Sultan Abulkhair, Nasser Madani

Stochastic modeling of iron in coal seams using two-point and multiple-point geostatistics: A case study

Mining Metallurgy and Exploration, 2022; 39(3):1313-1331

Stochastic modeling of iron in coal seams using two-point and multiple-point geostatistics: A case study

Sultan Abulkhair[1,2] and Nasser Madani[1]

[1]School of Mining and Geosciences, Nazarbayev University, Nur-Sultan, Kazakhstan.

[2]School of Civil, Environmental & Mining Engineering, The University of Adelaide, Adelaide, Australia.

Abstract

This work addresses the problem of quantifying iron content in a coal deposit in the Republic of Kazakhstan. The process of resource estimation in the mining industry usually involves building geological domains and then estimating the grade of interest within them. In coal deposits, the seam layers usually define the estimation domains. However, the main issue with the coal deposit in this study is that the iron dataset is solely based on data from three newly drilled drill holes located a significant distance apart and additional rock samples from stopes. A massive amount of geological information comes from legacy drill hole data sampled a long time ago, but there is no evidence of proper QA/QC being performed on those samples. For this reason, a workflow was introduced to construct a representative training image from legacy data and stochastically model geological domains within these three drill holes using a multiple-point geostatistics technique. Once the geological model was obtained, a two-point geostatistics algorithm was applied to model the iron inside each geological domain. The results showed that direct sampling (DeeSse) is a suitable multiple-point geostatistics algorithm that can reproduce the long-range connectivity and curvilinear features of seam layers. Furthermore, a sequential Gaussian simulation was used to model the iron in the corresponding domains. Both methods were extensively evaluated using different statistical tools and analyses.

Keywords: multiple-point statistics, direct sampling, training image, coal deposit, resource modeling, sequential Gaussian simulation

1 Introduction

The quality of coal is usually related to its composition. This refers to various minerals capable of either affecting coal utilization, improving gasification [1], and influencing its effectiveness as a heat energy source. In coal seams, certain elements, such as silicon, aluminum, sulfur, and iron, can significantly impact coal quality if their proportion is greater than 1% [2]. This can be explained by the wide variety of aluminum silicate and sulfide minerals present in Earth's crust. Particularly, iron-containing minerals play a significant role in coal quality. Iron minerals in coal can cause problems, including tube corrosion [3], boiler slagging [4], mine drainage damage [5], and health and environmental problems [6]. However, iron elements can also help with gasification [7] and geophysical monitoring due to their magnetic properties [8]. Frequently, a high accumulation of iron or other metalliferous minerals in coal can change its usage from one of heat energy to other industrial uses, such as manufacturing ceramic walls and tiles [9, 10].

Spatial modeling of iron grades is an essential component of coal resource estimation, which is used for downstream activities of coal mining projects, such as mine planning, coal preparation, and other analyses. In this context, geostatistical estimation and simulation algorithms are powerful techniques with which to model the iron in a coal deposit. These methods aim to use limited exploration information obtained from drill holes and other sources, such as geophysical investigations or even hand specimens, to produce unbiased and spatially reliable 2D or 3D models [11, 12]. Like other resource estimation workflows, a typical practice is to model the seam layers as an estimation domain and then separately model the mineral grades inside each domain. This method is known as cascade modeling [13–16]. Any two-point geostatistical algorithms can execute the second step, including deterministic and stochastic methods. However, deterministic geostatistical approaches overestimate low-grade and underestimate high-grade values (the smoothing effect), and screen out the influence of one data with another (the screening effect) [17]. On the other hand, the implementation of geostatistical simulations can overcome the smoothing effect and produce multiple unbiased scenarios. Popular univariate stochastic algorithms include the sequential Gaussian simulation (SGS) [18] and the turning bands simulation (TBS) [19, 20].

For modeling the seam layers, several deterministic and stochastic approaches exist. However, one of the challenges pertaining to coal deposits is that the seam layers represent complexity in the shape and stratigraphic positioning. In this respect, connectivity is a unique feature that manifests itself as long-range patterns, which require particular attention when opting for a stochastic geostatistical algorithm for modeling purposes. Two-point geostatistics, such as the sequential indicator simulation (SIS) [21, 22] and plurigaussian simulation [23], are inadequate for modeling such geological features with long-range connectivity. Compared to the traditional variogram based geostatistical approach, the multiple point geostatistics (MPS) have proven their applicability for modeling curvilinear patterns, especially in petroleum reservoir modeling, where the aim is to model the channelized reservoirs [24]. The premise of MPS-based approaches is to obtain spatial variability information from a conceptual training image (TI) instead of a two-point statistical function such as a variogram [25]. Doing so enables the multiple-point relation and complex curvilinear patterns that exist in the geological setting to be modeled [26]. Based on many examples of MPS applications, underinformed and overinformed cases are highly suited for MPS, while the covariance matrix cannot handle such datasets [27]. Several authors analyzed MPS and compared it to variogram-based methods by reviewing the available algorithms or applying statistical validation techniques [28–32]. Aside from petroleum and hydrogeology contexts, MPS is proven to be applicable in the mining industry, particularly in modeling slate deposits [33] and dykes in copper deposits [34].

There are several MPS-based algorithms. Since the first implementation of MPS by Guardiano and Srivastava [35], known as the extended normal equation simulation (ENESIM), many more advanced algorithms have been introduced. For example, adding a search tree into ENESIM addresses problems of CPU time limitations, which is the essence of the single normal equation simulation (SNESIM) [36]. Similar algorithms were also developed that simulate patterns instead of pixels, such as the filter-based simulation (FILTERSIM) [37] and simulation of pattern (SIMPAT) [38]. Recently, the application of machine learning in an MPS framework has become a focus of various research groups [39, 40]. This study was focused on applying one of the most recent MPS algorithms – direct sampling (DeeSse), which has similar features to both pixel-based and pattern-based methods [41]. DeeSse uses a distance function to scan the TI, and, unlike other MPS algorithms, it scans the TI directly, which significantly increases the simulation speed and lowers the load on memory. It has been modified and extended several times to overcome its limitations and enable its use in a broader range of cases [42–44].

2 Methodology and theoretical background

Cascade modeling workflow includes modeling geodomains and separately estimating the grade of interest inside each domain. These domains in coal deposits can be identified by seam layers, whereby the aim is to independently model the iron inside each domain (seam layer). In the former, an MPS-based algorithm known as DeeSse was selected for stochastic modeling of seam layers. For the latter, an SGS was introduced to model the iron inside each seam layer obtained from DeeSse. This enabled the quantification of the seam layer uncertainty and iron throughout the entire coal deposit. The following theoretical background and methodology were utilized for the training image, DeeSse, and SGS.

2.1 Training image

Every MPS-based simulation algorithm needs a training image (TI) to be inferred. The initial challenge is not the selection of a proper MPS algorithm but the construction of a convincing gridded training image that conceptually represents all the statistical features of the deposit. The main issue with gridded TI construction is that MPS requires it to be significantly larger than the target simulation grid [45]. The training images can be generated using physical principles, geostatistical algorithms, and regular statistics, and in some rare cases, using empirical methods [25]. According to Tahmasebi [31], various types of TI can be implemented using MPS algorithms, i.e., 2D images of outcrops or pictures of channels from airplanes [46, 47], gridded geological models of deposits with a similar geology [48], and conceptual models created using Boolean methods such as object-based [49] and process-based techniques [50]. There are many methodologies for constructing a TI, depending on the type and complexity of the deposit. For instance, the use of a deterministic geological block model is common in the mining industry [51–53]. These models can be obtained based on either a geological interpretation or following a model interpolated from the available data on the deposit, typically from exploration drill hole samples. These interpreted models provide a unique representation of the layout of the geological domains and boundaries within the deposit. The technique for establishing such models can be classified into two main families. The first involves hand contouring and wireframing [54–57], in which the polygons obtained from digitizing 2D cross-sections are connected to shape the ore body geometry into a 3D model. In contrast, indicator kriging [58] and the radial basis function (RBF) [59] are other deterministic alternatives that speed up manual digitization by the automatic generation of geological boundaries. Furthermore, different data types can be used to construct a TI, although there is a trade-off between building a geologically realistic conceptual model and generating a statistically representative one [31]. This depends on the availability of reliable information, wherein even historical data (i.e., legacy data) can be used to obtain a representative TI [25].

