Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Learning high-order spatial statistics at multiple scales: A kernel-based stochastic simulation algorithm and its implementation

Lingqing Yao[a,b,*], Roussos Dimitrakopoulos[b], Michel Gamache[a]

[a] Department of Mathematics and Industrial Engineering, Polytechnique Montreal, Montreal, Quebec, H3T 1J4, Canada
[b] COSMO – Stochastic Mine Planning Laboratory, Department of Mining and Materials Engineering, McGill University, 3450 University Street, Montreal, Quebec, H3A 2A7, Canada

ARTICLE INFO

Keywords:
High-order spatial statistics
Geostatistical simulation
High-order simulation software
Kernel
Statistical learning

ABSTRACT

This paper presents a learning-based stochastic simulation method that incorporates high-order spatial statistics at multiple scales from sources with different resolutions. Regarding the simulation of a certain spatial attribute, the high-order spatial information from different sources is encapsulated as aggregated kernel statistics in a spatial Legendre moment kernel space, and the probability distribution of the underlying random field model is derived by a statistical learning algorithm, which matches the high-order spatial statistics of the target model to the observed ones. In addition, a related software is developed as the SGeMS plugin. Case studies are conducted with a known data set and a gold deposit, demonstrating reproduction of high-order spatial statistics from the available data, as well as practical aspects in mining applications.

1. Introduction

High-order stochastic simulation methods are amongst the latest developments in geostatistical simulation (Journel and Huijbregts 1978; David 1988; Goovaerts 1997; Remy et al., 2009; Mariethoz and Caers 2014), aiming to reproduce complex spatial patterns from the available data. The spatial patterns represent the interaction of spatial attributes of certain natural phenomena among multiple locations and they can be characterized by the high-order spatial statistics defined in different ways such as high-order spatial cumulants or high-order spatial moments (Dimitrakopoulos et al., 2010; Mustapha and Dimitrakopoulos 2010b; De Iaco and Maggio 2011). High-order simulation methods contrast with the multiple point simulation approaches, where the multi-point interrelations are indirectly captured as either the frequency of data events occurring at multiple locations (Guardiano and Srivastava 1993; Strebelle 2002; Journel 2003; Remy et al., 2009) or as similarity measures amongst patterns (Arpat 2005; Zhang et al., 2006; Mariethoz et al., 2010; Mariethoz and Caers 2014). Instead, the high-order simulation methods explicitly build probabilistic models based on high-order spatial statistics. For instance, Legendre polynomial expansion series are used to approximate the probability distributions of spatial attributes where the expansion coefficients are determined by computing the spatial cumulants, leading to an early development of a high-order simulation algorithm known as HOSIM (Mustapha and Dimitrakopoulos 2010a, 2011). The concept of high-order spatial statistics has also been extended to multiple variables to develop joint simulation of spatially correlated attributes (Minniakhmetov and Dimitrakopoulos 2016). The probabilistic model of high-order simulation makes no parametric assumptions of the probability distribution and thus characterizes the non-gaussian and non-linear features of the spatial attributes. The estimation of the probability distribution yields a numerical model based on components link to the empirical high-order spatial statistics calculated from the available data. In practice, the input data for estimating the probability distribution may impact the numerical stability of the related estimation. The available sample data alone may not be sufficient to infer the high-order spatial statistics required and, thus, may influence the numerical model. This limitation is alleviated with the use of a training image (TI) as the complementary statistical analog (Mustapha and Dimitrakopoulos 2010a). Another approximation model of a high-order simulation that shows substantial improvement with regards to numerical stability is found in Minniakhmetov et al. (2018). The latter authors use the Legendre-like splines as the basis functions for the approximation series, which leads to a better reproduction of spatial data patterns, as compared to the previous HOSIM method.

A concern when using a TI as a statistical analog of the underlying

[*] Corresponding author. Department of Mathematics and Industrial Engineering, Polytechnique Montreal, Montreal, Quebec, H3T 1J4, Canada.
E-mail address: lingqing.yao@mcgill.ca (L. Yao).

https://doi.org/10.1016/j.cageo.2021.104702
Received 30 August 2020; Received in revised form 6 January 2021; Accepted 14 January 2021
Available online 21 January 2021

random field model is the possible statistical conflicts between the TI and sample data. This issue is more prominent in multiple point simulation methods as they are TI-driven, and thus related methods have been proposed to select TIs according to certain criteria of consistency with the sample data (Pérez et al., 2014; Feng et al., 2017). Yao et al. (2020) propose a statistical learning framework for high-order sequential simulation in a newly defined kernel space; the related learning algorithm shows generalization capacity to comply with the inferred model from the TI with the spatial statistics of the sample data, and thereby mitigates the possible statistical conflicts. The concepts of statistical machine learning and kernels have been brought up in some state-of-the-art geostatistical methods. For instance, the Gaussian kernels are used to define pattern similarity in kernel space for clustering spatial patterns in TIs (Scheidt and Caers 2009). New spatial clustering methods have been proposed to incorporate spatial correlations among multivariate geological attributes based on statistical learning/modeling (Martin and Boisvert 2018; Talebi et al., 2020). A so-called Stochastic Local Interaction model (Hristopulos 2015) is proposed for geostatistical prediction based on the optimization of a specific energy functional where the distance measure for optimization can be defined through kernels. The kernels are also used in developing a learning-based algorithm for geostatistical interpolation based on high-order spatial statistics through implicit volterra series (Gonbadi et al., 2019). Amongst these recent developments, the kernels are usually defined as classic kernel functions such as RBF or Gaussian kernel, while a new kernel function is constructed in Yao et al. (2020) to explicitly incorporate the high-order spatial statistics with a systemic learning framework provided for high-order simulation. Specifically, a spatial Legendre moment kernel is proposed by Yao et al. (2020) to define the associated kernel space. The replicates of the data events (conditioning data) retrieved from the available data are mapped into the spatial Legendre moment kernel space by a feature mapping function. Thereafter, the so-called empirical kernel statistics are defined by taking a sample average of the mapped elements in the kernel space corresponding to the replicates. As a result, the empirical kernel statistics carried high-order spatial statistics of the replicates. On the other hand, the target probability distribution from the related random field model can be embedded into the same kernel space through the termed expected kernel statistics. A kernelized learning algorithm is designed specifically to match the expected kernel statistics to the empirical kernel statistics, which results in a simulation model with a reproduction of high-order spatial statistics from the available data. Although the proposed statistical learning framework is general, one limitation of the application in Yao et al. (2020) is that the replicates retrieved from the TI act as the only training data in the related learning algorithm; this may influence the spatial continuity of the realizations given that the statistical conflicts between the TI and sample data are severe. Yao et al. (2021) propose a TI-free high-order simulation method based on the statistical learning framework. The concept of aggregated kernel statistics is defined such that the samples with different spatial configurations can be effectively utilized for statistical inference of the random field model. A limitation of the above-mentioned TI-free simulation method is that the quality of the realizations depends on the sampling density. While the sample data are relatively sparse, the fine-scale spatial structures of the spatial attributes of interest are not well represented. The limitations found in previous simulation methods motivate the present research to propose a new type of aggregated kernel statistics, which aims to incorporate the high-order spatial information at multiple scales. Specifically, the sample data are relatively sparse and thus carry high-order spatial information at coarse scales. On the other hand, the TIs are exhaustive and can provide high-order spatial information at finer scales. The general idea of the proposed aggregated kernel statistics in this paper is to exclude the influence of the TI from deriving the high-order spatial statistics at coarse scales by only utilizing the sample data, while complementing the high-order spatial information with TI. Thereafter, the aggregated kernel statistics are utilized in the statistical learning framework for further inference of the random field model. It should be noted that the aggregated kernel statistics proposed in this paper emphasize the incorporation of high-order spatial information from different sources across multiple spatial scales while minimizing their statistical inconsistency, which is very different from the application in Yao et al. (2021) where only sample data are considered in computing the kernel statistics. Although the present study considers only two different scales of data as the samples and the TI, the concept of the aggregated kernel statistics proposed herein can be generalized to multiple scales. The utilization of multi-scale information from various sources has recently drawn attention in mining and geographical applications (Neves et al., 2019; Rasera et al., 2020). In addition, a high-order simulation program is developed accordingly and described in this paper. The implementation is written in C++ language and is compatible to the SGeMS software.

In the following sections, Section 2 presents the high-order simulation method based on statistical learning and the concept of the aggregated kernel statistics. Section 3 describes a kernelized high-order simulation program and its implementation in C++ language. Section 4 contains two different case studies with a synthetic data set and at a gold deposit. Conclusions are presented in Section 5.

