DOI 10.4430/bgo00371

Presenting the attribute kriging algorithm for automatic domaining and simultaneous estimation

S. Mahmoudi and O. Asghari

Simulation and Data Processing Laboratory, School of Mining Engineering, College of Engineering, University of Tehran, Iran

(Received: 5 December 2020; accepted: 25 July 2021; published online: 27 January 2022)

ABSTRACT Due to the development of sampling facilities in the mining industry, the spatial reporting of data sets is increasing, more variables are being noted, and broader areas are being covered. Therefore, it is crucial to subdivide the study areas into smaller domains to clarify the computation of the different behaviour of natural phenomena. In the estimation of mineral resources, the process consists of partitioning the mineralised area into several domains defined by the ore grade differences. Defined domains can be considered as a clustering problem. Even so, non-spatial techniques of clustering do not guarantee the spatial continuity of geostatistical data sets. Multivariate spatial clustering methods must therefore be applied, which often indicate the specifics of spatial continuity and heterogeneity. As a result, the Geostatistical Hierarchical Clustering algorithm is proposed. The validation of the non-spatial and spatial clustering techniques is evaluated by a synthetic data set in which the acquired results highlight the necessity of applying the algorithm. The mentioned algorithm is used as a proper tool for automatic domaining in geostatistical data set estimation. Its effect on improving the results derived from the kriging estimator is analysed on the synthetic data set. Consequently, the Attribute Kriging algorithm is introduced for estimating mineral resources.

Key words: clustering, multivariate domaining, certainty, kriging estimation, geostatistics.

1. Introduction

Classifying data sets in various scientific fields to simplify descriptions, design a regular route for indicating structural models of data sets and related interpretations, facilitate the sampling process, and provide an appropriate context for accessing more information, is imperative (Oliver and Webster, 1989). According to the increasing development and evolution of sampling facilities in the mining industry, the compilation of spatial data sets is increasing, taking into account more variables and covering broader areas. Therefore, it is often necessary to divide the study area into domains to evaluate various behaviours of natural phenomena, identify hidden structures, examine the underlying information, and simplify mineral resource modelling (Romary *et al.*, 2015). In the estimation of mineral resources, the general process means dividing the mineralised area into domains defined by the ore grade differences in the univariate domaining. While it is advantageous to provide an appropriate mineralisation model in the region, corresponding to geological facies, each of the considered variables indicates different spatial behaviour in terms of mean, variability, and spatial structure (Emery and Ortiz, 2004). The domaining often regards the mineralogical classification. For example, high, medium, and low-grade domains are individually defined and analysed for each mineralisation unit, which is a time-consuming step in estimating mineral resources (Stegman, 2001). Besides, the domaining of mineralised regions solely using the ore grade does not consider the spatial continuity among adjacent domains, and the presence of no uncertainties on the boundaries by this univariate domaining implies decreasing estimation performance. A significant number of parameters should be combined to identify the domains in multivariate-domaining. Nevertheless, due to the number of qualitative and quantitative information, it quickly leads to a computationally complex problem. Fortunately, the number of domains is limited; however, the dependency among them and the reliability of each domain's variables must be guaranteed (Emery and Ortiz, 2004; Romary *et al.*, 2012).

Consequently, defining domains can be considered as a clustering problem. Clustering automatically classifies the mineralised region into homogeneous and independent clusters according to multivariate data sets. Nevertheless, traditional clustering approaches designed for independent observations do not guarantee the spatial continuity of the resulting clusters. Accordingly, to cluster geostatistical data sets that often show spatial continuity and heterogeneity behaviour in the study area, traditional clustering algorithms adapted to multivariate geostatistical data sets should be applied. Particular emphasis is given to procedures used to optimise traditional algorithms to apply the resulting clusters' spatial contiguity effect (Romary *et al.*, 2012, 2015). Clustering can be provided in a variety of ways, mainly dependent on similarity measures among the observations. The primary point is that the similarity of the attribute space does not guarantee a similar geographical space process. Therefore, in addition to determining these parameters in the attribute space, the geographical space's similarity should also be considered. The location of observations belonging to a cluster should therefore be related to each other in the geographical space (Fouedjio, 2016a, 2016b).

Oliver and Webster (1989) were the first to propose a method for clustering geostatistical data sets: a method based on a traditional hierarchical clustering algorithm with an adapted dissimilarity measurement among observations to determine the spatial continuity of resulting clusters. They calculated the dissimilarities among observations by multiplying the stationary variogram and the stationary variogram of the first principal component of observations to the dissimilarity matrix in the univariate and multivariate cases. Hence, it leads to smoothing the dissimilarity of adjacent pairs of observations, obscuring boundaries between different clusters. Ambroise *et al.* (1995) proposed a method upon the Markov random field based on the Expectation-Maximisation (EM) algorithm. It can be applied to irregularly spaced data sets while locations of observations in the geographical space are structured by the Delaunay triangulation graph of the data sets, and edge weights of the graph are assigned uniformly. Allard and Guillot (2000) investigated a mixture of Gaussian random field-based clustering algorithms in which an approximation of the EM algorithm is used. This approach relies on the following assumptions:

1. the data set is assumed to be Gaussian, and observations belonging to each cluster are considered independent;
2. the algorithm requires calculating the maximum likelihood estimator (MLE) in each realisation of the EM algorithm, including the covariance matrix's inversion.

Pawitan and Huang (2003) proposed two hierarchical and non-hierarchical clustering algorithms, which are spatially constrained. The constraint is structured by the Delaunay triangulation graph of the sample points, which lacks consideration of the lengths of the structure's edges. Romary *et al.* (2015) proposed two spatially constrained clustering algorithms. The first is an agglomerative hierarchical clustering algorithm with complete linkage. The spatial structure is based on the Delaunay triangulation, and the lengths of the edges are taken into account. The second algorithm is an alteration of the spectral clustering algorithm (von Luxburg, 2007) to partition the same graph. Fouedjio (2016a, 2016b) proposed a geostatistically consistent agglomerative hierarchical method, while the similarity among observations is a function of spatial correlation. A Gaussian mixture model (GMM) approach is investigated by Madenova and Madani (2021) in order to obtain clusters in the practice of geometallurgical modelling, which provides the uncertainty of each observation to the clusters.