2.2 Direct sampling simulation algorithm

In contrast to other pixel-based MPS methods, the direct sampling algorithm uses a distance function in the TI scanning process. Moreover, DeeSse samples the TI directly, rather than using a conditional cumulative probability distribution function (cpdf) in every step. This makes the algorithm faster and means it does not need to store scanning results in a separate database.

The rationale of DeeSse follows the basic simulation principles of MPS, i.e., obtaining the conditional data around the simulated node, then sampling the TI and moving onto the next node. However, instead of computing the cpdf and sampling from the produced distribution, DeeSse randomly samples the conditional value from the TI. Therefore, the sampling process in DeeSse is faster, and it requires less memory than other pixel-based MPSs (e.g., ENESIM and SNESIM) [41]. This workflow is shown in Fig. (1).

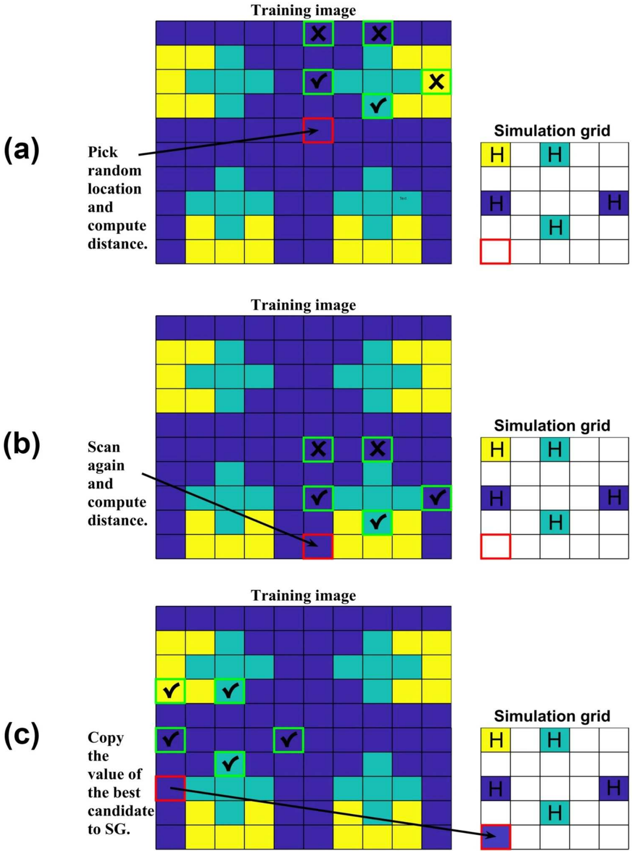

Training image
(a)
Pick random location and compute distance.
Simulation grid
H
H
H
H
H
Training image
(b)
Scan again and compute distance.
Simulation grid
H
H
H
H
H
Training image
(c)
Copy the value of the best candidate to SG.
Simulation grid
H
H
H
H
H

Fig. 1 Step-by-step illustration of the DeeSse algorithm. "H" – hard data; "✓" – node matches hard data; and "✗" – no match is found

The distance function is implemented to calculate the percentage mismatch between TI nodes and conditional data in the simulation grid (SG) [41]:

$$d\{d_n(x), d_n(y)\} = \frac{1}{n}\sum_{i=1}^{n} a_i \quad a_i = \begin{cases} 0 \text{ if } Z(x+h_i) = Z(y+h_i) \\ 1 \text{ if } Z(x+h_i) \neq Z(y+h_i) \end{cases} \tag{1}$$

where $d_n(x)$ and $d_n(y)$ are data events in SG and TI, respectively, and h_i is a lag distance.

Scanning for each node does not finish until the match between the TI and SG data events has occurred. In other words, until distance d is less than its acceptance threshold t. This can be expressed as $d\{d_n(x), d_n(y)\} < t$. The acceptance threshold here is a normalized parameter between 0 and 1, where 0 is for the same patterns and 1 is for completely different ones.

However, if there is no such data event with a computed distance less than the proposed acceptance threshold, sampling is stopped after scanning a certain fraction f of the TI. Then, the node with the lowest possible distance is copied from the TI to the SG target node as the best candidate. For detailed information related to the DeeSse algorithm, readers can refer to Mariethoz et al. [41]. In a nutshell, the DeeSse algorithm can be summarized as follows:

1. Assign conditioning points to their respective grid locations u_i;
2. Visit all locations u_i via a predefined simulation path (the path can be random or regular);
3. At each location u_i, find the neighborhood of nearest neighbors to u_i;
4. Scan the TI randomly and use the distance function $d\{d_n(x), d_n(y)\}$;
5. Assign a value of a data event such that $d\{d_n(x), d_n(y)\} < t$;
6. If step e is unsuccessful, assign another value with the smallest distance once the f fraction of TI is scanned;
7. Loop steps c–f for all locations u_i.

A maximum scan fraction f is a helpful parameter with which to avoid long searches, thus accelerating the scanning process. The whole fraction is only scanned when there are no nodes with a distance less than an acceptance threshold t. Another essential feature of this algorithm is the postprocessing to remove noisy data in the realizations, which is conducted by flagging unsuccessful nodes ($d > t$) and simulating them again.

2.3 Sequential Gaussian simulation

The steps involved in the sequential Gaussian simulation are described as follows [18]:

1. Transform the variable Y_j into its normal score Z_j, where j indicates the variable in case of multiple variables being available.
2. Assess the obtained normal scores for bivariate and multivariate Gaussianity.
3. Define a random or regular simulation path so that each grid node x_i is visited only once.
4. At each node x_i:
   Use simple kriging to determine the global parameters of the Gaussian conditional cumulative distribution (ccdf)

$$Z_j^n(x_i) = Z_j^{SK}(x_i) + \sqrt{\sigma_j^{SK}(x_i) U_i^n} \tag{2}$$

$$Z_j^{SK}(x_i) = \left[1 - \sum_{\alpha=1}^{k} \lambda_\alpha\right] m_j + \sum_{\alpha=1}^{k} \lambda_\alpha Z(x_i) \tag{3}$$

$$\sigma_j^{SK}(x_i) = C_j(0) - \sum_{\alpha=1}^{k} \lambda_\alpha C_j(x_\alpha - x_i) \tag{4}$$

where $Z_j^{SK}(x_i)$ is a simple kriging estimator, $\sigma^{SK}(x_i)$ is kriging variance, U_i^n is an independent random number generated between 0 and 1, λ_α is the weight assigned at location α, m_j is the mean value of the variable $Z_j, x_\alpha(\alpha = 1, \ldots, k)$ is the data location, and C_j is the covariance.

5. Loop until all grid nodes are simulated.
6. Apply step 4 to simulate the next variable (if applicable).
7. Back-transform simulated normal score values Z_j^n to the original scale Y_j^n.

3 Case study — coal deposit

3.1 Geological description of the study area

A coal deposit located approximately in the center of the Republic of Kazakhstan was selected as the case study. The relief of the deposit area is primarily flat, with surface elevation in the range of 450–490 m. The deposit basin is asymmetrical: the long axis is about 12 km in length, and the width is between 6 and 7 km. This is a shallow deposit with a maximum depth of 150 m. Fig. (2) represents the relative location and cross-section of the deposit, showing the essential features discussed in this study. The dip angles vary significantly over the whole basin area; however, a section of it, which is currently under operation, has a consistent strata dip of 30-40 degrees. There are three coal-bearing horizons in a Jurassic formation: upper, middle, and lower, but only the former is under production, using the open-pit method. The deposit's name cannot be disclosed for reasons of confidentiality [60].

The lower horizon is up to 50 m thick and consists of six seams, each about 0.2–1.5 m thick, while the middle horizon is only 2.8 m thick. The upper horizon has two main thick seams: 2b and 1b seam layers with thicknesses of 12.8-21.9 m and 8.3-12.0 m, respectively. Sediment input during coal deposition can be used to explain the several interlayers and low coal content on the left side of the cross-section (Fig. 2). Therefore, the original dataset consists of seams, such as 2b4, 2b3, 1b2, and other interlayers that are parts of the 2b and 1b seam layers. In this study, these interseams are considered as only 2b and 1b in order to reduce the number of possible categorical variables from about 30 to 3, i.e., the thin shale layer, and the 2b and 1b seam layers.