2. Method

In this section, concepts of high-order sequential simulation are first outlined, followed by a brief overview of the spatial Legendre moment kernel space. The concept of aggregated kernel statistics at different scales is then presented and utilized to develop a kernelized learning algorithm.

2.1. High-order sequential simulation

Suppose that the attributes of interest are modeled as a random field $\boldsymbol{Z}(\boldsymbol{u})$ where $\boldsymbol{u}$ represents locations at a certain spatial domain. The attributes at multiple locations within the spatial domain comprise a multivariate probability distribution. The multivariate probability distribution can be decomposed into a sequence of conditional probability distributions so that the random values can be sequentially drawn from the multivariate probability distribution to generate the simulated realizations. Without loss of generality, the conditional probability density functions (CPDF) can be approximated as $f(z_0|\zeta_1,\dots,\zeta_N\ )$ given that the node Z_0 to be simulated center at $\boldsymbol{u}_0$ and the conditioning data within its neighborhood located at $\boldsymbol{u}_1,\dots,\boldsymbol{u}_N$ with the value of attributes corresponding to $\zeta_1,\dots,\zeta_N$. In terms of high-order sequential simulation, the high-order spatial statistics are taken into account for approximating the CPDF $f(z_0|\zeta_1,\dots,\zeta_N)$, and the conditioning data $\zeta_1,\dots,\zeta_N$ are called as a data event associated with a spatial template defined by distance vectors of location $\boldsymbol{u}_1,\dots,\boldsymbol{u}_N$ to the location $\boldsymbol{u}_0$ of the center node. The high-order spatial statistics are contained in the replicates of a data event for inference. Note that the replicates of a data event in high-order simulation methods are not necessary to have identical or similar attribute values to the data event, but rather to have the same spatial template, i.e., the same data geometry. In general, the replicates from the sample data correspond to spatial template at coarse scales and the replicates from the TI provides spatial information at finer scales because of the sparsity of the sample data in contrast to the exhaustive TI.

2.2. Kernel space and spatial Legendre moment kernel

Suppose the original data space of the considered spatial attributes is represented by a nonempty set $\mathbb{E}$, then an element $x\in\mathbb{E}$ can be taken to a kernel space $\mathscr{H}$ by a so-called feature mapping function $\phi(x):\mathbb{E}\rightarrow\mathscr{H}$. The kernel space $\mathscr{H}$ is a Hilbert space with the inner product defined by a positive definite kernel function $K:\mathbb{E}\times\mathbb{E}\rightarrow\mathbb{R}$, where $\mathbb{R}$ is the set of the real numbers. Given a Hilbert space $\mathscr{H}$ with the kernel K, then for $x,y\in$

$\mathbb{E}$ and the corresponding features $\phi(x), \phi(y) \in \mathscr{H}$, the inner product on $\mathscr{H}$ can be defined as

$$\phi(x), \phi(y)_{\mathscr{H}} = K(x, y). \tag{1}$$

The symbol $< \cdot, \cdot >_{\mathscr{H}}$ represents the inner product of elements in the Hilbert space $\mathscr{H}$ throughout this paper. An interesting property with the kernel K is that the function $\phi(x) : \mathbb{E} \rightarrow \mathscr{H}, x \mapsto K(\cdot, x)$ also defines a feature map namely as reproduce kernel map or canonical feature map (Scholkopf and Smola 2001; Steinwart and Christmann 2008). The kernel function K has the reproducing property as

$$< f(x), K(\cdot, x)_{\mathscr{H}} = f(x), \tag{2}$$

$\forall x \in \mathbb{E}$ and $\forall f \in \mathscr{H}$, therefore there is

$$< K(\cdot, x), K(\cdot, y)_{> \mathscr{H}} = K(x, y). \tag{3}$$

This kind of reproduce kernel map is adopted throughout this paper as the feature mapping function from the original data space to the kernel space. It is obvious from Eq. (3) that the elements in the kernel space after the feature mapping from the original data space have the similarity measure defined as the distances between each other through the kernel function K.

The spatial Legendre moment kernel (Yao et al., 2020) allows to carry over the high-order spatial statistics information from the original data space to the newly defined kernel space with the definition as

$$K_V(\boldsymbol{X}, \boldsymbol{Y}) = \prod_{i=0}^{N} \left[\sum_{w=0}^{W} \left(w + \frac{1}{2} \right) P_w\ (x_i\ ) P_w\ (y_i\ ) \right], \tag{4}$$

where K_V is the kernel corresponding to the set of random variables associated with a spatial template of N distance vectors, and P_w is the Legendre polynomial of order w.

2.3. Aggregating kernel statistics at different scales

With the definition of spatial Legendre moment kernel in Eq. (4), the empirical kernel statistics can be defined accordingly based on the sample average of the elements in the kernel space mapped from samples in the original data space. The kernel function K_V depends on the spatial template involved, and so as the kernel statistics from the available data are related to the spatial templates of the data events. When both the sample data and the TI are available for retrieving the replicates and inferring the kernel statistics, the replicates from the two different sources generally carry high-order spatial statistics information at different scales. Specifically, the sample data are relatively sparse that frequently the spatial configuration of the replicates from them could only partially match to the spatial template of the data event, and theses replicates carry the spatial statistics at coarser scale with relatively higher compliance to the underlying random field. On the contrary, the TI are exhaustive data and the replicates from it can fully match the spatial template of the data event, thus the replicates provide spatial statistics at finer scale but possibly with less compliance to the underlying random field model.

Suppose that the spatial template of the data event be noted as v and the corresponding set of random variables be noted as V. Let v_s be the spatial template of the replicates of the data event retrieved from the sample data and the associated set of random variables as V_s, v_s and V_s are the subsets of v and V, respectively. Let $\mathscr{G}_{v_s}$ be the set of replicates from the sample data and the number of these replicates be n_s, the kernel statistics based on $\mathscr{G}_{v_s}$ can be defined as

$$\kappa[\mathscr{G}_{v_s}] = \frac{1}{n_s} \sum_{i=1}^{n_s} K_{V_s} \left(\boldsymbol{\zeta}^s_{i, v_s}, \cdot \right), \tag{5}$$

where ζ^s_{i,v_s} is the vector of the attribute values corresponding to the replicates in set $\mathscr{G}_{v_s}$. The kernel statistics of the replicates from the TI can be defined separately in a similar way. The motivation of aggregating kernel statistics at different scales is to utilize the part of high-order spatial information of the replicates from the sample data and in the meanwhile complement the rest part of high-order spatial information using the replicates from the TI. In other words, the spatial template v is divided into two sub-templates v_s and v_t respectively corresponding to the sample data and the TI, and so are the set of random variables are divided into V_s and V_t, respectively. Therefore, there are

$$v = v_s \cup v_t, \tag{6}$$

and

$$V = V_s \cup\ V_t. \tag{7}$$

The above subdivision regarding the spatial template also leads to kernel subspaces with kernels K_{V_s} and K_{V_t}. Suppose the ensemble of replicates from both the sample data and the TI denote as a set $\mathscr{G}_v$ and let n_t denote the number of replicates from the TI. The aggregated kernel statistics combining the replicates both from the TI and the sample data at different scales are defined as

$$\kappa[\mathscr{G}_v\ ] = \frac{1}{n_s} \sum_{i=1}^{n_s} K_{V_s}\ \left(\boldsymbol{\zeta}^s_{i,v_s}, \cdot \right) + \frac{1}{n_t} \sum_{j=1}^{n_t} \left[K_V\ \left(\boldsymbol{\zeta}^t_{j,v}, \cdot \right) - K_{V_s} \left(\boldsymbol{\zeta}^t_{j,v_t}, \cdot \right) \right], \tag{8}$$

where $\boldsymbol{\zeta}^s_{i,v_s}$, $\boldsymbol{\zeta}^t_{j,v}$ and $\boldsymbol{\zeta}^t_{j,v_t}$ represent the replicates from the sample data and the replicates from the TI with spatial template v and v_t, respectively.

