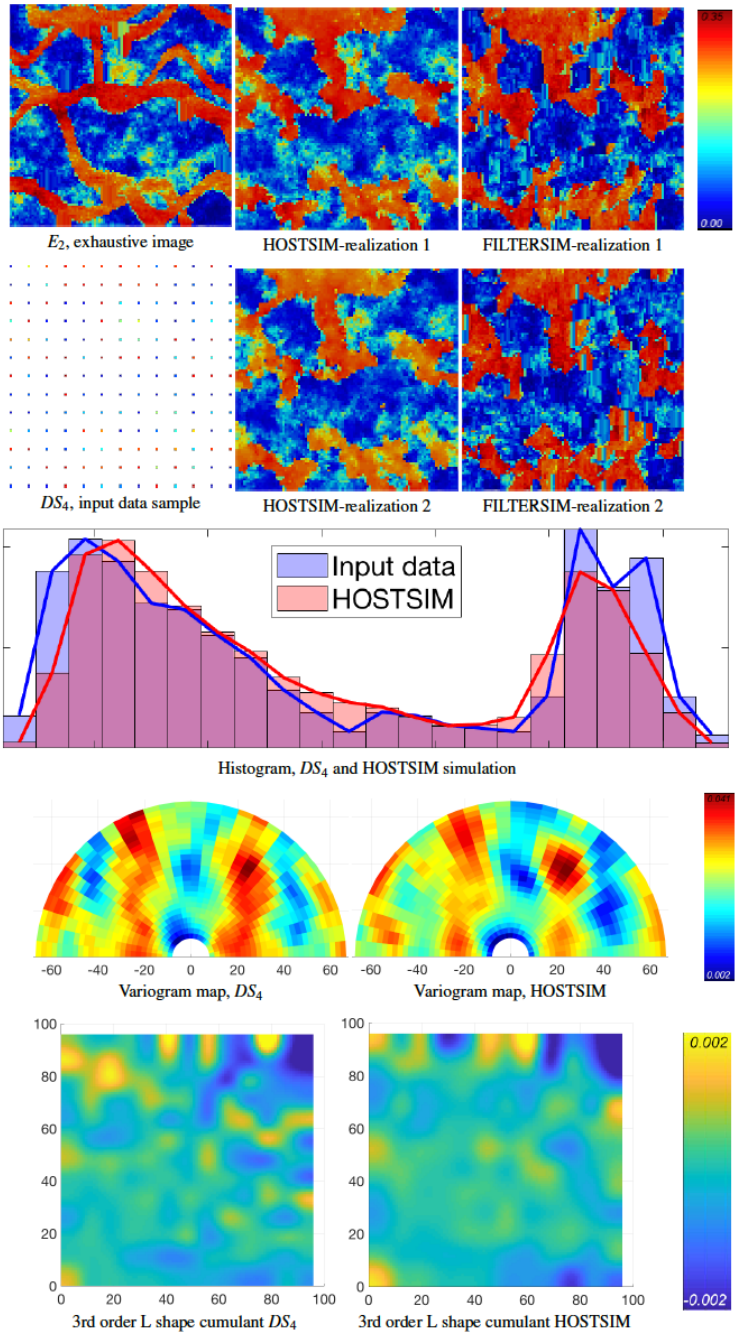

Figure 10: Example 2: The E_2 exhaustive image is shown (top row, left), and the input image DS_4 is provided on the second row. The two HOSTSIM and FILTERSIM simulations are in the second and third columns (above) followed by the histogram, variogram map and third-order spatial cumulant map (below).

Table 1: Table 1 PSNR (left) and SSIM (right) values for all four cases.

Case	HOSTSIM	FILTERSIM
DS_1	24.94	23.10
DS_3	23.10	21.29
DS_2	21.23	20.98
DS_4	20.17	18.76

Case	HOSTSIM	FILTERSIM
DS_1	0.61	0.50
DS_3	0.56	0.42
DS_2	0.47	0.43
DS_4	0.40	0.30

4 Conclusions

A new high-order, non-stationary stochastic simulation method, was introduced in this paper. Given a spatially sparse set of data and a training image, this method simulates realizations of attributes of interest on a finer spatial grid. The method is designed to use only training patterns that respect the local high-order statistics of the simulation pattern by comparing the spatial high-order statistics of the data-events of both training and the target simulation patterns.