3.2 Training image

In this deposit, records of past geological exploration activities are primarily handwritten and do not present sufficient evidence of proper QA/QC. However, even though the validity of this dataset is questionable, a geological model produced from legacy data can act as a representation of the deposit's geology.

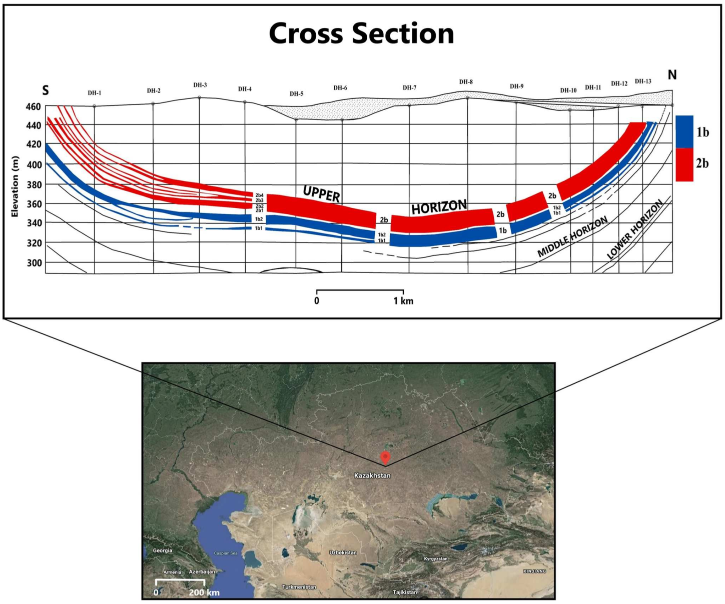


Fig. 2 Approximate location of the study area and cross-section of the coal deposit

In this regard, the aim of MPS is not to replicate TI but rather to reproduce higher-order statistics from it. Therefore, this study used an interpretive geological model of the seam layers obtained by wireframing as a TI.

Seam number data were combined into three categories, but each interlayer was named differently in the original dataset by adding an index to the main layer, e.g., 2b4 or 1b2 (Fig. 2). Handling and merging the interlayers into the 2b and 1b seams reduced the number of possible categories. As a result, the categories for the TI can be identified as follows:

$$Z(u) = \begin{cases} 3 \text{ if node belongs to 1b seam or its interlayers} \\ 2 \text{ if node belongs to 2b seam or its interlayers} \\ 1 \text{ if node belongs to shale} \\ 0 \text{ if node does not belongs to any seam number} \end{cases} \tag{5}$$

The last category (0) was considered waste and represents all surrounding rocks in the final block model. Moreover, there were undefined zones between layers that can affect the MPS simulation; thus, all undefined nodes were assigned to the waste category. Fig. (3) shows a 3D view of the TI used in this study, which is massive in terms of its real size and the number of nodes, making it suitable for use in the MPS simulation [25]. The training image is available in the GitHub repository, and its link is available in the Supplementary Information section.

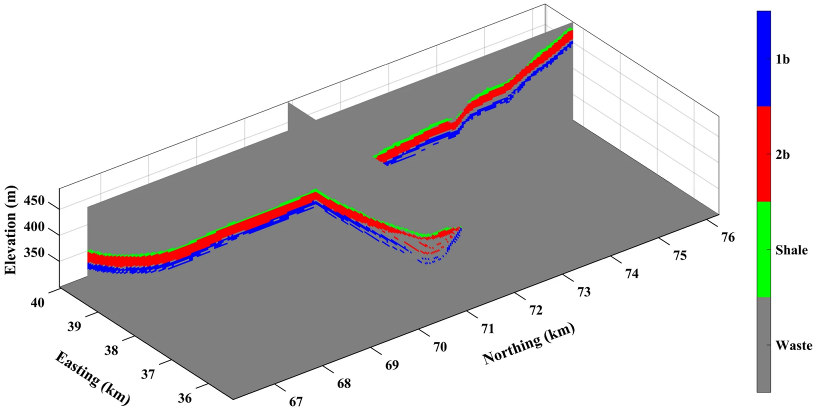


Fig. 3 3D representation of the training image for this study, illustrating the stratigraphic model of the seam layers

3.3 Hard data

Reliable exploration data in this deposit were solely obtained through three recently drilled drill holes and so are limited. Therefore, the greatest challenge in this study was the scarcity of the conditioning data for modeling iron in the seam layers. The three drill holes are located at a significant distance from each other (Fig. 4 a). The categorical proportions in the three drill holes are: waste: 72.92%; shale: 3.22%; 2b seam: 15.55%; and 1b seam: 8.31%. To address the data scarcity, additional samples from stopes consisting of both seam layers (only 1b and 2b without shale) and iron grade information were used. Therefore, a total of 13,379 sample points was divided into two groups based on the seam layer category, i.e., iron in the 1b seam and iron in the 2b seam. Fig. (4 b and c) demonstrates the location map of the conditional hard data, in which the target SG is depicted with a dashed line.

The cell-declustering [61, 62] technique was used to address the problem of widely dispersed drill holes and densely distributed stope samples. A cell size of 80 m × 80 m × 8 m was selected for declustering after checking the effect of different cell dimensions on the global statistics. As a result, the mean grades of the iron in the 1b and 2b seams increased after declustering, while their variances decreased (Table 1). Overall, the average iron grade and its variance in 1b were considerably higher as compared to those in the 2b seam layer.

3.4 Geostatistical simulation of layers

Geostatistical simulation algorithms produce the uncertainty model. This can be characterized as multiple sets of possible values regularly distributed in the region under study. One set of these possible outcomes is a realization, and the number of realizations should be large enough to produce a reliable assessment of joint uncertainty [63]. Therefore, this number was selected to be 100 for this study.

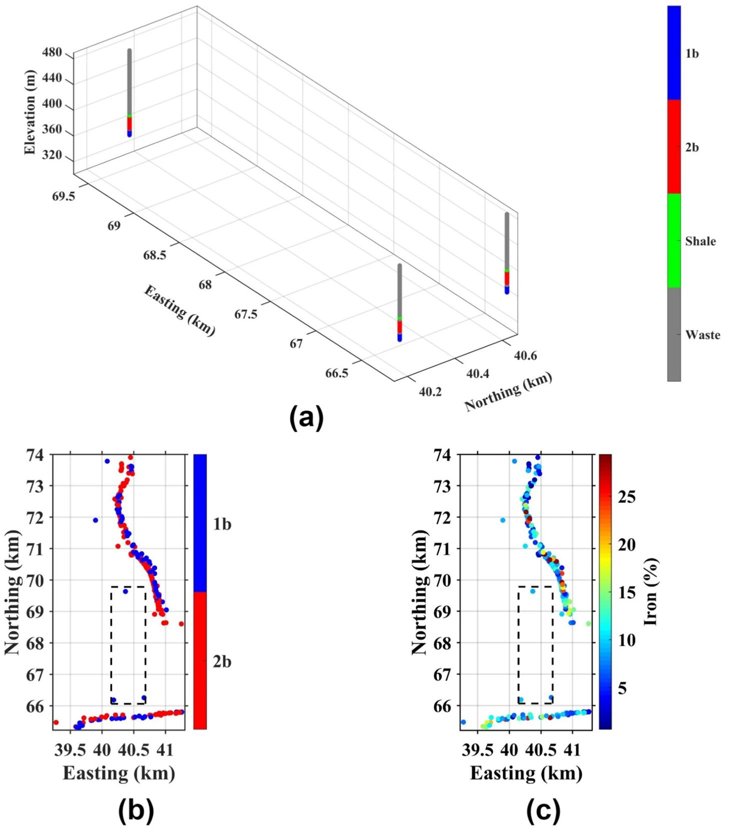


Fig. 4 Location maps of (a) three drill holes inside SG, (b) seam layers and (c) iron grades. The dashed line represents the target SG

Table 1 Summary statistics of the iron dataset before and after declustering (grades are expressed in %)

		Before declustering		After declustering	
Group	Number of samples	Average grade	Variance	Average grade	Variance
Iron in 1b	3,212	13.00	58.52	13.43	58.06
Iron in 2b	10,167	8.68	19.80	8.77	19.36
Total	13,379	9.72	32.49	10.17	35.64