2.4. Kernelized high-order sequential simulation algorithm

The high-order spatial information from both the sample and the TI can be represented by the aggregated kernel statistics at two scales. In terms of high-order sequential simulation, the target is to obtain conditional probability distributions which match the high-order spatial statistics of available data. This matching of high-order spatial statistics can be conveniently achieved by a statistical learning algorithm in kernel space. Suppose that the target probability density function $\widehat{p}$ lies in the convex space of certain prototype probability density functions p_i as

$$\widehat{p} = \sum_{i=1}^{n} \alpha_i p_i, \tag{9}$$

where $\sum_{i=1}^{n} \alpha_i = 1$ and $\alpha_i \geq 0, \forall\ 1 \leq i \leq n$. It is straightforward that the expected kernel statistics with regards to the probability distribution can be defined as

$$\kappa_0\ [\widehat{p}] = E_{z_0 \sim \widehat{p}}\ [K_0\ (z_0, \cdot)] \tag{10}$$

where Z_0 is the center node to be simulated and K_0 is the corresponding kernel function. The aggregated kernel statistics defined in Eq. (8) can be projected to the same kernel space through marginalization, and therefore the expected kernel statistics can be matched to the observed kernel statistics from the available data simply by minimizing the distance of two elements in the kernel space. Given that the conditioning data as $\Lambda = \{\zeta_1, \ldots, \zeta_N\}$ and the evaluation of $\kappa[\mathscr{G}_v]$ on Λ as $\kappa[\mathscr{G}_v|\Lambda]$, the projection of the aggregated kernel statistics can be defined as

$$\kappa_0\ [\mathscr{G}_v\ |\Lambda] = \frac{\kappa[\mathscr{G}_v; \Lambda]}{\int_{[-1,1]} \kappa[\mathscr{G}_v\ ; \Lambda] dz_0}. \tag{11}$$

Specifically, the statistical learning of high-order spatial statistics leads to a minimization problem

$$\min_{\widehat{p}} \|\ \kappa_0\ [\mathscr{G}_v\ |\Lambda] - \kappa_0\ [\widehat{p}] \|^2_{\mathscr{H}}. \tag{12}$$

The minimization in Eq. (12) amounts to solve a quadratic problem in a general form Song et al. (2008) as

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T\left(\boldsymbol{Q}+\lambda \boldsymbol{I}\right)\boldsymbol{\alpha}-\boldsymbol{q}^T\boldsymbol{\alpha}$$

$$\sum_{i=1}^{n}\alpha_i = 1 \tag{13}$$

$$\alpha_i \geq 0, \forall 1 \leq i \leq n.$$

$\boldsymbol{I}$ is the identity matrix and λ is a regularization constant. Matrix $\boldsymbol{Q}$ and vector $\boldsymbol{q}$ differ as the kernel function varies. For deriving the entries of the matrix $\boldsymbol{Q}$ and vector $\boldsymbol{q}$, as well as solving the quadratic programming problem in the SLM-kernel space, the readers are referred to Yao et al. (2020).

As long as the target conditional probability density functions are determined through the above learning process, the rest of simulation follows the general procedure of sequential simulation. Hence, the kernelized high-order sequential simulation algorithm can be described as follows

(1) Transform sample data and TI to the domain of Legendre polynomials (interval [-1, 1]).
(2) Create a random path to visit the simulation grid.
(3) Find the conditioning data inside the neighborhood of the current node to simulate as the data event, the spatial template of the data event is used to retrieve replicates from the sample data and TI.
(4) Compute the aggregated kernel statistics defined in Eq. (8) from the replicates retrieved from the sample data and TI.
(5) Match the kernel statistics of the target CPDF to the aggregated kernel statistics and build the quadratic programming problem through Eqs. (12) and (13). Solve the quadratic programming problem to derive the target CPDF.
(6) Generate a random value from the target CPDF and add it to the simulation grid.
(7) Repeat from steps (3) to (6) until all the nodes on the simulation grid are simulated.
(8) Back transform the node attributes of the simulation from the interval [-1, 1] to the original data space.

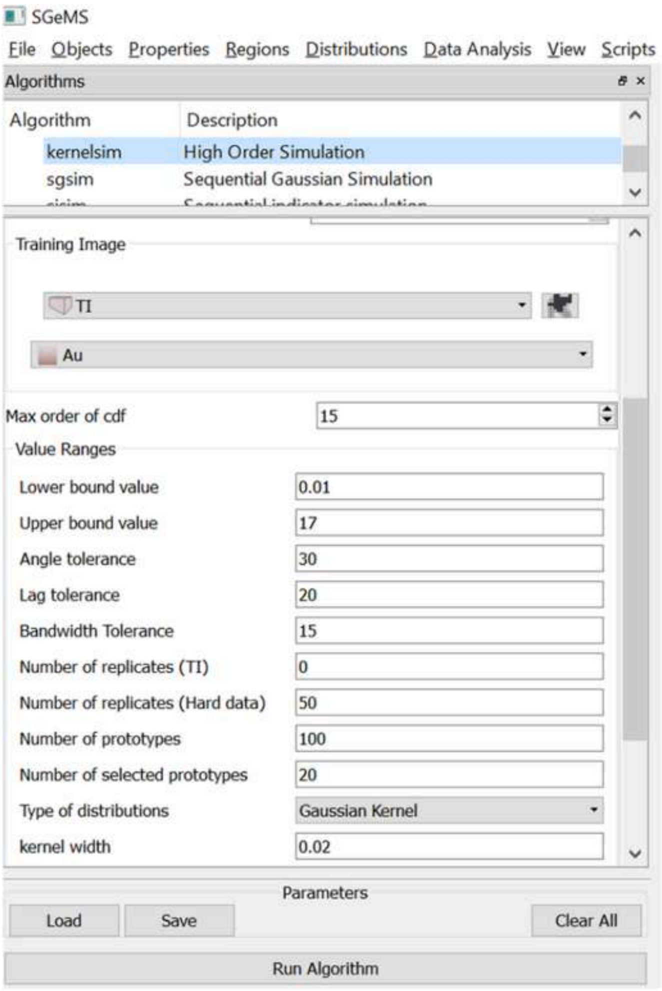
For the general computational complexity of the computation of kernel statistics and statistical learning algorithm, readers are referred to Yao et al. (2020). The additional computational cost in the current method is from the computation of the aggregated kernel statistics, which depends linearly on the number of the scales encountered as well as the size of the related training data.

3. A kernelized high-order simulation program

The kernelized high-order simulation program is developed as a software compatible with the SGeMS platform (Remy et al., 2009). The GUI is developed on top of the SGeMS as a plugin that includes the selection of the algorithm and the related input parameters (Fig. 1). All the file formats comply to the convention of SGeMS and thus can be visualized through it. The program is written in C++ language and follows the generic programming paradigm adopted in the design of GsTL, a geostatistical template library (Remy et al., 2002). The main workflow contains three major C++ classes which are described as the following.

3.1. Class kernelsim

This class is the application class communicating with the SGeMS platform through the user parameters, as well as running the simulation algorithm from the GUI. The class is derived from a predefined interface from the SGeMS platform so that it is compatible to the function calling convention of SGeMS. The object from the class *kernelsim* calls the sequential simulation function to start the high-order simulation procedure. The parameters of the proposed simulation algorithm can either be input from the GUI by the user or can be loaded from an XML file. The parameters are described in Table 1.


SGeMS
File Objects Properties Regions Distributions Data Analysis View Scripts
Algorithms
Parameters
Load
Save
Clear All
Run Algorithm

Algorithm	Description
kernelsim	High Order Simulation
sgsim	Sequential Gaussian Simulation
sisim	Sequential indicator simulation

Training Image
TI
Au

Max order of cdf: 15
Value Ranges
Lower bound value: 0.01
Upper bound value: 17
Angle tolerance: 30
Lag tolerance: 20
Bandwidth Tolerance: 15
Number of replicates (TI): 0
Number of replicates (Hard data): 50
Number of prototypes: 100
Number of selected prototypes: 20
Type of distributions: Gaussian Kernel
kernel width: 0.02

Fig. 1. GUI of the algorithm panel.

Table 1
Parameters description.

Parameter	Range
Maximum order of Legendre polynomials	10–20
Maximum number of conditioning data	10–30
Number of replicates from the TI	−1: take all the replicates n > 0: n replicates from the TI
Hard data usage	0: only use hard data 1: incorporate both the hard data and the TI −1: not using the hard data (only use the TI)
Angle tolerance	15–45
Lag tolerance	Application dependent
Bandwidth	Application dependent
Dimensions of searching window	Application dependent
Number of prototype distributions	10–20
Number of divisions on the interval	100–200
Scale parameter of the prototype distribution	0.01–0.05

3.2. Class SLM_kde_estimator

This class serves as the role to estimate the conditional probability density function through the learning algorithm. The class SLM_kde_estimator first calls the other function class to process the replicates which returns the aggregated kernel statistics. The main functions inside the class include the selection of the prototype distributions, construction of quadratic programming problem in Eq. (13), solving the quadratic programming problem to obtain the target conditional probability density function.