The high-order statistics of the data-event of each simulation pattern is first calculated and compared to high-order statistics of the data-event calculated for the training patterns with a similar spatial template. The contribution of each training pattern in simulating a pattern is determined by a weight calculated based on the similarity of the high-order statistics of the training and simulation patterns. The calculated weights are designed to be invariant against ordering the data-event nodes. The high-order statistics of the simulation pattern is then estimated by calculating the non-stationary maximum likelihood estimate over the training patterns by giving more weight to the most similar patterns, after which a sample is drawn from the distribution function fitted to the estimated high-order statistics. The use of the high-order statistics in this simulation method ensures capturing and reproducing complex patterns in the generated simulations. Meanwhile, using the non-stationarity similarity measure ensures that only training patterns that respect the local statistics of the simulation are used in the estimation of the high-order spatial statistics of the pattern. Notable are that (a) non-stationary weight reduces the numerical instabilities that may be encountered in stationary high-order simulation methods; and (b) the transformation-invariant quality of the measure ensures that different orientations of each TI are considered for the simulation of the pattern. The latter allows for data-driven simulations even without using an training image. Two set of examples have been used for testing the proposed method. The first contains 625 data points and the second contains a very sparse set of 156 data points. The simulation results are calculated on a 100x100 size spatial grid. The training patterns are extracted from a 100x100 size grid image. Despite the sparsity of the data in both examples, the method is able to simulate the channels in the initial exhaustive image. This is due to the utilization of a non-stationary rotation invariant weighting for the contribution of the patterns based on the similarity of the high-order statistics of their data-events.

Appendix

PSNR and SSIM (Salomon 2007) are two image similarity measures that are used in this paper to evaluate the simulation quality. PSNR (peak signal-to-noise ratio) is a measure calculated by comparing a reconstructed image to the original image. Assuming two $m \times n$ images denoted I (reference) and K (reconstructed), first the mean-square-root error is calculated according to the following equation

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \tag{22}$$

The PSNR is then calculated as

$$PSNR = 10\log\left(\frac{\max(I)^2}{MSE}\right) . \tag{23}$$

This function returns a real value between zero and infinity. Greater values of PSNR imply better quality.

The structural similarity index (SSIM) is another quality measure used in this paper; it evaluates similarity according to a relative measure varying between 0 and 1 that compares an image to a reference image. Images that are more similar to the reference image result in SSIM values closer to 1. For the calculation of SSIM, two images I (reference) and K (reconstructed) of size $m \times n$ are considered. Multiple smaller-size windows are generated from both images. Consider x and y representing two windows from I and K. SSIM is calculated by

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c2)}{({\mu_x}^2 + {\mu_y}^2 + c_1)({\sigma_x}^2 + {\sigma_y}^2 + c_2)}, \tag{24}$$

where μ_x and μ_y are the mean values over x and y windows. The parameters σ_x, σ_y and σ_{xy} are the variance of x and y and the covariance between x and y, respectively. The values c_1 and c_2 are constants.

5 References

Abolhassani AAH, Dimitrakopoulos R, Ferrie FP (2017) A new high-order, nonstationary, and transformation invariant spatial simulation approach. In: Gómez-Hernández J, Rodrigo-Ilarri J, Rodrigo-Clavero M, Cassiraga E, Vargas-Guzmán J (eds), Geostatistics Valencia 2016. Springer, Dordrecht

Abramowitz M (1974) Handbook of mathematical functions. Dover Publications

Arpat G.B, Caers J (2007) Conditional simulation with patterns. Mathematical Geology 39(2):177–203

Boyd J P, Ong J R (2011) Exponentially-convergent strategies for defeating the runge phenomenon for the approximation of non-periodic functions, part two: Multi-interval polynomial schemes and multi-domain chebyshev interpolation. Applied Numerical Mathematics 61(4):460–472

Chatterjee S, Dimitrakopoulos R, Mustapha H (2012) Dimensional reduction of pattern-based simulation using wavelet analysis. Mathematical Geosciences 44(3):343–374

Chiles J P, Delfiner P (2012) Geostatistics: Modeling spatial uncertainty. Second Edition, Wiley

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

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

de Vries L M, Carrera J, Falivene O, Gratacos O, Slooten L J (2009) Application of multiple point geostatistics to non-stationary images. Mathematical Geosciences 41(1):29–42

Dimitrakopoulos R, Mustapha H, Gloaguen E (2010) High-order statistics of spatial random fields: Exploring spatial cumulants for modeling complex non-Gaussian and non-linear phenomena. Mathematical Geosciences 42(1):65–99

Funkhouser H G (1930) A short account of the history of symmetric functions of roots of equations. The American Mathematical Monthly 37(7):357–365

Goodfellow R, Consuegra F A, Dimitrakopoulos R, Lloyd T (2012) Quantifying multi-element and volumetric uncertainty, Coleman McCreedy deposit, Ontario, Canada. Computers & Geosciences 42:71–78