The stochastic modeling of seam layers was performed by the DeeSse algorithm using the Ar2GEMS software. Simulation parameters were chosen based on the instructions given in Meerschman et al. [64] and the notes in Mariethoz et al. [41]. Since the TI was quite large and taking into account the restriction in the resource computations, in order to reduce the simulation time, a maximum scan fraction f of 0.5 was considered, meaning that only half of the TI was scanned to find a node with the smallest distance $d\{d_n(x), d_n(y)\}$. To choose the acceptance threshold t, a sensitivity analysis was implemented for this parameter over different values of 0.01, 0.1, 0.2, and 0.3. For this purpose, 10 nonconditional realizations were produced by DeeSse using each different threshold over the same target grid. Then, the indicator covariance was calculated for each realization and then averaged over the lag distances. The results showed that the low values of t produce better connected patterns, while high values of t produce patchy and unstructured results. The lower the t, the more structured the indicator covariance, which implies a superior reproduction of connectivity. An example of this graph is shown in Fig. 5 for shale (see Fig. A1 in Appendix A to find the same graphs for the 2b and 1b seam layers). Therefore, in this study, the acceptance threshold t was selected as 0.01. This is in accordance with the statement of Meerschman et al. [64], in which they suggest making this value as small as possible in order to reproduce the geological domains with long-range connectivity.

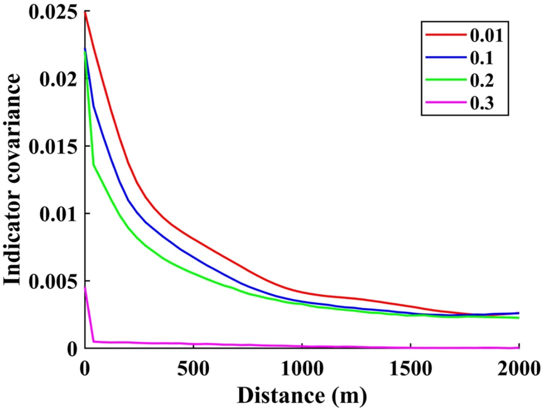
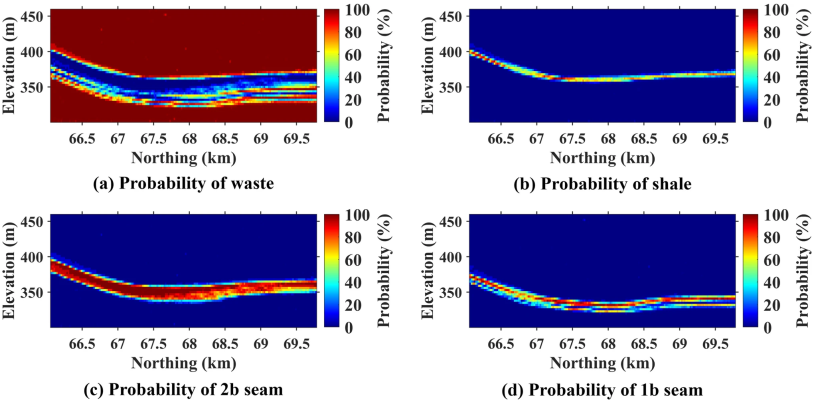


Fig. 5 Indicator covariances of the shale in the northing direction using different acceptance thresholds. For brevity, only the indicator covariance of the shale is presented

The produced realizations show that DeeSse reproduced the layered structure pattern of the TI (see Fig. A2 in Appendix A to find cross-sections of the four random DeeSse realizations). However, as compared to the other layers produced in the simulation results, the successful reproduction of seam 1b was challenging as the layer is disrupted and consists of multiple interlayers. Furthermore, probability maps were produced by calculating the probabilities of categories in each node over multiple realizations (Fig. 6). As can be seen, the probability maps obtained from 100 realizations prove that the layer patterns remain consistent throughout all realizations. Moreover, it can be stated that the desired patterns maintained their connectivity, shapes, and dimensions.

The most probable map was also calculated using probability maps to discover the potential deterministic positioning of the seam layer boundaries (Fig. 7 a). The most probable values were obtained from the probabilities of each category. For instance, if the probability of the 1b seam was the highest



(a) Probability of waste

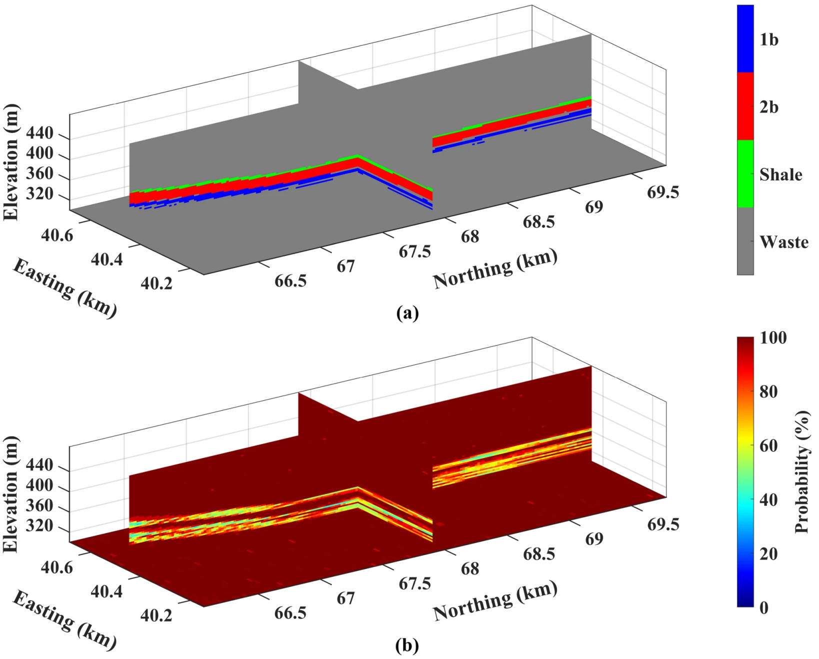
(b) Probability of shale

(c) Probability of 2b seam

(d) Probability of 1b seam

Fig. 6 Probability maps of (a) waste, (b) shale, (c) 2b seam, and (d) 1b seam in a cross-sectional view

among all categories in one particular node, then the category 1b seam was assigned to this node. The same process was applied to waste, shale, and the 2b seam if they had the highest probability. Additionally, Fig. (7 b) shows the probabilities of the most probable categories. For example, a thin shale layer had lower probabilities than the 2b seam, while the 1b seam demonstrated lower probabilities because of the multiple interlayers in the realizations.



(a)

(b)

Fig. 7 3D view: (a) most probable model over 100 realizations of DeeSse and (b) probability of the most probable category

3.5 Statistical validation

A critical feature of this deposit is that seam layers are inconsistent and divided into interlayers, as stated in the geological description of this deposit. Therefore, local proportions of the TI vary drastically in different parts of the deposit, while the conditional data are not dense enough to derive proportions from it. For this reason, this information was not utilized during the DeeSse simulation. Nevertheless, the proportion or histogram reproduction is the part of simulation validation that compares first-order statistics of the TI with the produced realizations. In this respect, the waste category in the TI and SG was not considered in the validation or further analyses because the waste category was not considered for modeling the iron. Fig. (8) plots the seam layer proportions over 100 realizations, where dashed lines represent the proportions of each particular seam layer in the TI, while dotted line denote the proportions obtained from hard data. As can be seen, the proportions of each layer fluctuate in a similar manner to the original proportions from the TI and hard data. These fluctuations are related to the ergodic property [17] of each seam layer and allow one to indirectly quantify the uncertainty of the original proportions. Furthermore, proportions of 2b seam produced by DeeSse fluctuate slightly above TI proportion, while it is opposite for 1b seam and shale. The possible reasons for this are limited input data and larger thickness and continuity of 2b seam compared to other layers.