3.3. Class replicate_processor

This class is designed for processing the replicates. The conditioning data, the sample data, and the TI are used as input to this class. The spatial templates of the data events are constructed from the spatial configuration of the conditioning data. There are two major member functions defined in this class. The first function retrieves the replicates from both the sample data and the TI, respectively. The other function computes the aggregated kernel statistics from the retrieved replicates according to Eq. (8). The aggregated kernel statistics are passed to the object of the class.

SLM_kde_estimator to estimate the target probability density function.

4. Numerical results

Two separate case studies are carried out to test the developed simulation program. The first case study is conducted with a synthetic data set to verify the performance of the proposed simulation method. The other case study carries out the stochastic orebody modeling at a gold deposit, aiming to test the proposed method in a three-dimensional space, as well as its practical aspects in real-life mines.

4.1. Case study with a synthetic data set

The Stanford V Reservoir data set available in Mao and Journel (1999) are used to conduct the simulation in this case study. Specifically, two sections are extracted from the data set and the sections consist of 100 × 100 cells. One section is regarded as the exhaustive image where 200 points are randomly drawn from this image. The other section is rotated 45° clockwise so that the channels have distinct preferential directions from the exhaustive image after the rotation. The rotated section acts as the TI in this case study to represent the situation of the statistical conflicts existing between the TI and sample data. The exhaustive image, the TI and the sample data are shown in Fig. 2.

Two realizations using the above sample and the TI are displayed in Fig. 3. The visualization of the simulated results demonstrates good reproduction of the channels in the preferential orientation along the vertical direction from the exhaustive image. In addition, 10 realizations are generated to evaluate the overall performance of the simulation method in reproducing the low-order statistics. The latter includes the proportions and the second-order spatial statistics from the sample data, where the histograms and variograms of the 10 realizations are compared to those of the sample data, TI and exhaustive image, as shown in Fig. 4 and Fig. 5, respectively. The comparison of the histograms shows that the proposed simulation method has a reasonable reproduction of proportions from the sample data as well as the exhaustive image. The comparison of variograms clearly shows that the simulated realizations tend to have the similar second-order spatial statistics to the sample data instead of the TI. A further comparison of the third- and fourth- order cumulant maps of two separate realizations

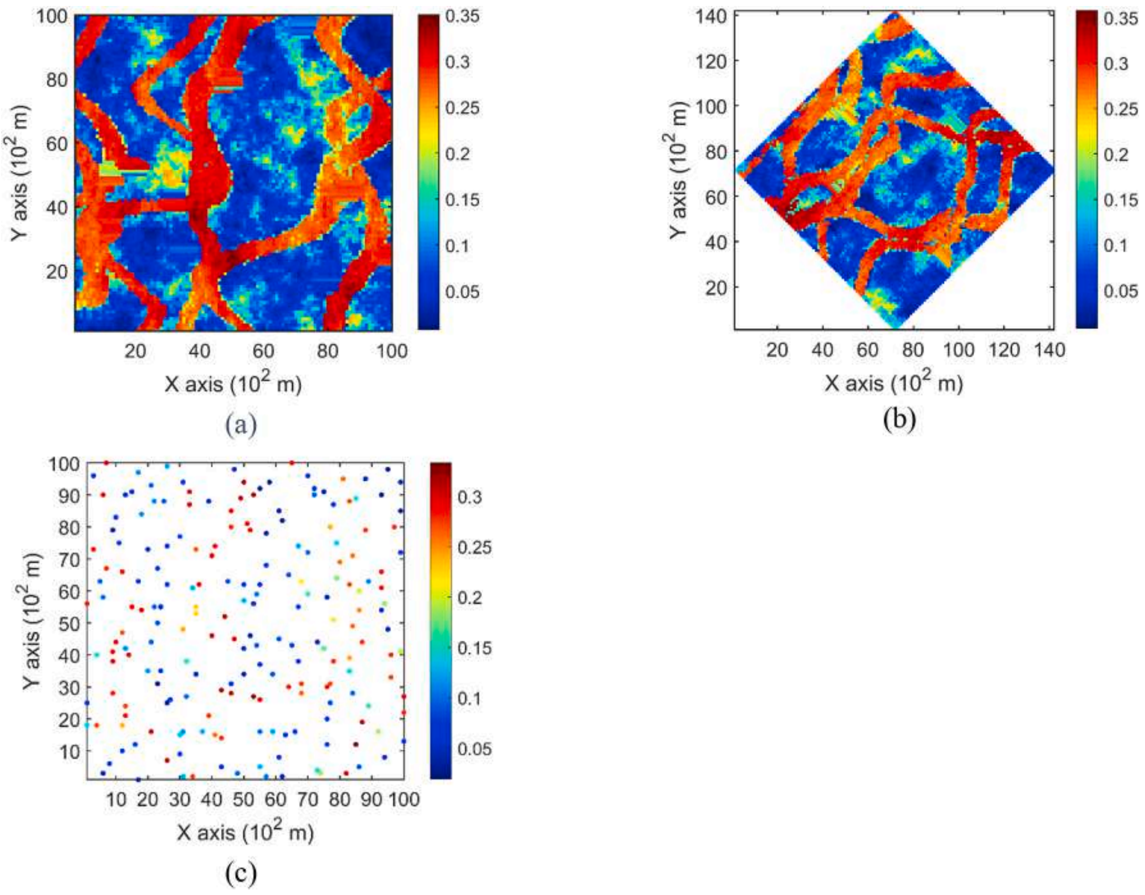


Fig. 2. (a) Exhaustive image; (b) training image; (c) sample data drawn from the exhaustive image.

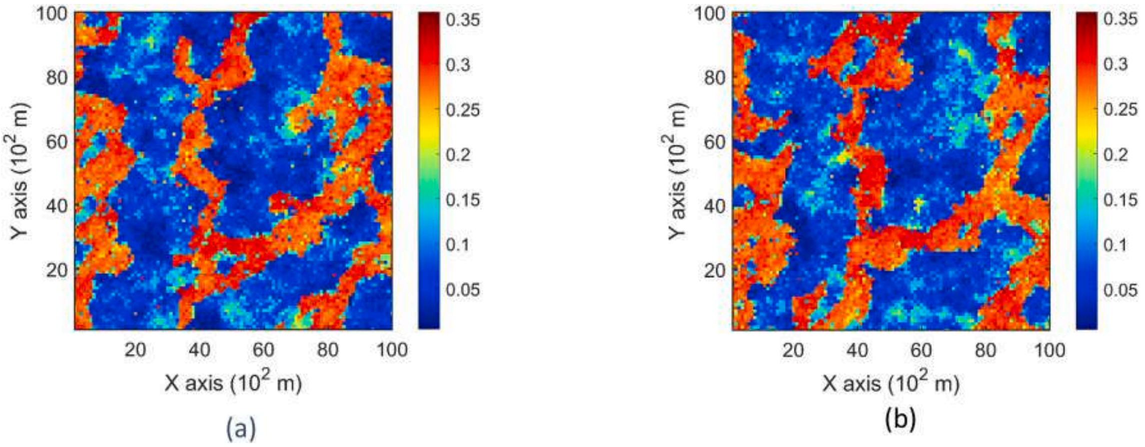


Fig. 3. Two simulated realizations using the samples and the TI shown in Fig. 2.

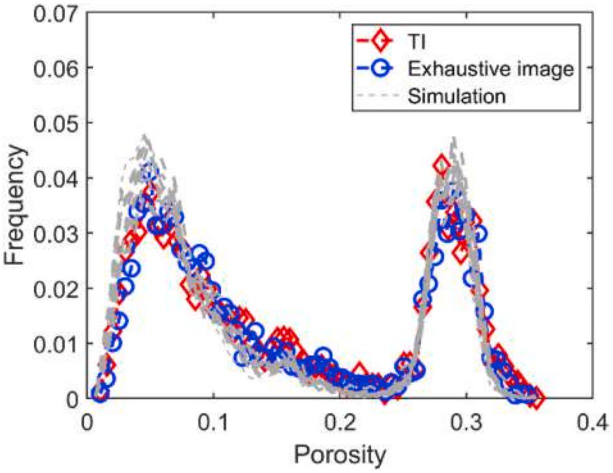


Fig. 4. Histograms of 10 simulated realizations using the samples and the TI shown in Fig. 2.

with those of the sample data, the TI and the exhaustive image are illustrated in Fig. 6 and Fig. 7, respectively. The spatial template used in the third-order cumulant maps includes directions along the X-axis and Y-axis and the spatial template of the fourth-cumulant maps includes an additional direction along the diagonal. The fourth-order cumulant maps are normalized to visually highlight the spatial patterns. Significant difference can be seen between the cumulant maps of the TI and those of the sample data and exhaustive image. The similarity between the cumulant maps of the realizations and the exhaustive image implies that the simulation method is able to mitigate the statistical conflicts between the samples and the TI, maintaining reasonable reproduction of both low-order and high-order spatial statistics from the sample data.