In this paper, the proposed clustering algorithm is performed by a ward-like hierarchical clustering method with spatial constraint, relying on a spatial structure of data sets and overall similarity among variables defined in the attribute space. Section 2 outlines the basics of clustering and kriging algorithms. Then, in Section 3, the efficiency of the algorithms is analysed and evaluated on a synthetic data set. Finally, using the clustering algorithm as a proper tool for automatic domaining, the proposed kriging algorithm is applied to the real data set of the Mehdi Abad lead and zinc deposit. According to the algorithm's performance and results, it significantly improves the quality of the estimation in both synthetic and real data sets.

2. Methodology

The following section introduces K-Means, Hierarchical, Spectral, and the Geostatistical Hierarchical Clustering (GHC) methods. Further, determining the appropriate number of clusters based on the cluster's intrinsic validation index, calculating the resulting clusters' certainty, and the validation of clustering algorithms are also considered in this section. Finally, the GHC and the Attribute Kriging (AK) estimator's association and the estimation validation procedure are demonstrated.

2.1. K-Means

Generally, the steps of the K-Means clustering method are as follows:

1. determining the centres of cluster number (*k*);
2. random selection of primary cluster centres (μ);
3. calculating the Euclidean distance between the random centres and the observations in pairs, designating the observations to the nearest selected centre, and finally assigning the designated observations to the clusters associated with the same centre;
4. calculating the clusters' average values based on the observations within them in accordance with Eq. 1, and reproducing new clusters:

$$\bar{\mu}_r = \frac{1}{size\ (C_r)} \sum_{i=1, \bar{x}_i \in\ C_r}^{n} \bar{x}_i \tag{1}$$

5. repeating the last two steps and calculating the error function value according to Eq. 2 after step 4 for the reproduced clusters; until the value of this function is fixed and reaches the minimum:

$$E = \sum_{r=1}^{k} \sum_{i=1, \bar{x}_i \in C_r}^{n} (\bar{x}_i - \bar{\mu}_r)^2 \tag{2}$$

where C_r corresponds to cluster r in which $1 \leq r \leq k$ and the size of C_r is equal to the number of objects within it while n represents the total number of observations and $1 \leq i \leq n$.

2.2. Hierarchical

This method measures the dissimilarity existing among the clusters based on the dissimilarity between pairs of observations. The most popular variants are:

- single linkage: in this case, two separate C_i and C_j, clusters are merged based on the minimum distance between the two elements *P* and *P'* within them:

$$d_{min}(C_i,C_j) = \min_{P \in C_i, P' \in C_j} |P - P'| \tag{3}$$

- complete linkage: in this case, two separate C_i and C_j, clusters are merged based on the maximum distance between the two elements *P* and *P'* within them:

$$d_{max}(C_i,C_j) = \max_{P \in C_i, P' \in C_j} |P - P'| \tag{4}$$

- average linkage: in this case, two separate C_i and C_j, clusters are merged based on the average distance between the two elements *P* and *P'* within them:

$$d_{avg}(C_i,C_j) = \frac{1}{n_i n_j} \sum_{P \in C_i} \sum_{P' \in C_j} |P - P'|. \tag{5}$$

Then, due to the results, a new cluster is built. This process goes so far when all observations are placed in one cluster. It should be noted that hierarchical clustering does not require determining the number of clusters in the early stages.

2.3. Spectral

The spectral clustering method classifies the data sets by modifying their primary structure and using other clustering methods such as K-Means to determine clusters. The basis of the mentioned method is the proximity of observations to each other, resulting in identifying complex and intertwined structures with a high separability due to modifying the structure. The steps of the spectral clustering method are as follows:

- calculation of the similarity matrix between observations (*S*) based on the Euclidean distance and, consequently, the proximity matrix estimation based on Delaunay triangulation ($W \sim S$);
- calculating the degree matrix (*D*) according:

$$D_{ii} = \sum_{j=1}^{n} W_{ij} \tag{6}$$

- estimating the Laplacian matrix:

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{7}$$

- calculating *k* largest eigenvalues of the matrix *L* and determining the matrix $F \in R^{nxk}$, so that the columns of the matrix are related to *k* number of the matrix *L*'s initial eigenvectors;
- normalising rows of matrix *F*;
- performing the K-Means clustering method on the matrix *F* rows and forming clusters $C_1, ..., C_k$.
- assigning observations to the cluster in which row *I* of matrix *F* is located.

This method requires determining the number of expected clusters in the early stages. The stated method's convergence becomes an optimisation problem based on selecting appropriate eigenvalues and the eigenvectors associated with them:

$$\min_{F \in R^{n \times k}} Tr\ (F^T L F)\ subject\ to\ F^T\ F\ = I \tag{8}$$

where *I* is the identity matrix, and *Tr* is the effect of the matrix.

2.4. Geostatistical hierarchical clustering

The primary purpose of the GHC algorithm is to provide concepts related to spatial continuity among observations. These concepts should be useful in adapting traditional clustering algorithms in order to provide geostatistical data-based approaches. In this case, a clustering algorithm considering the basic needs of mineral resource modelling projects can follow a smooth path to unsupervised classification and improve mineral resource estimation quality. The basis of the proposed algorithm generally consists of two steps:

1. spatial structuring is created by a graph function according to the observations' spatial location set on various criteria and desirability of measuring the spatial similarity;
2. the use of adapted traditional clustering methods, while the mentioned structure controls the observations' spatial continuity.

Consequently, the spatial similarity among observations leads to understanding the spatial interactions among the resulting clusters.

2.4.1. Spatial structure

The spatial structure of the data set is based on the observations' adjacency in terms of geographical space represented in Fig. 1 by the graph function *G*, according to Eq. 9. If observations x_i and $x_{i'}$ are adjacent to each other, the corresponding value in the matrix $[G_{ii'}]$ is equal to 1, otherwise 0. In this regard, 1 ≤ *i*, *i* ≤ *N*, *i* ≠ *i'*, and *N* indicates the total number of observations.