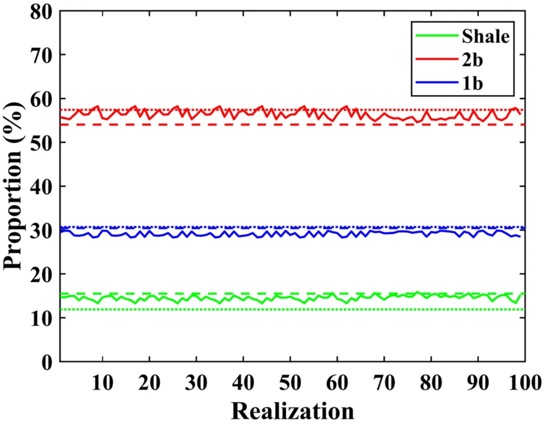


Fig. 8 Proportions of the DeeSse realizations in comparison with the TI and hard data. Dashed lines belong to TI proportions and dotted lines are proportions derived from the three drill holes

It is also of interest to check whether or not the simulation algorithm can reproduce the spatial continuity of the seam layers imposed by the TI. Experimental variograms of 100 realizations in the northing direction were computed and are illustrated in Fig. (9 left). The spatial continuity of each seam layer introduced by their corresponding indicator variograms roughly follows the spatial continuity of the TI. However, the TI is generated through an interpretive geological model obtained using a wireframing approach, where no variogram model or other spatial continuity tools were involved. Therefore, it is not necessary that the produced realizations exactly mimic the spatial continuity of the TI.

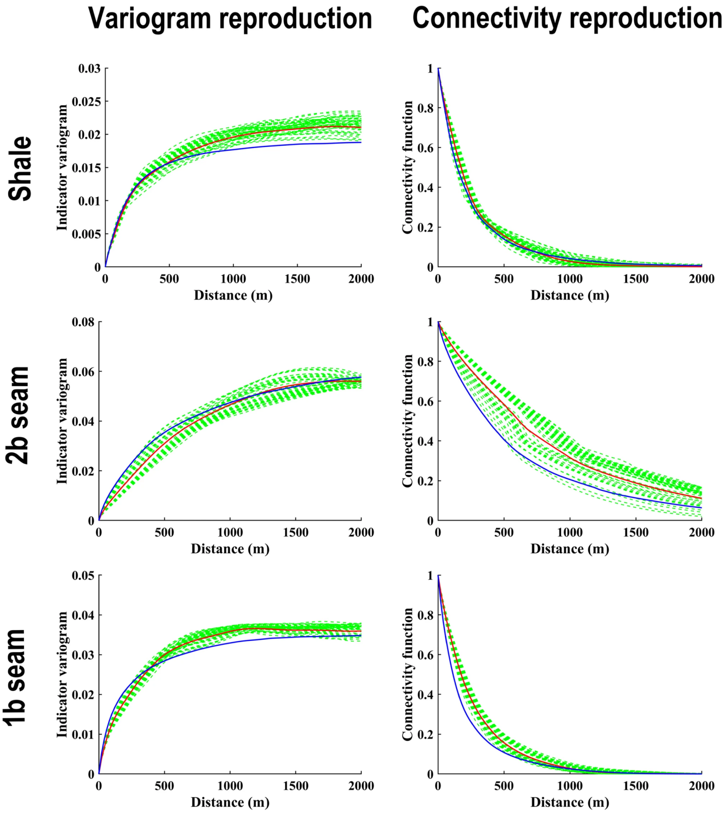


Fig. 9 Indicator variograms and connectivity functions of shale and the 2b and 1b seams in the northing direction for variogram and connectivity validation. Green: 100 realizations; red: average over realizations; and blue: TI

The main advantage of multiple-point geostatistics over variogram-based methods is the ability to reproduce curvilinear geological patterns. This can be expressed in the connectivity of the pattern, in which the results can be validated mathematically and statistically. In brief, connectivity is the probability that two separate nodes are connected in a particular direction [65]. One way to inspect pattern connectivity is to compute connectivity functions [66]. Connectivity functions of shale and the 2b and 1b seams in the northing direction are shown in Fig. (9 right) and compared to the connectivity functions from the TI. It can be seen that the results obtained from DeeSee are in reasonable agreement with the TI in this particular connectivity analysis.

However, DeeSse did not seem to reproduce the 2b seam connectivity of the TI correctly and slightly overestimated the connectivity of the 1b seam.

Each seam layer in this deposit represents a unique and independent pattern with extended structural connectivity. Therefore, their connectivity and variograms were analyzed over all the realizations and for each category separately. This type of validation makes sense since the data from the three drill holes and the TI belong to the same coal deposit. Therefore, it is expected that the simulated results approximately follow the spatial continuity and multiple-point relationship that inherently exists in the TI.

3.6 Simulation of iron in seam layers

After completing seam layer modeling, the iron grades in corresponding layers were identified. Iron in this deposit is disseminated throughout the coal layers and is not shown as a separate layer. Therefore, its separation from coal needs further processing after the exploitation of the coal layers. For simulation implementation, the spatial continuity of the variables must be derived. Before this, however, it is necessary to identify the presence of anisotropy. For this purpose, different directions were examined to determine potential anisotropy. As a result, an anisotropic variogram was placed in the horizontal and vertical directions. Finally, after transforming both original variables (i.e., $Y_{Fe_{1b}}$ and $Y_{Fe_{2b}}$) independently into normal scores (i.e., $Z_{Fe_{1b}}$ and $Z_{Fe_{2b}}$), sample variograms were identified in the horizontal and vertical directions and then were fitted automatically to obtain the corresponding theoretical variograms (see Fig. A3 in Appendix A to find fitted theoretical variograms):

$$\begin{pmatrix} \gamma_{Z_{Fe_{2b}}} \\ \gamma_{Z_{Fe_{1b}}} \end{pmatrix} = \begin{pmatrix} 0.50 \\ 0.36 \end{pmatrix} Exp\left(250m, 250m, 23m\right) + \begin{pmatrix} 0.14 \\ 0.20 \end{pmatrix} Exp\left(735m, 735m, 33m\right) \tag{6}$$

Before proceeding with simulation, cross-validation was implemented to ensure the unbiasedness of the covariance functions. The cross-validation procedure involves removing the actual data points one by one and finding the predicted value using the rest of the data [58]. Predicted and actual data are then compared to find the coefficient of variation R^2 and mean squared error MSE (Fig. 10 a). The SGS cross-validation results were also compared to the estimation results using simple kriging (Fig. 10 b). For the sake of comparison, an e-type of the predicted results was used to demonstrate the SGS cross-validation. Overall, both methods demonstrated reasonable validity, and the difference in MSE was very low, i.e., 0.21 for the simulation and 0.20 for the estimation. Furthermore, uncertainty was also validated using an accuracy plot [67] between probability intervals and fractions of data that belong to those intervals (Fig. 10 c). Considering the data scarcity, the validation of uncertainty demonstrated an acceptable match. However, the fact that the accuracy plot is slightly below the identity line is a sign that the uncertainty model is slightly inaccurate.

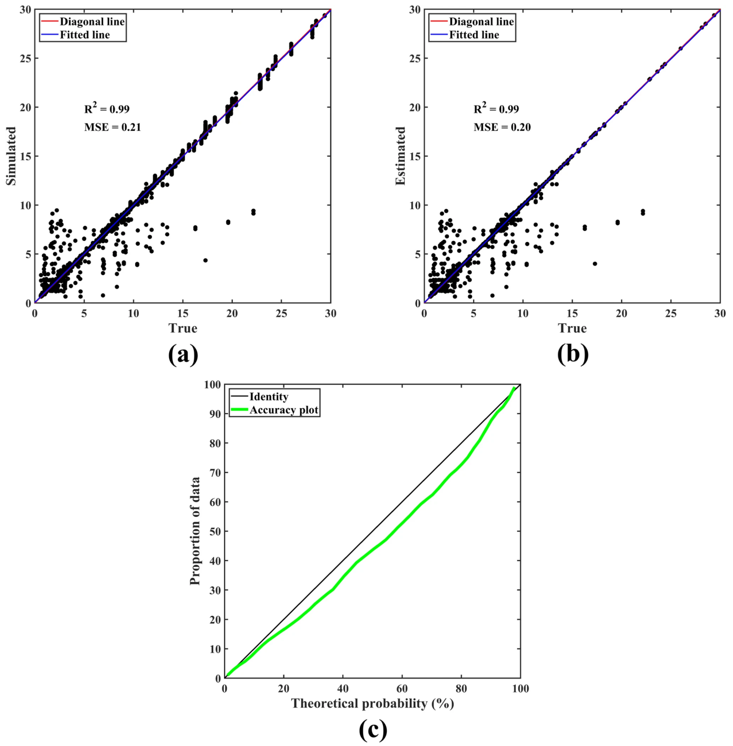


Fig. 10 Cross-validation: predicted vs. actual plots for (a) simulated results (e-type of 100 realizations from SGS), (b) estimated results, and (c) accuracy plot