4.2. Case study at a gold deposit

In this section, a case study at a gold deposit is presented to document practical aspects of the developed simulation program in stochastic orebody modeling. The gold deposit contains samples spatially distributed in 407 exploration drill holes as shown in Fig. 8a. The samples are composited to 10 m. The simulation grid is defined as blocks of size 5 m × 5 m × 10 m. The TI is generated from blasthole data in a mined out area of the deposit, and a cross-section is shown in Fig. 8b.

Cross-sections of two different realizations are shown in Fig. 9. The histograms of 10 different realizations are shown in Fig. 10 and the comparison shows that the simulation method reproduces the histogram of the Au grades from the drill hole samples. The variograms of the same set of 10 realizations are shown in Fig. 11. The comparison results also show that the variograms of the simulated realizations resemble more

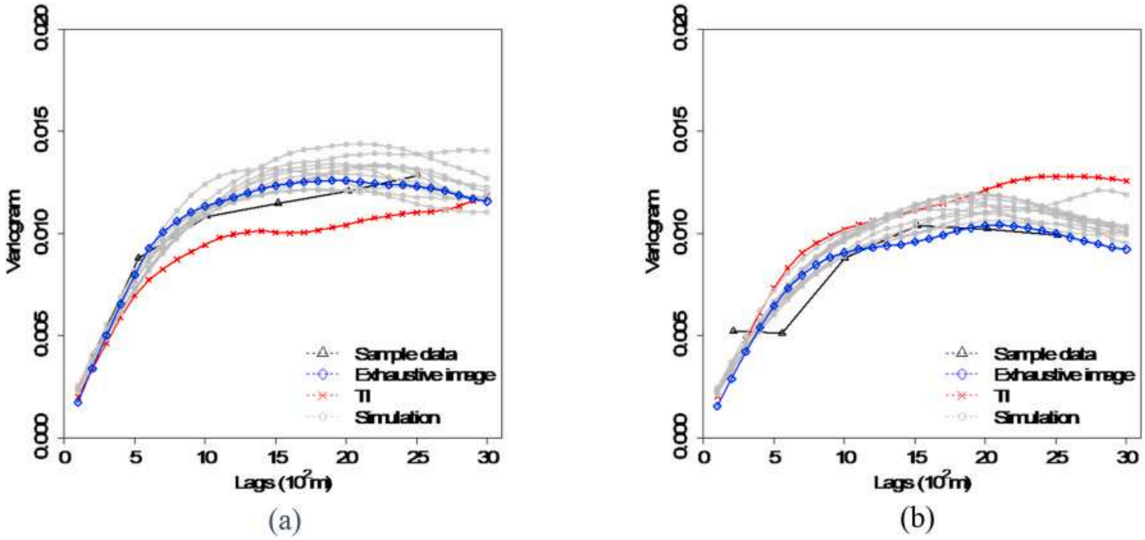


Fig. 5. Variograms of 10 simulated realizations along (a) X-axis and (b) Y-axis, using the samples and the TI shown in Fig. 2.

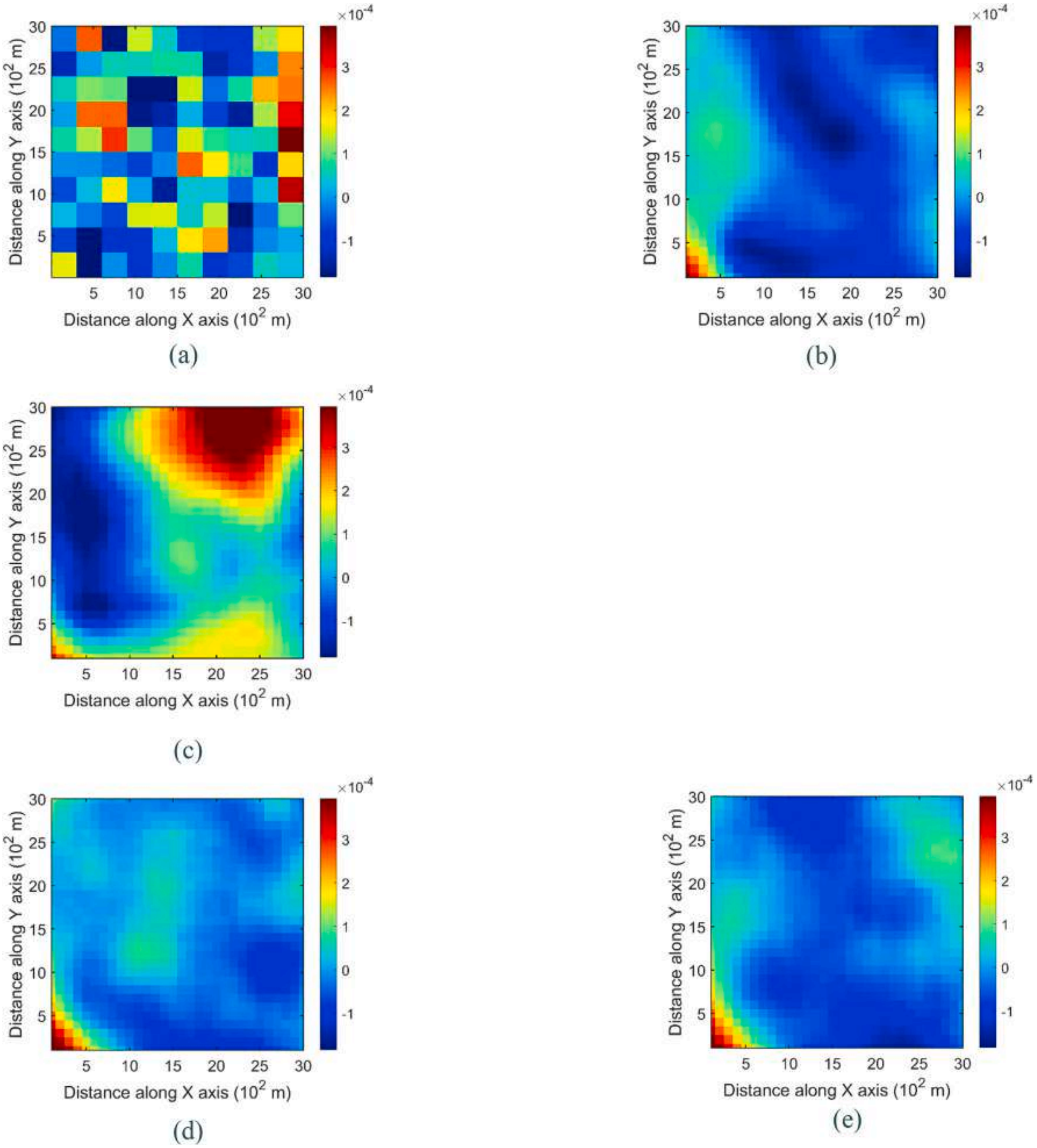


Fig. 6. Third-order cumulant maps of (a) sample data; (b) exhaustive image; (c) TI; (d) realization in Fig. 3a; (e) realization in Fig. 3b.

closely the variograms of the sample data, instead of those of the TI. Fig. 12 shows the comparison of third-order cumulant maps of the two realizations displayed in Fig. 9 to the third-order cumulant maps of the sample data and TI. The fourth-order cumulant maps are compared in the same manner and are shown in Fig. 13. Both the third-order or the fourth-order cumulant maps demonstrate distinct patterns compatible with the corresponding cumulant maps of the sample data. The high-order spatial information from the TI is only partly incorporated to complement the fine spatial structures of the stochastic orebody models generated with the proposed simulation method. Therefore, the high-order spatial statistics from the simulated realizations retain the main features from sample data, reducing the influence of the possible statistical conflicts from the TI.