While the observations' adjacency has already been controlled by the dissimilarity function (*S*) in Eq. 10 according to a normalised kernel function (α = 1), the higher the spatial dissimilarity of observations, the closer the corresponding value in the matrix [$S_{ii},$] to 0; otherwise, it is closer to 1 ($0 \leq S_{ii}, \leq 1$):

$$G_{ii'} = \begin{cases} 1, & if \;\; x_i \longleftrightarrow x_{i'} \\ 0, & otherwise \end{cases} \tag{9}$$

$$S_{ii'} = exp(-\alpha \|x_i - x_{i'}\|^2). \tag{10}$$

So each observation connects to the optimal number of the most adjacent neighbours of its own in the geographical space. The optimal number is set according to the Rand index (*RI*) over neighbour number 1 to 100 for 1000 samplings.

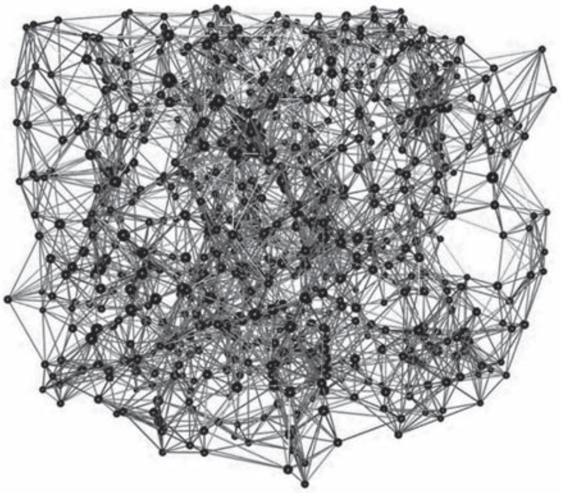


Fig. 1 - Spatial structure of 800 observations based on 8 adjacent neighbours.

2.4.2. Overall dissimilarity

Given the spatial data $(x_1, \ldots, x_n) \in R^{nxp}$, *n* represents the number of observations, and *p* represents the number of variables in the attribute space (*j*). As illustrated in Eq. 11, while *j* = 1, 2, ..., *p*, the dissimilarity (*D*) is calculated by the usage of the Euclidean distance (*d*) and the mean weights (*w*) of every attribute for each pair of the standardised variable in the mentioned space:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} w_j . d_j(\, x_{ij}, x_{i'j}) \,; \sum_{j=1}^{p} W_j = 1 \tag{11}$$

where W_j corresponds to the weights of each attribute. Thus, the summed weights used for variables must be equal to 1. Note that despite considering the dissimilarity of variables in attribute space, dissimilarity in geographical space is also calculated by Eq. 11 as an independent parameter with proportional weight. The latter is added to the attribute dissimilarities, which results in an overall dissimilarity (*D**). Due to the RI analysis, the mentioned proportional weight for K-Means and Hierarchical is 5 and 30%, respectively, which leads to improving the performance of these methods according to the synthetic data set. Moreover, it is 100% for Spectral and 0% for the GHC algorithm.

2.4.3. The adapted clustering algorithm

Non-spatial clustering is the core of the proposed spatial clustering algorithm based on hierarchical clustering using the Ward method as a distance measurement. This method is cumulative and uses two dissimilarity matrices to compute the dissimilarity of observations. The first one is obtained from the spatial structure, which affects the algorithm by the spatial continuity of observations. The overall dissimilarity matrix (*D**), which is determined in the attribute space, is selected as the second dissimilarity matrix, affecting the algorithm by observation relationships in the attribute space.

The algorithm calculates the Ward distance (W_{ij}) between the observations *i* and *j*:

$$w_{ij} = \frac{{D_{ij}}^2}{2n} \tag{12}$$

where D_{ij} is the value corresponding to the degree of dissimilarity between the two observations *i* and *j*, and *n* is equal to the number of observations.

Finally, Ward's ultimate value (W_{ij}) is reached:

$$W_{ij} = \alpha \times w_G + (1 - \alpha) \times w_{D*} \tag{13}$$

where W_G represents the Ward value calculated for the dissimilarity matrix obtained from the dissimilarity function *S*, and W_{D*} refers to the Ward value calculated by the overall dissimilarity matrix. The value of *α* determines the importance of each input matrix, and its best value is a compromise between the loss of W_{D*} and the gain of W_G set up on the quality criterion of the clusters obtained with different values of *α* ($0 \leq \alpha \leq 1$) performed by the function "choicealpha" from the "ClustGeo" package in R (Chavent *et al.*, 2018).

2.5. Number of clusters

The Caliński-Harabasz (*CH*) index is selected to determine the appropriate number of clusters based on the cluster's internal-external instability and variability (Caliński and Harabasz, 1974). Given different clusters *k* = 1, 2, 3,.., *n* – 1, the appropriate number of them is the one that maximises the *CH* index:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{14}$$

The total variance among clusters *B(k)*, and the total intra-cluster variance *W(k)* are calculated according to Eqs. 15 and 16:

$$B(k) = \sum_{m=1}^{k} n_m \|\bar{y}_m - \bar{y}\|^2 \tag{15}$$

$$W(k) = \sum_{m=1}^{k} \sum_{t \in C_m} n_m \|\bar{y}_m - \bar{y}\|^2 . \tag{16}$$

While $y_t \in R^k$ is the vector corresponding to the *t*-th row of the matrix F, n_m is the number of points in the cluster C_m and $\bar{y}_m$ is the average of points in the cluster C_m, which is determined by:

$$\bar{y}_m = \frac{1}{n_m} \sum_{t \in C_m} y_t \tag{17}$$

and $\bar{y}$ is the overall average:

$$\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t . \tag{18}$$

2.6. Clustering certainty

The certainty of observations is calculated based upon the matrix [$G_{ii'}$] and [$S_{ii'}$]. The procedure adopts the connection quality between observations, governed by the eligible number of neighbours estimated by the *RI*. The results of certainty for each observation are expressed as a degree of belonging to each cluster. For example, given two different clusters $k = C', C''$ observations x_i, $x_{i'}$ and $x_{i''}$ are connected according to the matrix [$G_{ii'}$] and $x_{i'} \in C'$, $x_{i''} \in C''$. Therefore, this connection indicates that the observation *i* belongs to the cluster *C′* and *C″* by an equal probability of 50%. Nonetheless, the mentioned probability should also be controlled by the spatial similarity of observations, hence, the matrix [$S_{ii'}$] governs the last step. Therefore, the higher the spatial similarity among observations x_i and $x_{i'}$, the greater the probability of belonging the *i* to the cluster *C′*. Although the certainty calculates the probability of each observation belonging to the resulting clusters, it can be used as a proper