3.7 Cascade modeling

To combine the categorical and continuous variables, cascade modeling was implemented. To do so, simulated iron in 1b and iron in 2b were juxtaposed into each target location x_0 for each realization with the corresponding modeled seam layers in the following fashion:

$$Y_i\left(x_0\right) = \begin{cases} Y_{Fe_{2b}}(x_0) \text{ if location belongs to 2b seam} \\ Y_{Fe_{1b}}(x_0) \text{ if location belongs to 1b seam} \\ NaN \text{ if node is undefined or belongs to shale} \end{cases} \quad (7)$$

where Y_i is the final simulated iron value for each realization i at the target location x_0, and $Y_{Fe_{2b}}(x_0)$ and $Y_{Fe_{1b}}(x_0)$ are simulation results of iron in the 1b and 2b seams, respectively.

After both iron in 2b and iron in 1b were independently simulated using SGS, cascade modeling was performed by juxtaposing each SGS realizations into the corresponding coal seams in DeeSse realizations (see Fig. A4 in Appendix A to find four random realizations of cascade modeling). Table (2) compares the statistical parameters obtained from cascade modeling with the original declustered parameters. As observed, cascade modeling can reproduce statistical parameters similar to those found in the original data.

Table 2 Statistical parameters of cascade modeling (iron grades are expressed in %)

Parameter	Realizations	Original dataset
Mean	10.05	10.17
Variance	34.14	35.64
COV	0.59	0.58

The most probable e-type map was produced by averaging 100 simulation results and juxtaposing them into the most probable seam layer model. Fig. (11) shows the produced model. It can be seen that the lower layer (1b seam) has significantly higher iron grades as compared to the upper layer (2b seam).

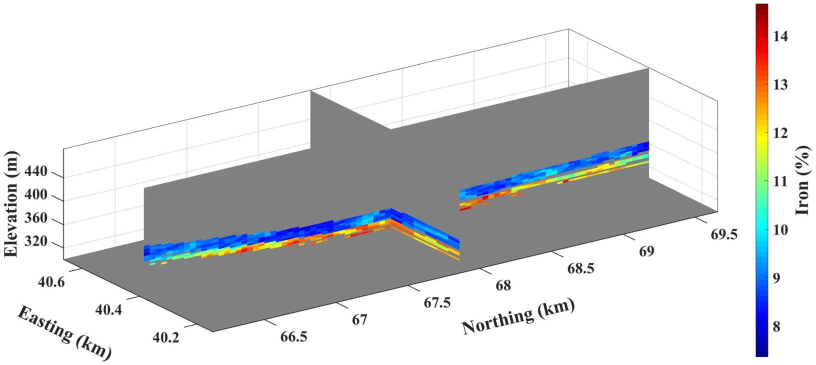


Fig. 11 Most probable e-type model over 100 realizations of DeeSse and SGS in 3D view

Distributions of realizations produced by cascade modeling and the original distribution of iron were compared using quantile-quantile (Q-Q) plots (Fig. 12). Produced Q-Q plots show that realizations are close to the diagonal line with significant deviations in the upper quantiles. However, this can be explained by the scarcity of high-grade samples, particularly inside the simulation grid.

The reproduction of spatial continuity for the final iron model after cascade modeling is presented in Fig. (13). The average and confidence limits (± 2 standard deviations around the average) of the indicator variograms calculated over the 100 realizations are also depicted. This step is necessary to examine whether the proposed algorithm can reproduce the local statistical variability of iron in the region. If, on average, the variogram of the realizations matches the original variogram, it signifies that the proposed algorithm worked adequately [68]. This figure shows that although, on average, the variogram converges to the original experimental points, the fluctuations depicted

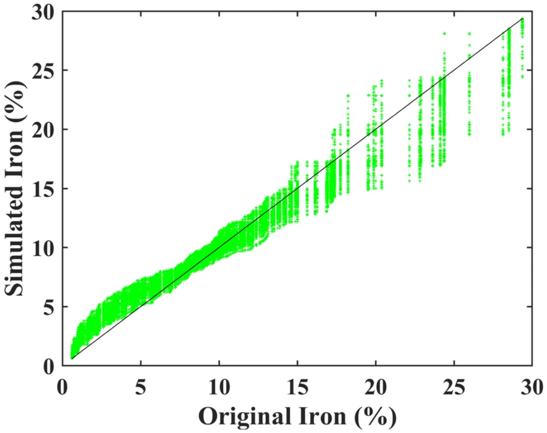


Fig. 12 Quantile-quantile plots between original and simulated iron grades. Green points: realizations; black line: identity line

in the variograms of individual realizations are high and slightly exceed the confidence interval. There are various reasons for these relatively high fluctuations, which are also referred to as ergodic fluctuations [17], i.e., 1) DeeSse uses a sequential simulation paradigm, in which large fluctuations are expected [58]; and 2) the density of the conditioning data in this study is small. Furthermore, the more conditioning data used in constructing the realizations, the fewer fluctuations expected [11].

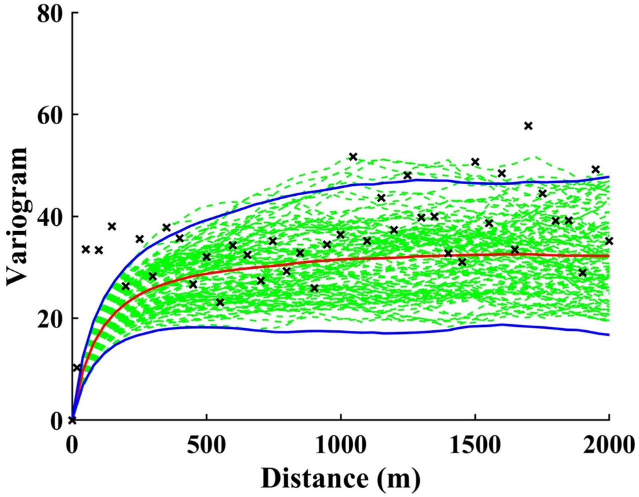


Fig. 13 Variogram reproduction of the proposed cascade modeling in the northing direction. Green: variograms of 100 realizations; red: average over realizations; blue: confidence intervals (mean ± 2 std); crosses: experimental variogram in the same direction

4 Discussion and conclusion

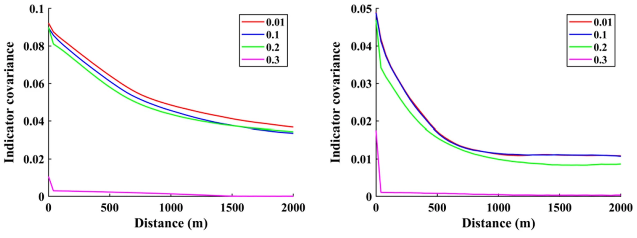
The purpose of this study was to model the layered structure of coal seams with long-range connectivity using a multiple-point geostatistical simulation and the obtained realizations as stochastic domains for the cascade modeling of iron. In this case study, the reliable data come from three drill holes located a significant distance from each other. Moreover, a legacy dataset was available with a large number of samples taken a long time ago. Therefore, this dataset was used to build a TI, which is a representative interpreted model of the whole deposit. The main reasons for using MPS instead of traditional geological modeling techniques were the scarcity of reliable drill hole data and no evidence of QA/QC in the legacy dataset. In this regard, MPS can sample key statistical parameters from the interpretive TI without replicating the exact model [25] and work with underinformed data [27].

The direct sampling MPS method was used in this study because of its ability to implement a distance function and sample directly from the TI, which increases computation speed and decreases the load on memory. DeeSse can reproduce the proportions, indicator variograms, and connectivity of seam layers to a certain extent. However, the results showed an overestimation of connectivity for the 2b seam as compared to the TI, which can be addressed by using more hard data during the conditional simulation.