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

Guardiano FB, Srivastava RM (1993) Multivariate geostatistics: Beyond bivariate moments. In: Soares A (ed), Geostatistics Tróia '92. Springer, Dordrecht

Honarkhah M, Caers J (2010) Stochastic simulation of patterns using distance-based pattern modeling. Mathematical Geosciences 42(5):487–517

Hu LY, Chugunova T (2008) Multiple-point geostatistics for modeling subsurface heterogeneity: A comprehensive review. Water Resources Research, 44, DOI: 10.1029/2008WR006993

Journel AG (2005) Beyond covariance: The advent of multiple-point geostatistics. In: Leuangthong O, Deutsch CV (eds) Geostatistics Banff 2004. Springer, Dordrecht

Journel A G, Huijbregts C J (1978) Mining Geostatistics. Academic Press, London

Landau L, Lifshitz E (1980) Statistical physics. 3rd Ed. Pergamon Press, Oxford

Liu D, Wang Z, Zhang B, Song K, Li X, Li J, Li F, Duan H (2006) Spatial distribution of soil organic carbon and analysis of related factors in croplands of the black soil region, northeast china. Agriculture, Ecosystems & Environment 113(1):73–81

Lochbuhler T, Doetsch J, Brauchler R, Linde N (2013) Structure-coupled joint inversion of geophysical and hydrological data. Geophysics 78(3):ID1–ID14

Mariethoz G, Renard P, and Straubhaar j (2010) The direct sampling method to perform multiple-point geostatistical simulations, Water Resources Research 46, W11536, DOI: 10.1029/2008WR007621

Minniakhmetov I, Dimitrakopoulos R (2017) Joint high-order simulation of spatially correlated variables using high-order spatial statistics. Mathematical Geosciences 49(1):39–66, DOI:10.1007/s11004-016-9662-x

Minniakhmetov I, Dimitrakopoulos R, Godoy M (2018) High-order spatial simulation using Legendre-like orthogonal splines. Mathematical Geosciences DOI: 10.1007/s11004-018-9741-2

Mustapha H, Chatterjee S, Dimitrakopoulos R (2014) Cdfsim: Efficient stochastic simulation through decomposition of cumulative distribution functions of transformed spatial patterns. Mathematical Geosciences 46(1):95–123

Mustapha H, Dimitrakopoulos R (2010a) Generalized Laguerre expansions of multivariate probability densities with moments. Computers & Mathematics with Applications 60(7):2178–2189

Mustapha H, Dimitrakopoulos R (2010b) High-order stochastic simulation of complex spatially distributed natural phenomena. Mathematical Geosciences 42(5):457–485

Mustapha H, Dimitrakopoulos R (2011a) Hosim: A high-order stochastic simulation algorithm for generating three-dimensional complex geological patterns. Computers & Geosciences 37(9):1242–1253

Mustapha H, Dimitrakopoulos R (2011b) A new approach for geological pattern recognition using high-order spatial cumulants. Computers & Geosciences 36(3):313–334

Mustapha H, Dimitrakopoulos R, Chatterjee S (2011), Geologic heterogeneity representation using high-order spatial cumulants for subsurface flow and transport simulations, Water Resources Research 47, W08536, DOI: 10.1029/2010WR009515

Osterholt V, Dimitrakopoulos R (2007) Simulation of wireframes and geometric features with multiple-point techniques: application at Yandi iron ore deposit. In, Orebody Modelling and Strategic Mine Planning, 2nd ed., AusIMM, Spectrum Series 14:95–124

Rezaee H, Mariethoz G, Koneshloo M, Asghari O (2013) Multiple-point geostatistical simulation using the bunch-pasting direct sampling method. Computers & Geosciences 54:293–308

Salomon D (2007) Data compression: The complete reference. Springer Science & Business Media DOI: 10.1007/978-1-84628-603-2

Stien M, Kolbjørnsen O (2011). Facies modeling using a Markov mesh model specification. Mathematical Geosciences 43(6):611–624

Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. Mathematical Geology 34(1):1–21

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

Tamayo-Mas E, Mustapha H, Dimitrakopoulos R (2016) Testing geological heterogeneity representations for enhanced oil recovery techniques. Journal of Petroleum Science and Engineering 146:222–240

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

Yao L, Dimitrakopoulos R, Gamache M (2018) A new computational model of high-order stochastic simulation based on spatial legendre moments. Mathematical Geosciences DOI: 10.1007/s11004-018-9744-z

Zhang T, Gelman A, Laronga R (2017) Structure-and texture-based fullbore image reconstruction. Mathematical Geosciences 49(2):195–215

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