5. Conclusions

The present paper presents an extension of the high-order simulation method based on the statistical learning framework (Yao et al., 2020). A modified concept of aggregated kernel statistics is proposed to incorporate the high-order spatial information at two different scales from the sample data and TI. Specifically, the aggregated kernel statistics proposed herein contain the high-order spatial information at the coarse scales from the sample data with high-order spatial information at the finer scales complemented by the TI. These aggregated kernel statistics are utilized in a kernelized learning algorithm to develop the high-order simulation method, which incorporates high-order spatial statistics from both the sample data and the TI. Although the present study only considered the data at two different scales, the proposed aggregated kernel statistics can be easily extended to scales of more than two, given that the resolutions of data sets at different scales progressively increase. In practice, it is suitable for applications where data are progressively expanding along certain time periods. A high-order simulation program based on the above paradigm is developed and described. The simulation program is integrated into the SGeMS platform for a user-friendly

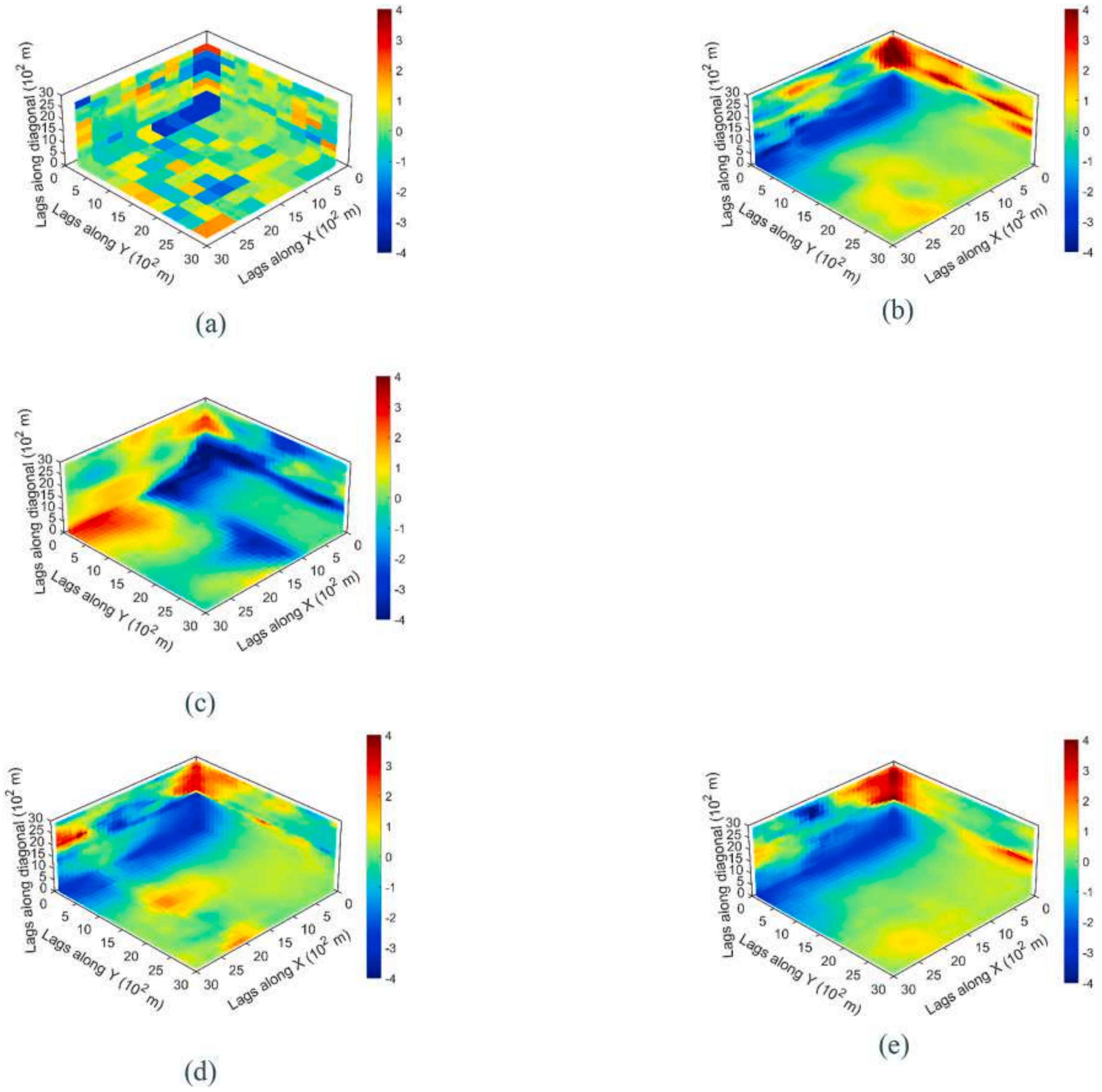


Fig. 7. Fourth-order cumulant maps of (a) sample data; (b) exhaustive image; (c) TI; (d) realization in Fig. 3a; (e) realization in Fig. 3b.

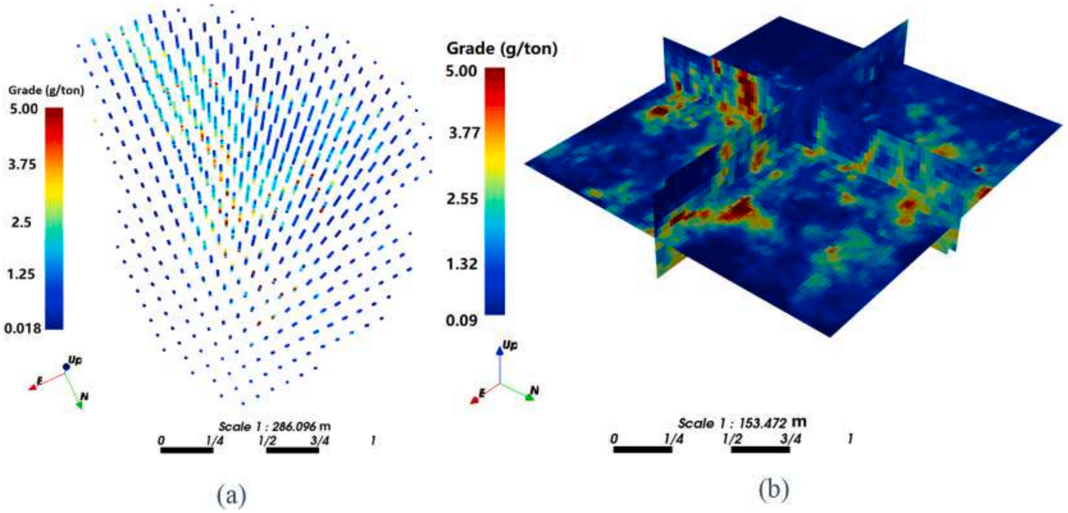


Fig. 8. (a) Drill hole samples at a gold deposit; (b) TI derived from the blasthole data in an adjacent area. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

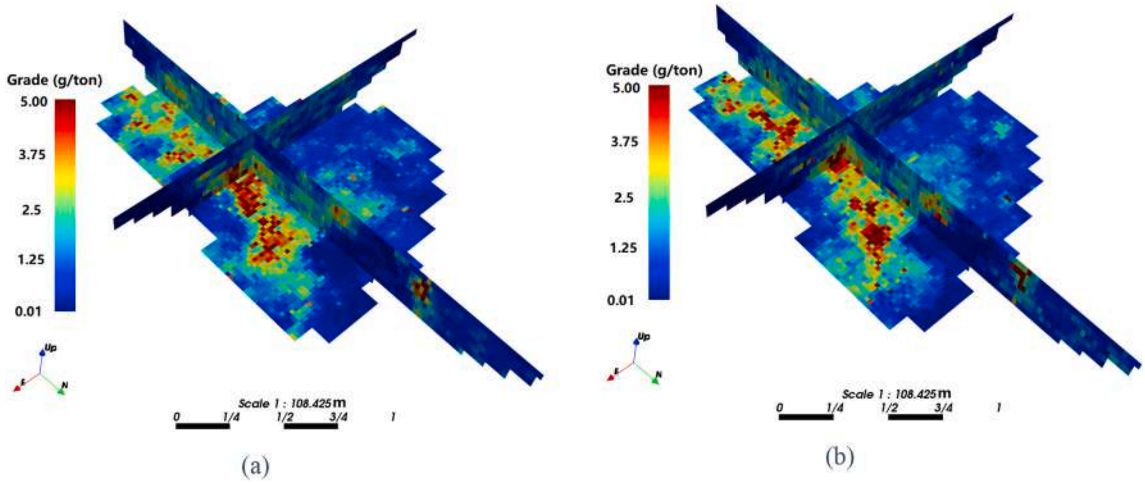


Fig. 9. Two simulated simulations using the sample data and the TI shown in Fig. 8.