The second part of the study involved the cascade modeling of the continuous variable, i.e., the iron grade. A sequential Gaussian simulation was used to independently model 100 realizations of the iron in the 1b and 2b seams. These iron realizations were then juxtaposed in their corresponding seam layers from DeeSse realizations to obtain the final iron model for the coal seams. The final realizations were assessed based on cross-validation, reproduction of global statistical parameters, and histogram and variogram validations. Despite the scarcity of conditional data, the iron simulation results demonstrated an acceptable quality with good reproduction of histogram and spatial continuity.

Overall, MPS can produce geologically realistic seam layers with acceptable reproduction of first, second, and higher-order statistics. First-order statistics were assessed by checking the reproduction of seam layer proportions and the second-order statistics were assessed by variogram validation. The connectivity functions were used to assess the higher-order statistics by checking the reproduction of curvilinear patterns. Moreover, in the proposed combination with the SGS algorithm in the cascade modeling framework, multiple reliable and unbiased realizations can be obtained for use in further mining processes. Nevertheless, DeeSse algorithms produce few noises and inconsistencies, which image cleaning approaches can remove. In addition, the lack of proper dip and connectivity reproduction can be avoided by using more informed sampling patterns as hard data. It is also recommended to compare the DeeSse simulation with pattern-based MPS methods, such as SIMPAT and FILTERSIM, in order to check the validity of the variogram and connectivity reproduction obtained in this study. However, more dense conditional data are required to make a fully valid comparison. The application of MPS methods to model seam layers is not well researched; for further studies, the training image used in this study is available as supplementary information through the GitHub link.

Appendix A Additional figures



(a) 2b seam (b) 1b seam

Fig. A1 Indicator covariances of (a) the 2b seam and (b) the 1b seam in the northing direction using different acceptance thresholds

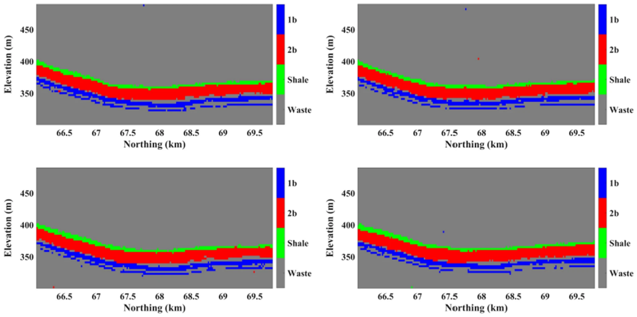


Fig. A2 Four random realizations of seam layers produced by DeeSse

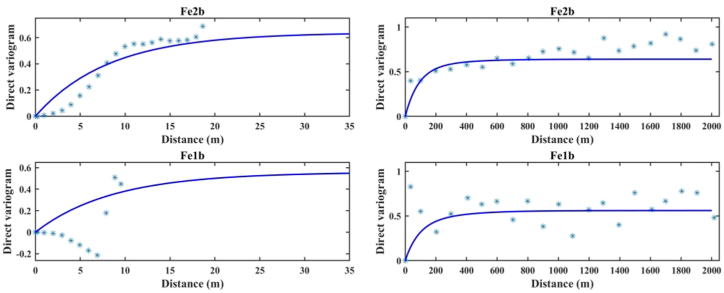


Fig. A3 Theoretical variograms of normal scores in (left) the vertical and (right) horizontal directions. Line: theoretical variograms; points: sample variograms

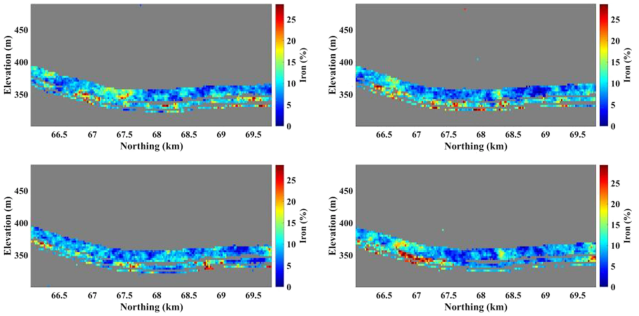


Fig. A4 Four random realizations of cascade modeling

Supplementary information. The training image used in this study is available for free at https://github.com/sultanabulkhair/Training-image.

Acknowledgments. Both authors are grateful to Nazarbayev University for funding this work via Faculty Development Competitive Research Grants for 2018–2020 under Contract No. 090118FD5336 and 2021–2023 under Contract No. 021220FD4951. The research of the first author was partially supported by the Australian Research Council Integrated Operations for Complex Resources Industrial Transformation Training Centre (project number IC190100017) and funded by universities, industry and the Australian Government. The authors also thank the ERG Company in Kazakhstan for providing the coal dataset and acknowledge the kind cooperation of Dr. Philippe Renard for permission to use the AR2GEMS software. The editorial team and two anonymous reviewers are gratefully acknowledged for helping to improve the paper.

References

[1] Finkelman RB, Dai S, French D (2019) The importance of minerals in coal as the hosts of chemical elements: A review. Int J Coal Geol 212:103251.

[2] Finkelman RB (1993) Trace and Minor Elements in Coal. In: Engel MH, Macko SA (eds) Organic Geochemistry. Topics in Geobiology, vol 11. Springer, Boston, pp 593-607.

[3] Dai B, Wu X, Zhang J, Ninomiya Y, Yu D, Zhang L (2020) Characteristics of iron and sulphur in high-ash lignite (Pakistani lignite) and their influence on long-term T23 tube corrosion under super-critical coal-fired boiler conditions. Fuel 264:116855.

[4] Bool LE, Peterson TW, Wendt JO (1995) The partitioning of iron during the combustion of pulverized coal. Combust Flame 100(1-2):262-270.

[5] Silva LF, Wollenschlager M, Oliveira ML (2011) A preliminary study of coal mining drainage and environmental health in the Santa Catarina region, Brazil. Environ Geochem Health 33:55–65.

[6] Huang X, Gordon T, Rom WN, Finkelman RB (2006) Interaction of Iron and Calcium Minerals in Coals and their Roles in Coal Dust-Induced Health and Environmental Problems. Rev Mineral Geochem 64(1):153-178.

[7] Liu J, Ward RC, Graham IT, French D, Dai S, Song X (2018) Modes of occurrence of non-mineral inorganic elements in lignites from the Mile Basin, Yunnan Province, China. Fuel 222:146-155.

[8] Hatherly P (2013) Overview on the application of geophysics in coal mining. Int J Coal Geol 114:74-84.

[9] Stolboushkin AY, Karpacheva AA, Ivanov AI (2011) Wall ceramic products based on waste coal and iron-containing additives. Inter-Kuzbass, Novokuznetsk.

[10] Namkane K, Naksata W, Thiansem S, Sooksamiti P, Arqueropanyo O (2016) Utilization of coal bottom ash as raw material for production of ceramic floor tiles. Environ Earth Sci 75:386.

[11] Goovaerts P (1997) Geostatistics for Natural Resource Evaluation. Oxford University Press, New York.

[12] Rossi ME, Deutsch CV (2014) Mineral resource estimation. Springer, Berlin.

[13] Boucher A, Dimitrakopoulos R (2012) Multivariate Block-Support Simulation of the Yandi Iron Ore Deposit, Western Australia. Math Geosci 44:449-468.

[14] Jones P, Douglas I, Jewbali A (2013) Modeling Combined Geological and Grade Uncertainty: Application of Multiple-Point Simulation at the Apensu Gold Deposit, Ghana. Math Geosci 45:949–965.

[15] Talebi H, Sabeti EH, Azadi M, Emery X (2016) Risk quantification with combined use of lithological and grade simulations: Application to a porphyry copper deposit. Ore Geol Rev 75:42-51.

[16] Mery N, Emery X, Cáceres A, Ribeiro D, Cunha E (2017) Geostatistical modeling of the geological uncertainty in an iron ore deposit. Ore Geol Rev 88:336-351.

[17] Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty. Wiley, New York.

[18] Isaaks EH (1990) The application of Monte Carlo methods to the analysis of spatially correlated data. PhD Thesis, Stanford University.

[19] Matheron G (1973) The Intrinsic Random Functions and Their Applications. Adv Appl Probab 5(3):439-468.