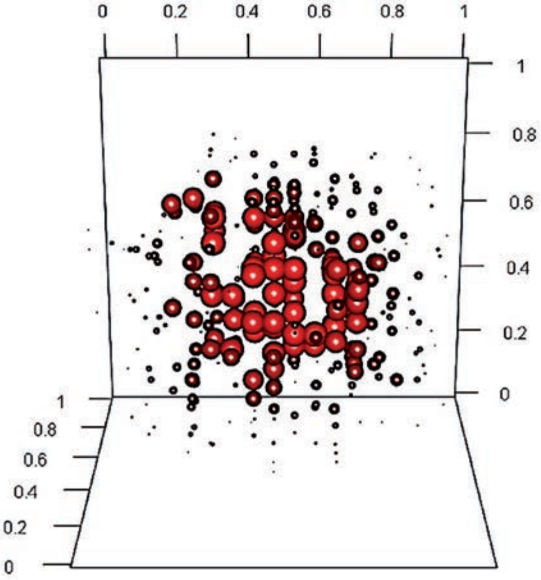


Fig. 2 - A schematic of the red cluster certainty (the smaller and darker the red observations, the lower the certainty of the red cluster).

tool to enhance clustering results at the boundaries. The low certainty observations relative to a cluster were removed from that cluster and assigned to another cluster with maximum certainty, which leads to correct misclassified observations.

2.7. Clustering validation

The validation of clustering methods is based on a synthetic data set. Thus, the variogram and *RI* validate the accuracy of the clustering methods.

2.7.1. Spatial validation

Using the variogram for each variable in predetermined domains in a synthetic data set, and comparing its parameters with the conclusions and interpretations of the variogram results in obtained clusters, leads to spatially validating clustering methods.

2.7.2. Rand index

RI is applied to estimate the clustering accuracy. The following parameters are calculated from clustering methods and predetermined domains in a synthetic data set:

- A: pairs of number of observations, labelled similarly in both the resulting clusters and predefined domains;
- B: pairs of number of observations, labelled differently in both the resulting clusters and predefined domains.

Therefore, according to Eq. 19, the *RI* is calculated so that the *n* is the number of observations:

$$RI = \frac{A+B}{\binom{n}{2}} . \tag{19}$$

The closer the *RI*'s value to 1, the higher the accuracy of the clustering method (0 ≤ *RI* ≤ 1).

2.8. Kriging estimator

According to Eq. 20, the ordinary Kriging estimator is applied as a suitable tool to estimate the study area:

$$Z(x_0) = \sum_{i=1}^{N} \lambda_i Z(x_i) . \tag{20}$$

Thus, Λ_i represents the Kriging weights, $Z(x_i)$ is the variable's value at the premeasured points, and $Z(x_0)$ is the variable's estimated value. According to the spatial correlation between the observations defined as a variogram, kriging estimates the variables' values at known coordinates using the same values in other points with known coordinates:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_{i+h})]^2 \tag{21}$$

where $N(h)$ represents the number of pairs of points whose distance from each other is equal to h. $Z(x_i)$ and $Z(x_{i+h})$ are regional variables with distance h from each other. The spherical model is used as an appropriate fitted model for variograms:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C\left(\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\right) & 0 < h \leq a \\ C_0 + C, & h > a \end{cases} \tag{22}$$

where C_0 corresponds to the nugget effect value. The parameter C is the structure's variance so that $C_0 + C$ represents the sill of the variogram and its range shown by a.

2.8.1. Estimation validation

Estimation Validation is derived from various methods, such as analysing primary and estimated statistics of the data sets and estimation variance. Also, cross-validation is used in each domain; five regions are randomly removed from the primary data set after estimating variables and re-estimated according to the estimation parameters. Finally, the correlation of the re-estimated and the primary data sets in the same regions are applied to determine the estimation's validity.

3. Performance and results

3.1. Synthetic data set

The synthetic data set environment is a cube with 1-m sides in 3D space with a grid size of 20×20×20 simulated in 3 variables. The smaller cube as the internal domain with sides of 0.5 m is located in the centre of this environment and covers the range of [0.25, 0.75] in directions X, Y, Z.

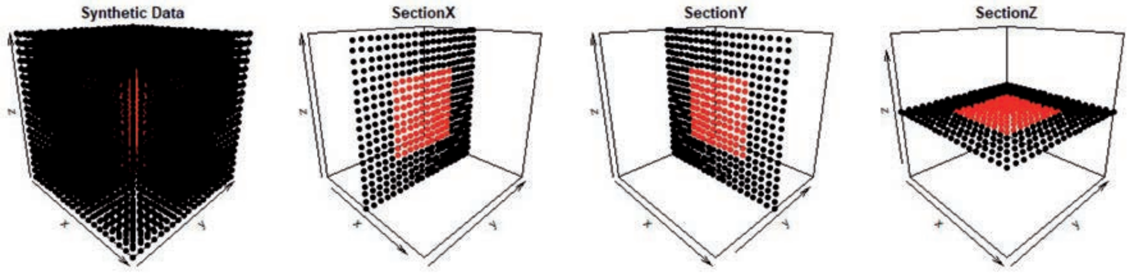


Fig. 3 - Predefined domains and the grid of synthetic data set.