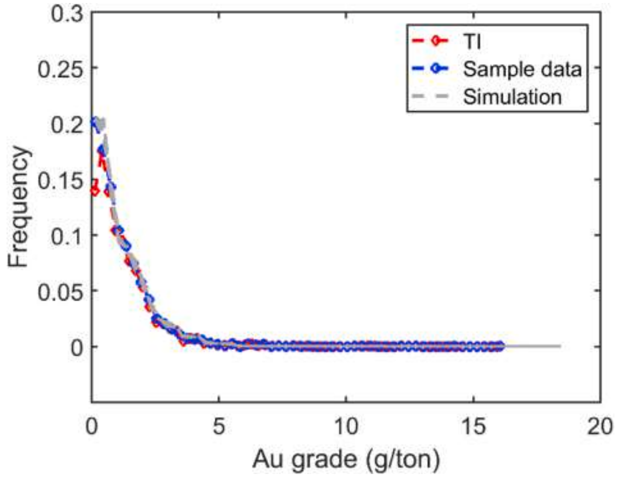


Fig. 10. Histograms of 10 simulated realizations using the samples and the TI shown in Fig. 8.

parameter selection and visualization in three-dimensional space. This simulation program is utilized here to carry out two different case studies. The first case study with the synthetic data set demonstrates the capacity of the proposed simulation method in reproducing the low- and high-order spatial statistics from the sample data, while significantly mitigating the statistical conflicts between the samples and the TI. The study at a gold deposit demonstrates the applied aspects of the simulation program when used to simulate pertinent properties of actual mineral deposits.

Computer code availability

- Name of code: kernelsim
- Developer: Lingqing Yao
- Contact details: COSMO – Stochastic Mine Planning Laboratory, Dept. of Mining and Materials Engineering, McGill University, 3450 University Street, Montreal, QC H3A 2A7, Canada
- E-mail: lingqing.yao@mcgill.ca
- Year first available: 2020
- Hardware required: Run on a computer with 4 cores (2.4 GHz each) and 8 GB.
- Software required: Needs SGeMS software
- Program language: C ++
- Program size: 122 kb
- Details on how to access the source code: the source files of kernelsim can be downloaded from github: https://github.com/yaolq/kernelsim

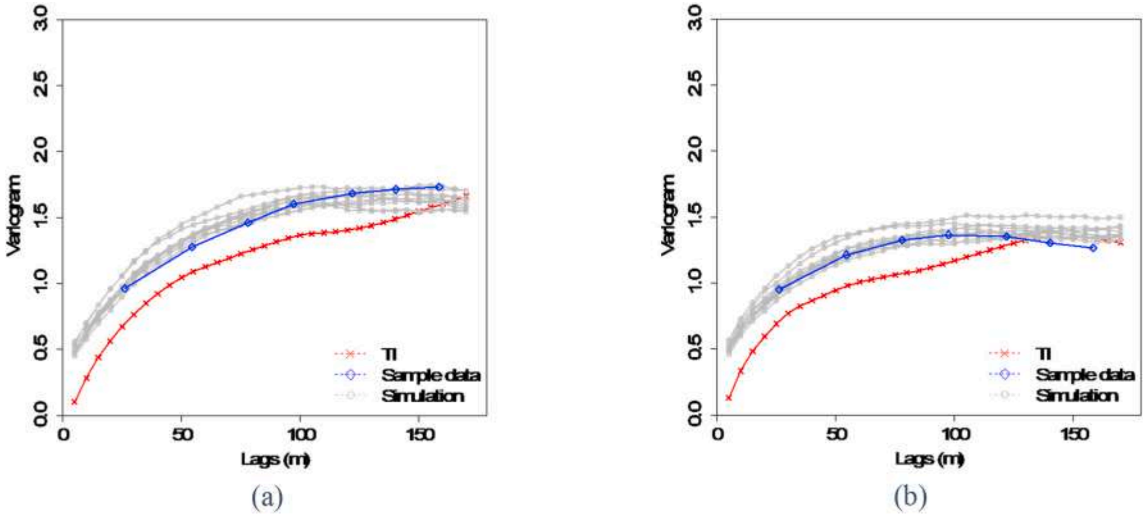


Fig. 11. Variograms of 10 simulated realizations along (a) E-W and (b) N–S direction, using the samples and the TI shown in Fig. 8.

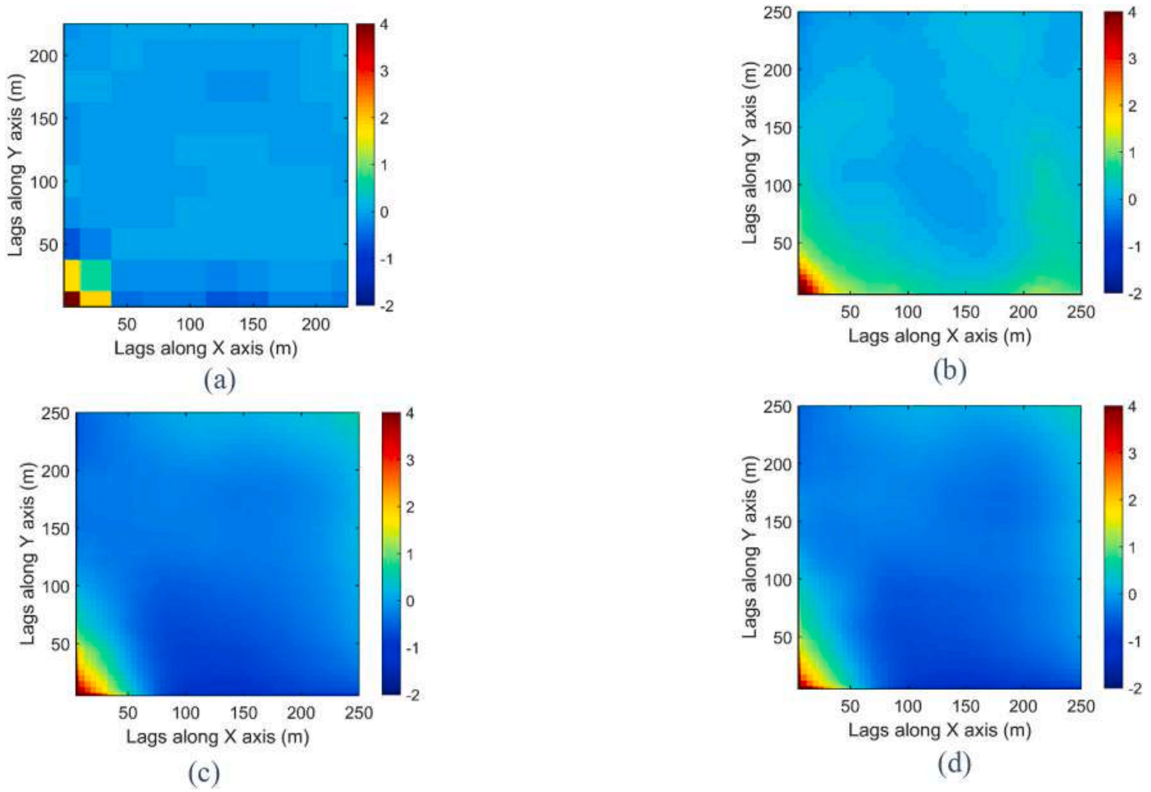


Fig. 12. Third-order cumulant maps of (a) sample data; (b) TI; (c) realization in Fig. 9a; (d) realization in Fig. 9b.