[20] Emery X, Lantuéjoul C (2006) TBSIM: A computer program for conditional simulation of three-dimensional Gaussian random fields via the turning bands method. Comput Geosci 32(10):1615-1628.

[21] Journel AG (1983) Nonparametric estimation of spatial distributions. J Int Assoc Math Geol 15:445-468.

[22] Deutsch CV (2006) A sequential indicator simulation program for categorical variables with point and block data: BlockSIS. Comput Geosci 32(10):1669-1681.

[23] Emery X (2007) Simulation of geological domains using the plurigaussian model: New developments and computer programs. Comput Geosci 33(9):1189-1201.

[24] Al-Mudhafar WJ (2017) Multiple-Point Geostatistical Lithofacies Simulation of Fluvial Sand-Rich Depositional Environment: A Case Study From Zubair Formation/South Rumaila Oil Field. SPE Reserv Eval Eng 21(1):39-53.

[25] Mariethoz G, Caers J (2014) Multiple-Point Geostatistics: Stochastic Modeling with Training Images. Wiley, New York.

[26] Boisvert JB, Pyrcz MJ, Deutsch CV (2007) Multiple-Point Statistics for Training Image Selection. Nat Resour Res 16:313-321.

[27] Mariethoz G (2018) When Should We Use Multiple-Point Geostatistics? In: Daya Sagar B, Cheng Q, Agterberg F (eds) Handbook of Mathematical Geosciences. Springer, Cham, pp 645-653.

[28] Boisvert JB, Pyrcz MJ, Deutsch CV (2010) Multiple Point Metrics to Assess Categorical Variable Models. Nat Resour Res 19:165-175.

[29] De Iaco S, Maggio S (2011) Validation Techniques for Geological Patterns Simulations Based on Variogram and Multiple-Point Statistics. Math Geosci 43:483-500.

[30] Tan X, Tahmasebi P, Caers J (2014) Comparing Training-Image Based Algorithms Using an Analysis of Distance. Math Geosci 46:149-169.

[31] Tahmasebi P (2018) Multiple Point Statistics: A Review. In: Daya Sagar B, Cheng Q, Agterberg F (eds) Handbook of Mathematical Geosciences. Springer, Cham, pp 613-643.

[32] Madani N, Maleki M, Emery X (2019) Nonparametric Geostatistical Simulation of Subsurface Facies: Tools for Validating the Reproduction of, and Uncertainty in, Facies Geometry. Nat Resour Res 28:1163-1182.

[33] Bastante FG, Ordóñez C, Taboada J, Matías JM (2008) Comparison of indicator kriging, conditional indicator simulation and multiple-point statistics used to model slate deposits. Eng Geol 98(1-2):50-59.

[34] Rezaee H, Asghari O, Koneshloo M, Ortiz JM (2014) Multiple-point geostatistical simulation of dykes: application at Sungun porphyry copper system, Iran. Stoch Environ Res Risk Assess 28:1913-1927.

[35] Guardiano FB, Srivastava RM (1993) Multivariate geostatistics: beyond bivariate moments. In: Soares A (ed) Geostatistics Tróia '92. Quantitative Geology and Geostatistics, vol 5. Springer, Dordrecht, pp 133-144.

[36] Strebelle S (2002) Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics. Math Geol 34:1-21.

[37] Zhang T, Switzer P, Journel A (2006) Filter-Based Classification of Training Image Patterns for Spatial Simulation. Math Geol 38:63-80.

[38] Arpat GB, Caers J (2007) Conditional Simulation with Patterns. Math Geol 39:177-203.

[39] Avalos S, Ortiz JM (2020) Recursive convolutional neural networks in a multiple-point statistics framework. Comput Geosci 141:104552.

[40] Bai T, Tahmasebi P (2020) Hybrid geological modeling: Combining machine learning and multiple-point statistics. Comput Geosci 142:104519.

[41] Mariethoz G, Renard P, Straubhaar J (2010) The Direct Sampling method to perform multiple-point geostatistical simulations. Water Resour Res 46(11).

[42] Huang T, Li X, Zhang T, Lu DT (2013) GPU-accelerated Direct Sampling method for multiple-point statistical simulation. Comput Geosci 57:13-23.

[43] Rezaee H, Mariethoz G, Koneshloo M, Asghari O (2013) Multiple-point geostatistical simulation using the bunch-pasting direct sampling method. Comput Geosci 54:293-308.

[44] Straubhaar J, Renard P, Mariethoz G (2016) Conditioning multiple-point statistics simulations to block data. Spat Stat 16:53-71.

[45] Emery X, Lantuéjoul C (2014) Can a Training Image Be a Substitute for a Random Field Model? Math Geosci 46:133-147.

[46] Anderson KS, Hickson TA, Crider JG, Graham SA (1999) Integrating Teaching with Field Research in The Wagon Rock Project. J Geosci Educ 47:227-235.

[47] Bayer P, Huggenberger P, Renard P, Comunian A (2011) Three-dimensional high resolution fluvio-glacial aquifer analog: Part 1: Field study. J Hydrol 405(1-2):1-9.

[48] Pyrcz MJ, Boisvert JB, Deutsch CV (2008) A library of training images for fluvial and deepwater reservoirs and associated code. Comput Geosci 34(5):542-560.

[49] Deutsch CV, Wang L (1996) Hierarchical object-based stochastic modeling of fluvial reservoirs. Math Geol 28:857-880.

[50] Pyrcz MJ, Boisvert JB, Deutsch CV (2009) ALLUVSIM: A program for event-based stochastic modeling of fluvial depositional systems. Comput Geosci 35(8):1671-1685.

[51] Goodfellow R, Consuegra FA, Dimitrakopoulos R, Lloyd T (2012). Quantifying multi-element and volumetric uncertainty, Coleman McCreedy deposit, Ontario, Canada. Comput Geosci 42:71-78.

[52] Boucher A, Costa JF, Rasera LG, Motta E (2014) Simulation of Geological Contacts from Interpreted Geological Model Using Multiple-Point Statistics. Math Geosci 46:561-572.

[53] Paithankar A, Chatterjee S (2018) Grade and Tonnage Uncertainty Analysis of an African Copper Deposit Using Multiple-Point Geostatistics and Sequential Gaussian Simulation. Nat Resour Res 27:419-436.

[54] Vistelius AB (1989) Principles of Mathematical Geology. Springer, Dordrecht.

[55] Houlding S (1994) 3D Geoscience Modeling, Computer Techniques for Geological Characterization. Springer, Berlin.

[56] Mallet JL (1992) Discrete smooth interpolation in geometric modelling. Comput Aided Des 24(4):178-191.

[57] Mallet JL (2002) Geomodeling. Oxford University Press, New York.

[58] Deutsch CV, Journel AG (1992) Geostatistical software library and users guide. Oxford University Press, New York.

[59] Hardy RL (1990) Theory and applications of the multiquadric-biharmonic method 20 years of discovery 1968–1988. Comput Math Appl 19(8-9):163-208.

[60] SRK Consulting (Kazakhstan) (2018) Competent persons report on the coal assets of Shubarkol Komir JSC, Republic of Kazakhstan.

[61] David M (1977) Geostatistical ore reserve estimation. Elsevier, New York.

[62] Deutsch CV (1989) DECLUS: a fortran 77 program for determining optimum spatial declustering weights. Comput Geosci 15(3):325-332.

[63] Leuangthong O, McLennan JA, Deutsch CV (2004) Minimum Acceptance Criteria for Geostatistical Realizations. Nat Resour Res 13:131-141.

[64] Meerschman E, Pirot G, Mariethoz G, Straubhaar J, Van Meirvenne M, Renard P (2013) A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. Comput Geosci 52:307-324.

[65] Journal AG, Alabert A (1989) Non-Gaussian data expansion in the earth sciences. Terra Nova 1:123-134.

[66] Pardo-Igúzquiza E, Dowd PA (2003) CONNEC3D: a computer program for connectivity analysis of 3D random set models. Comput Geosci 29(6):775-785.

[67] Goovaerts P (2001) Geostatistical modelling of uncertainty in soil science. Geoderma 103(1-2):3-26.

[68] Emery X (2004) Testing the correctness of the sequential algorithm for simulating Gaussian random fields. Stoch Environ Res Risk Assess 18:401-413.