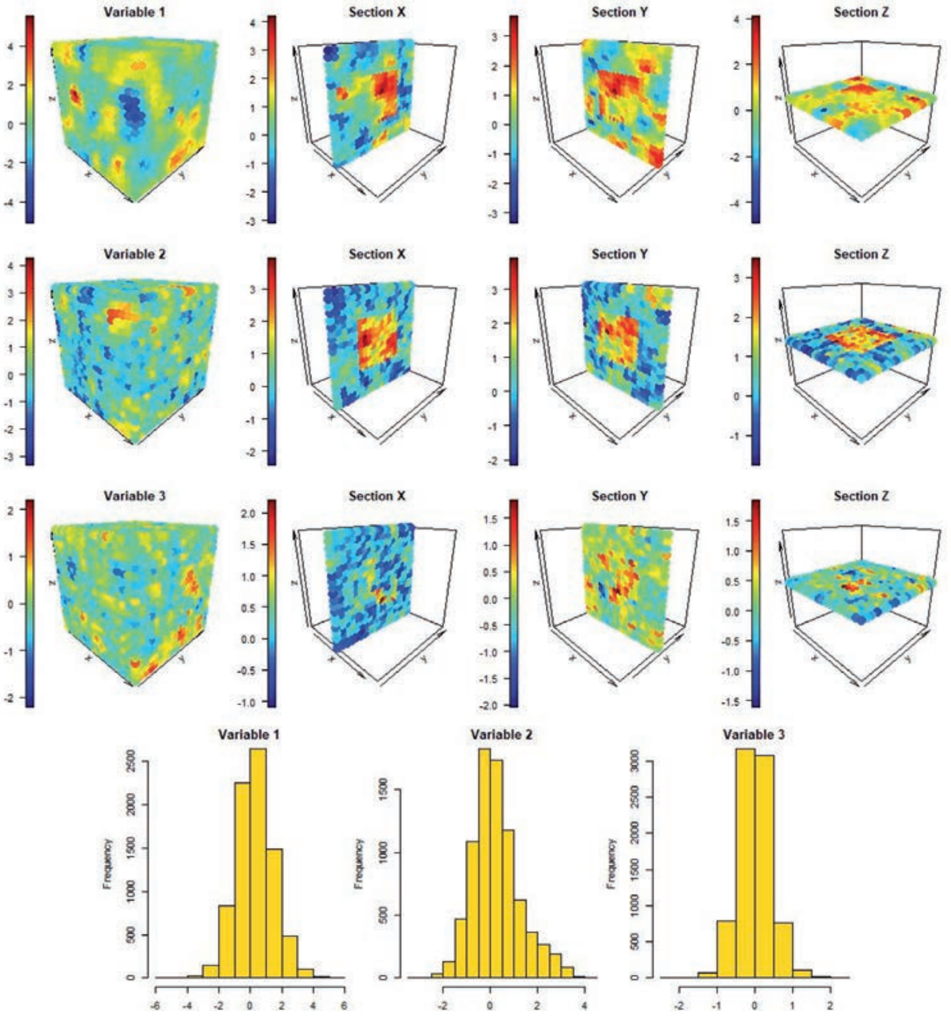
The internal domain's surrounding area is considered to be the external domain and covers the two intervals [0, 0.25) and (0.75, 1] in each direction. According to the grid, the internal domain includes 1000 points, and the external domain has 7000 points. Each domain's variables are simulated separately using the Sequential Gaussian Simulation (SGS) method. Fig. 3 displays the environment.

3.1.2. Variables

To generate the first variable in the internal domain, 100 random points have been used. The mean and standard deviation are both equal to 2. The spherical variogram model with variance 1 and range 1 is applied. The same method with different parameters has been implemented on 700 points for the external domain. Table 1 represents the parameters of each variable.

Table 1 - Simulation parameters of synthetic data set.

Variables	Domain	Mean	Standard deviation	Variogram model	Variogram sill	Variogram range	Number of points
1	Internal	2	2	Spherical	1	1	100
	External	0	2	Exponential	1	1	700
2	Internal	2	0.5	Spherical	1	0.25	100
	External	0	0.5	Spherical	1	0.25	700
3	Internal	0	0	Spherical	1	0.5	100
	External	0	0	Exponential	1	0.5	700


Variable 1

Section X

Section Y

Section Z

Variable 2

Section X

Section Y

Section Z

Variable 3

Section X

Section Y

Section Z

Variable 1

Variable 2

Variable 3

Fig. 4 - Generated variables in the synthetic data set and the histogram of each variable.

According to Fig. 4, the first variable's predetermined domains are not separated. So, there are overlaps in parts of the study area. Besides, the distribution of this variable does not indicate distinguished populations. The mentioned procedure challenges the clustering process. The second variable is considered the target variable. The histogram of this variable also does not represent the difference in populations entirely. Even so, the variable has not been generated in an intertwined way. The scattered and relative values of this variable in both domains can reduce the accuracy of the clustering algorithms. The third variable is simulated in both domains similarly and plays the role of noise in the clustering process, setting a significant challenge to the clustering procedure.

3.1.3. Number of clusters

The *CH* index is calculated according to 3 variables of the synthetic data set.

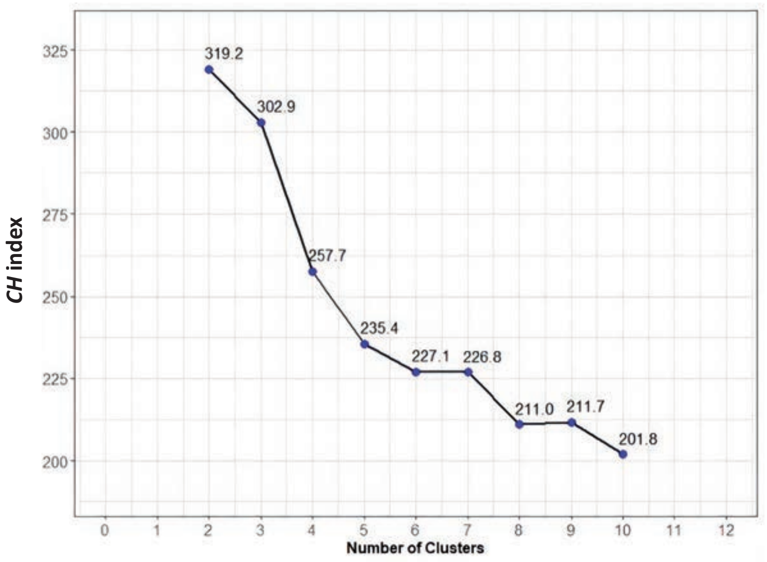


Fig. 5 - *CH* index of cluster numbers.

Fig. 5 presents the *CH* index based on the number of different clusters. Regarding the obtained results, 2 clusters with the *CH* value of 319.2 are specified as a suitable number of clusters.