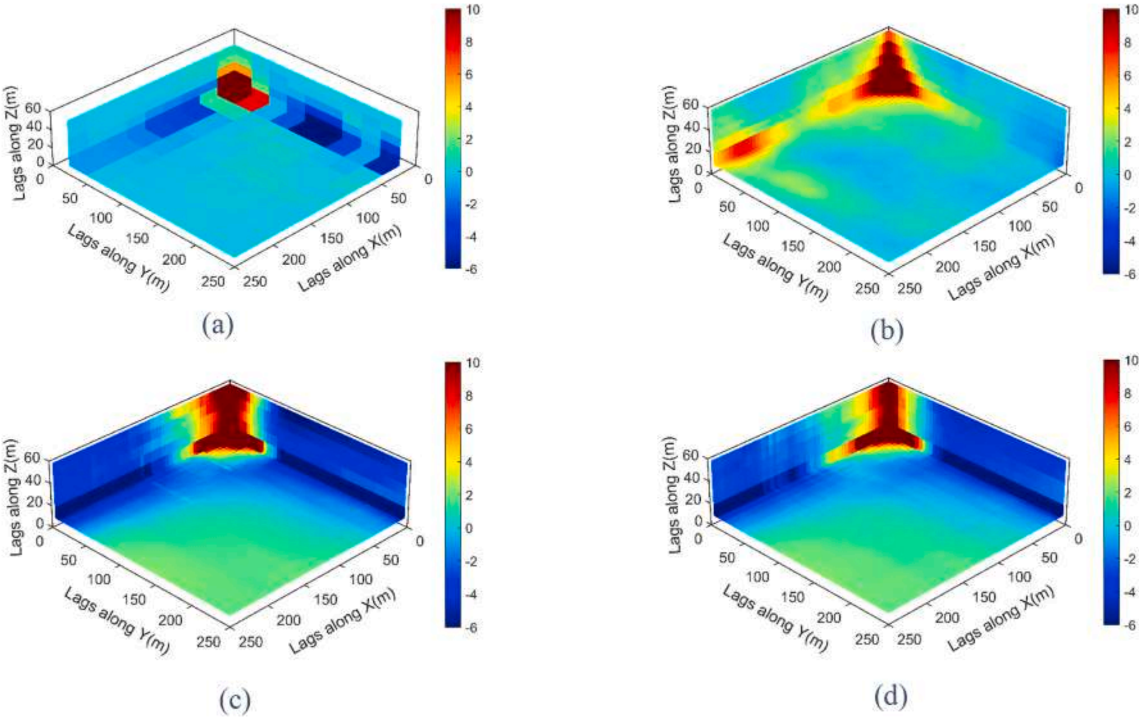


Fig. 13. Fourth-order cumulant maps of (a) sample data; (b) TI; (c) realization in Fig. 9a; (d) realization in Fig. 9b.

CRediT authorship contribution statement

Lingqing Yao: Conceptualization of this study, Methodology, Software, Writing - Original draft preparation, Writing – review and editing. **Roussos Dimitrakopoulos:** Funding acquisition, supervision, Writing - review and editing. **Michel Gamache:** Supervision, Writing - review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, CRD Grant CRDPJ 500414-16, the COSMO Mining Industry Consortium (AngloGold Ashanti, Anglo American, BHP, De Beers, IAMGOLD, Kinross, Newmont Mining and Vale), and NSERC Discovery Grant 239019.

References

Arpat, G.B., 2005. Sequential Simulation with Patterns. PhD. Stanford University, CA, USA.

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. Math. Geosci. 43 (4), 483–500. https://doi.org/10.1007/s11004-011-9326-9.

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. Math. Geosci. 42 (1), 65–99. https://doi.org/10.1007/s11004-009-9258-9.

Feng, W., Wu, S., Yin, Y., Zhang, J., Zhang, K., 2017. A training image evaluation and selection method based on minimum data event distance for multiple-point geostatistics. Comput. Geosci. 104, 35–53. https://doi.org/10.1016/j.cageo.2017.04.004.

Gonbadi, A.M., Tabatabaei, S.H., Fathianpour, N., 2019. A new multiple-point grade estimation method by implicit volterra series. Comput. Geosci. 129, 69–81. https://doi.org/10.1016/j.cageo.2019.05.005.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Applied Geostatistics Series. Oxford University Press, New York ; Oxford, England.

Guardiano, F., Srivastava, R.M., 1993. Multivariate geostatistics: beyond bivariate moments. In: Soares, A. (Ed.), Geostatistics Tróia `92, Vol 5. Quantitative Geology and Geostatistics. Kluwer Academic, Dordrecht, pp. 133–144. https://doi.org/10.1007/978-94-011-1739-5_12.

Hristopulos, D.T., 2015. Stochastic Local Interaction (SLI) model: bridging machine learning and geostatistics. Comput. Geosci. 85, 26–37. https://doi.org/10.1016/j.cageo.2015.05.018.

Journel, A.G., 2003. Multiple-point Geostatistics: A State of the Art. Stanford Center for Reservoir Forecasting, Stanford, CA, USA. Report no. 16.

Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, London; New York.

Mao, S., Journel, A., 1999. Generation of a Reference Petrophysical/seismic Data Set: the Stanford V Reservoir. 12th Annual Report. Stanford Center for Reservoir Forecasting, Stanford, CA, USA.

Mariethoz, G., Caers, J., 2014. Multiple-point Geostatistics: Stochastic Modeling with Training Images. Wiley, Hoboken.

Mariethoz, G., Renard, P., Straubhaar, J., 2010. The direct sampling method to perform multiple-point geostatistical simulations. Water Resour. Res. 46 (11), W11536. https://doi.org/10.1029/2008WR007621.

Martin, R., Boisvert, J., 2018. Towards justifying unsupervised stationary decisions for geostatistical modeling: ensemble spatial and multivariate clustering with geomodeling specific clustering metrics. Comput. Geosci. 120, 82–96. https://doi.org/10.1016/j.cageo.2018.08.005.

Minniakhmetov, I., Dimitrakopoulos, R., 2016. Joint high-order simulation of spatially correlated variables using high-order spatial statistics. Math. Geosci. 49 (1), 39–66. https://doi.org/10.1007/s11004-016-9662-x.

Minniakhmetov, I., Dimitrakopoulos, R., Godoy, M., 2018. High-order spatial simulation using Legendre-like orthogonal splines. Math. Geosci. 50 (7), 753–780. https://doi.org/10.1007/s11004-018-9741-2.

Mustapha, H., Dimitrakopoulos, R., 2010a. High-order stochastic simulation of complex spatially distributed natural phenomena. Math. Geosci. 42 (5), 457–485. https://doi.org/10.1007/s11004-010-9291-8.

Mustapha, H., Dimitrakopoulos, R., 2010b. A new approach for geological pattern recognition using high-order spatial cumulants. Comput. Geosci. 36 (3), 313–334. https://doi.org/10.1016/j.cageo.2009.04.015.

Mustapha, H., Dimitrakopoulos, R., 2011. HOSIM: a high-order stochastic simulation algorithm for generating three-dimensional complex geological patterns. Comput. Geosci. 37 (9), 1242–1253. https://doi.org/10.1016/j.cageo.2010.09.007.

Neves, J., Pereira, M.J., Pacheco, N., Soares, A., 2019. Updating mining resources with uncertain data. Math. Geosci. 51 (7), 905–924. https://doi.org/10.1007/s11004-018-9759-5.

Pérez, C., Mariethoz, G., Ortiz, J.M., 2014. Verifying the high-order consistency of training images with data for multiple-point geostatistics. Comput. Geosci. 70, 190–205. https://doi.org/10.1016/j.cageo.2014.06.001.

Rasera, L.G., Gravey, M., Lane, S.N., Mariethoz, G., 2020. Downscaling images with trends using multiple-point statistics simulation: an application to digital elevation models. Math. Geosci. 52 (2), 145–187. https://doi.org/10.1007/s11004-019-09818-4.

Remy, N., Boucher, A., Wu, J., 2009. Applied Geostatistics with SGeMS : a User`s Guide. Cambridge University Press, Cambridge, UK.

Remy, N., Shtuka, A., Levy, B., Caers, J., 2002. GSTL: the geostatistical template library in C++. Comput. Geosci. 28 (8), 971–979. https://doi.org/10.1016/S0098-3004(02)00021-3.

Scheidt, C., Caers, J., 2009. Representing spatial uncertainty using distances and kernels. Math. Geosci. 41 (4), 397–419. https://doi.org/10.1007/s11004-008-9186-0.

Scholkopf, B., Smola, A., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond. MIT Press, Cambridge, Mass.

Song, L., Zhang, X., Smola, A., Gretton, A., Schölkopf, B., 2008. Tailoring density estimation via reproducing kernel moment matching. In: Proceedings of the 25th International Conference on Machine Learning. ACM, pp. 992–999, 2008.

Steinwart, I., Christmann, A., 2008. Support Vector Machines. Springer, New York.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. 34 (1), 1–21. https://doi.org/10.1023/A:1014009426274.

Talebi, H., Peeters, L.J.M., Mueller, U., Tolosana-Delgado, R., van den Boogaart, K.G., 2020. Towards geostatistical learning for the geosciences: a case study in improving the spatial awareness of spectral clustering. Math. Geosci. 52 (8), 1035–1048. https://doi.org/10.1007/s11004-020-09867-0.

Yao, L., Dimitrakopoulos, R., Gamache, M., 2020. High-order sequential simulation via statistical learning in reproducing kernel Hilbert space. Math. Geosci. 52 (5), 693–723. https://doi.org/10.1007/s11004-019-09843-3.

Yao, L., Dimitrakopoulos, R., Gamache, M., 2021. Training image free high-order stochastic simulation based on aggregated kernel statistics. Math. Geosci. https://doi.org/10.1007/s11004-021-09923-3.

Zhang, T., Switzer, P., Journel, A., 2006. Filter-based classification of training image patterns for spatial simulation. Math. Geol. 38 (1), 63–80. https://doi.org/10.1007/s11004-005-9004-x.