3.1.4. Clustering procedure

In this section, the results achieved from the clustering of the synthetic data set are discussed. Firstly, the performance of the K-Means, Hierarchical and Spectral clustering methods are illustrated. The clustering process is applied to 10% of the standardised synthetic data set. Before analysing the results, the target variable's variograms in predetermined domains need to be estimated. The mentioned variograms are implemented as a criterion for spatially validating the clustering algorithms in the next steps. Fig. 6 shows the results of the clustering methods, and Fig. 7 displays the average *RI* in the implementation of each clustering algorithm per 100 random samplings. The K-Means clustering method has not segregated the clusters practically due to the lack of spatial continuity, and it has significant sensitivity to the values associated with the attribute space. Increasing the spatial dissimilarity proportional weight (>5%) in the overall dissimilarity measurements reduces the method's accuracy, so the best response is obtained considering a weight of 5%.

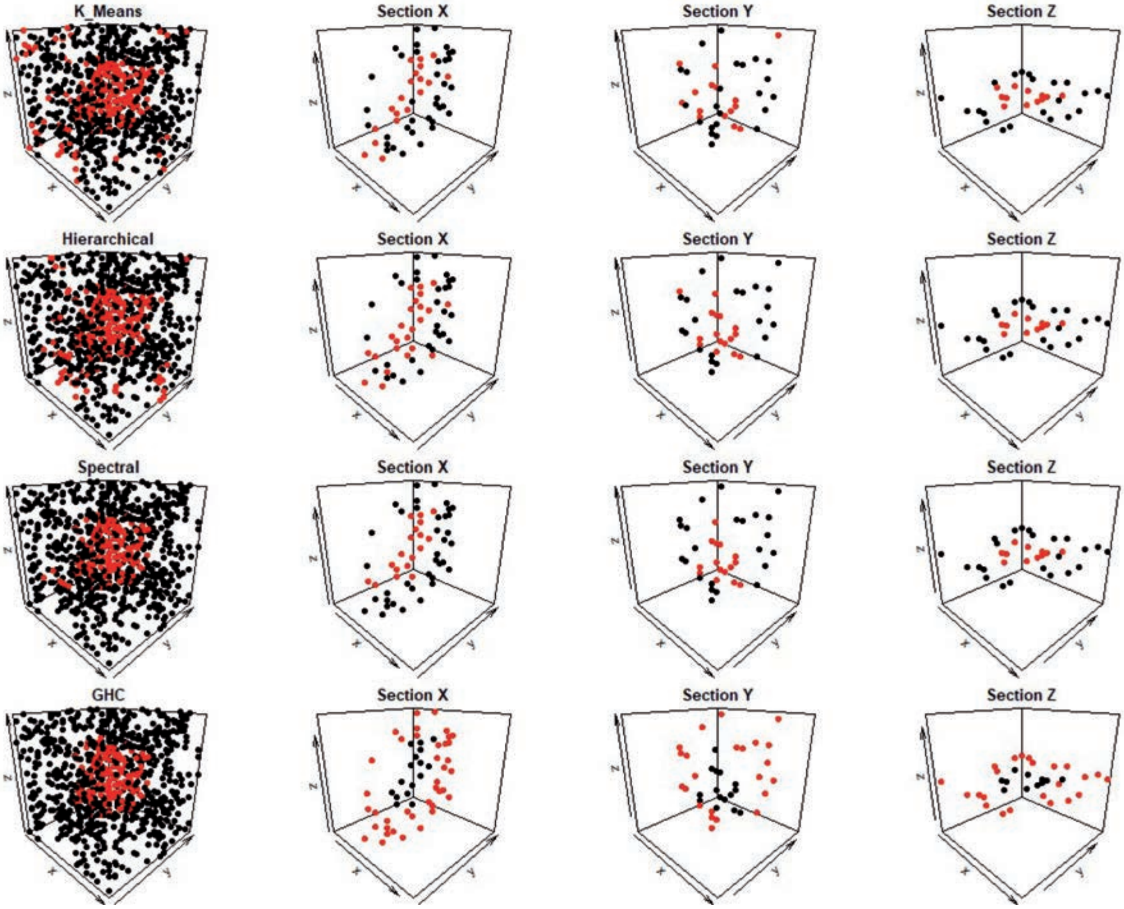


Fig. 6 - Clustering results of the synthetic data set performed by K-Means, Hierarchical, Spectral, and GHC algorithm.

On the other hand, the Hierarchical method has not effectively identified the internal and external clusters, and the lack of spatial continuity in the resulting clusters is apparent. This method is significantly sensitive to samples, so, in some sampling, it correctly classifies the observations and distinguishes the variables' high and low values into meaningful clusters. Considering the appropriate spatial dissimilarity proportional weight (30%) improves the Hierarchical clustering quality. Next, clusters extracted from the Spectral method are more meaningful and efficient than the latter ones and almost reveal the hidden structure of the synthetic data set. However, even in this case, the lack of spatial continuity is observed in the study area. It is worth mentioning that the spatial dissimilarity proportional weight is 100%, which is a complementary parameter in this method's overall dissimilarity measurement.

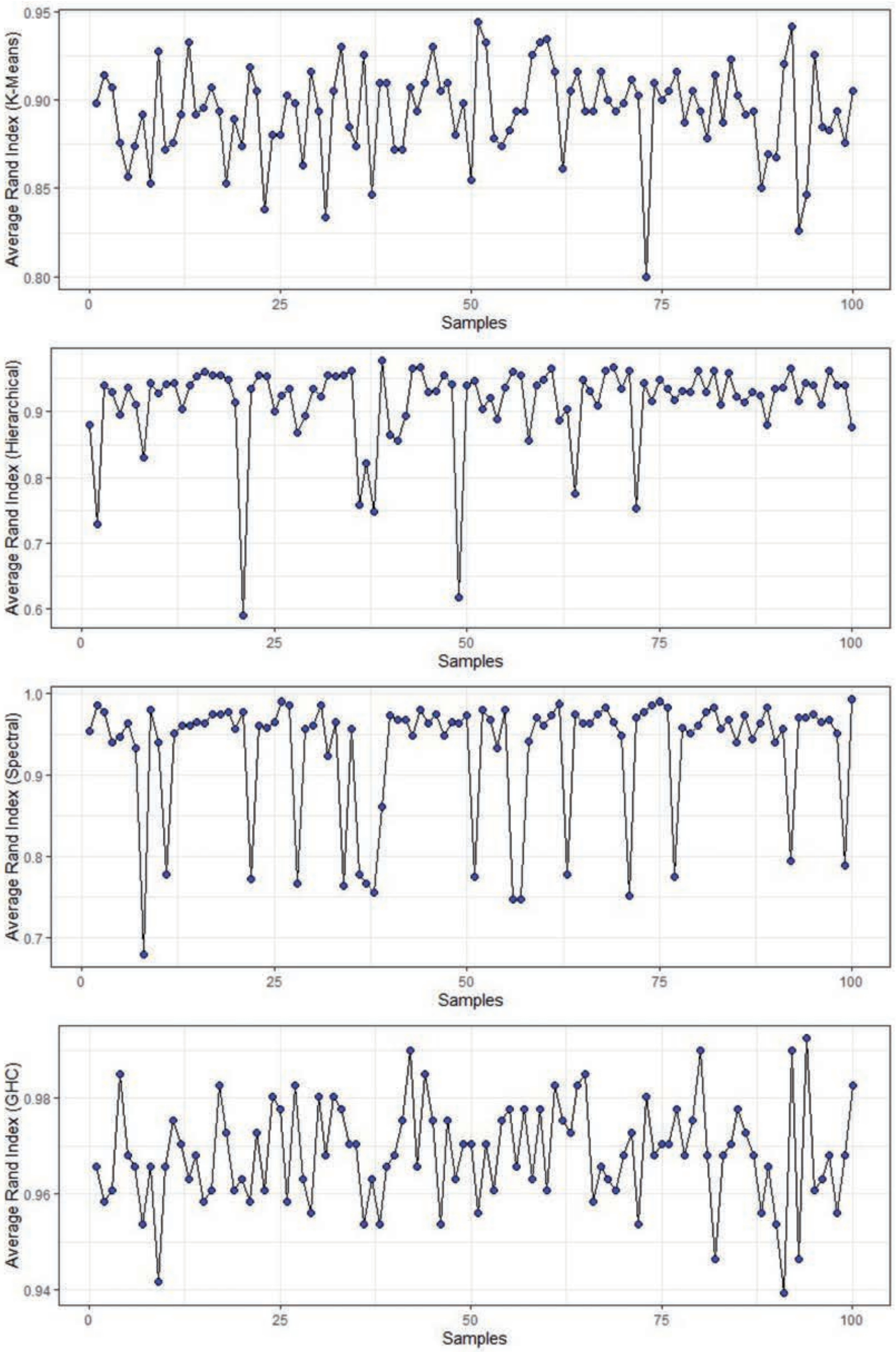


Fig. 7 - The average *RI* of the clustering methods with the optimal dynamic neighbours per sampling for the GHC method.

Finally, the clusters extracted from the GHC are significant, which reveal the hidden structure of the synthetic data set perfectly. According to Fig. 6, the resulting clusters' spatial continuity in the study area is entirely apparent. The maximum value of the *RI* is 0.99, and the minimum value is 0.94. The index variance is 0.0001, and the standard deviation is equal to 0.01 (Table 2), which indicates the algorithm's excellent performance. In this case, spatial structuring is calculated separately per sampling while the *RI* governs the optimal number of neighbours to estimate structures. These numbers are not necessarily the same for each sampling.

Table 2 - *RI* parameters of the clustering methods with the optimal dynamic neighbours per sampling for the GHC method.

RI parameters	K-Means	Hierarchical	Spectral	GHC
Mean	0.89	0.91	0.93	0.97
Variance	0.0004	0.0049	0.0049	0.0001
Standard deviation	0.02	0.07	0.07	0.01
Maximum	0.94	0.97	0.99	0.99
Minimum	0.80	0.59	0.67	0.94

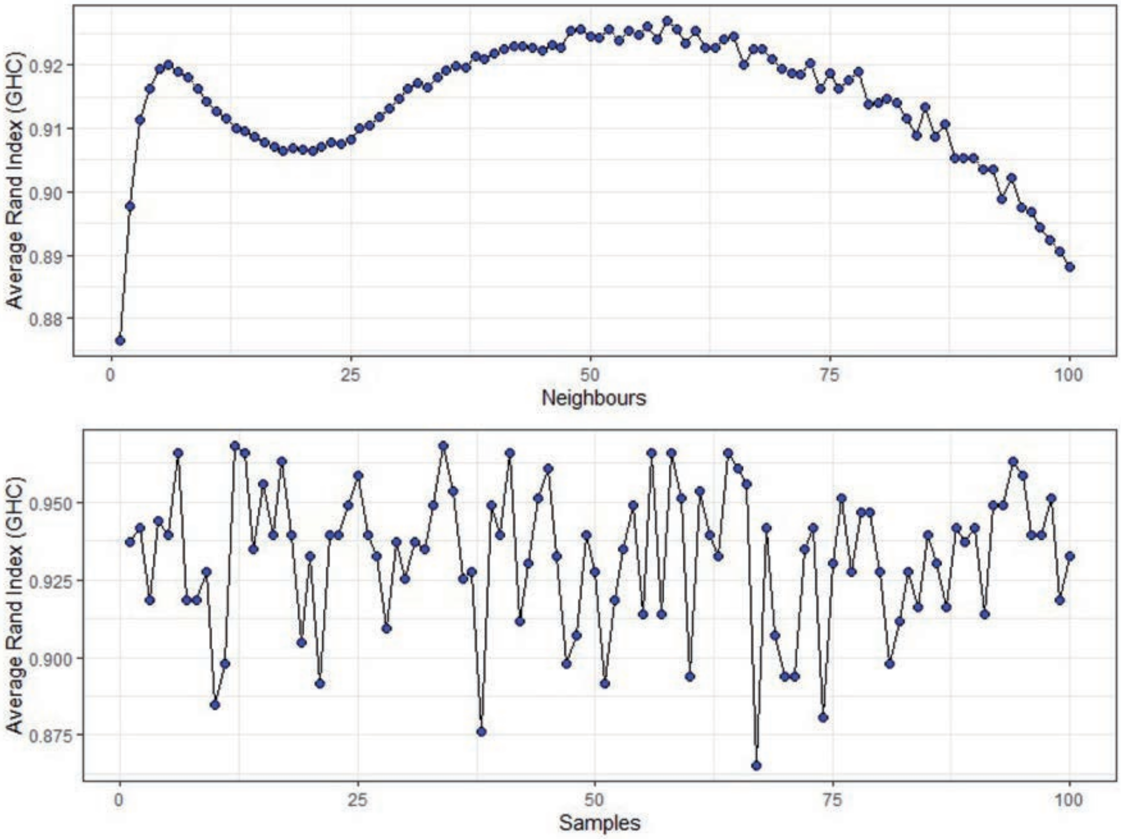


Fig. 8 - The average *RI* of the GHC considering the optimal number of 58 neighbours.

However, in real data sets, predetermined domains for estimating the *RI* to evaluate clustering performance are unavailable. Therefore, based on the synthetic data set's random samplings and performing implementations of the clustering algorithm in different neighbours (1-100) per 1000 samplings, the neighbours' optimal number is determined. According to Fig. 8, the 58 neighbours represent the acceptable performance. Therefore, this number of neighbours is used to cluster the real data set.

Table 3 - GHC`s *RI* parameters considering the optimal number of 58 neighbours.

RI parameters	GHC
Mean	0.93
Variance	0.0004
Standard deviation	0.02
Maximum	0.97
Minimum	0.86

RI's average is calculated over 100 samples by considering the optimal number of neighbours (58), represented in Fig. 8. According to Table 3, the *RI*'s maximum value, in this case, is 0.97, and its minimum is 0.86. The variance and standard deviation are 0.0004 and 0.02, respectively. The mean value of 0.93 indicates the proposed algorithm's desirable performance.

Table 4 - Variogram-based spatial validation of clustering methods for Internal/External Domains.

	Predefined	K-Means	Hierarchical	Spectral	GHC
Sill (%^{2})	0.71 (0.78*)	13 (0.65*)	1.5 (0.71*)	0.81 (0.77*)	0.75 (0.78*)
Range (m)	0.09 (0.13*)	7.5 (0.12*)	0.69 (0.15*)	0.14 (0.13*)	0.11 (0.13*)
Nugget effect (%^{2})	0.0 (0.18*)	0.23 (0.16*)	0.0 (0.22*)	0.0 (0.2*)	0.0 (0.17*)

* External domain

The spatial validation studies of the target variable in obtained clusters are based on the variogram results. The fitted model for the internal domain has a sill of 0.71, a range of 0.09, and a nugget effect of 0.0. The external domain's parameters are equal to 0.78, 0.13, and 0.18, respectively (Table 4). Comparing the variogram results of predetermined domains and the resulting clusters illustrates their spatial continuity as expected. Thus, the GHC algorithm has the highest adaptation in the target variable's variogram model in clusters and predetermined domains. Moreover, the variogram model parameters obtained in the Spectral method indicate the excellent performance of this method compared to the Hierarchical and K-Means clustering.

3.1.5. Certainty application

Fig. 9 shows the internal cluster certainty where smaller observations have a lower certainty while black observations belong to the external cluster. According to the *RI's* average value, 16 neighbours are selected to estimate the certainty of clusters. In addition to providing the concept of certainty, this tool can improve the algorithm's results.

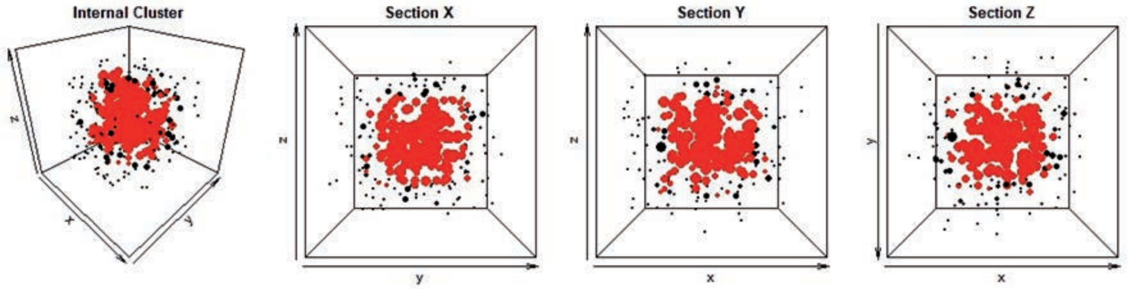


Fig. 9 - The certainty of the internal cluster.

3.1.6. Ordinary kriging estimator

The process of the ordinary kriging estimator is performed on standardised samples and clusters obtained from the GHC algorithm in the following stages:

1. without domaining,
2. with domaining.

The results are analysed using cross-validation, the squares sum (*SS*) of the estimation variance, and the correlation between the estimation results and the primary values. In the first stage, the ordinary kriging estimator directly estimates the study area without considering the clusters. It should be pointed out that a grid with a size of 20×20×20 is used for the estimation procedure in the first step. In the second stage, the domains generated from the clustering algorithm are analysed separately. Therefore, clusters are estimated according to their variogram model, and the estimation grid is calculated individually for each domain.

Table 5 - Estimation validation for the second variable.

Data set	Max.	Min.	Mean	Standard deviation	*SS* of estimation variance	Cross-validation of internal cluster	Cross-validation of external cluster	Correlation
Original	4.72	-3.11	0	1	-	-	-	-
Estimated (1)	3.90	-2.54	-0.001	0.79	2461	0.59	0.68	0.72
Estimated (2)	3.90	-2.54	-0.002	0.80	1664	0.65	0.72	0.80

As Table 5 reports, both stages have similar performance in estimating maximum and minimum values. The *SS* of estimation variance in the first and second cases is 2461 and 1664, respectively, which shows the second method's superiority. The correlation between the estimation results and the primary values in the second stage has increased noticeably, which was equal to 0.72 in the first case and has reached 0.80 with the domaining procedure's aid in the second case. The results of cross-validation also guarantee the positive effect of domaining on the estimation process. Thus, cross-validation results on the second stage for internal and external clusters are equal to 0.65 and 0.72; plus, these values are 0.59 and 0.68 for the first stage.

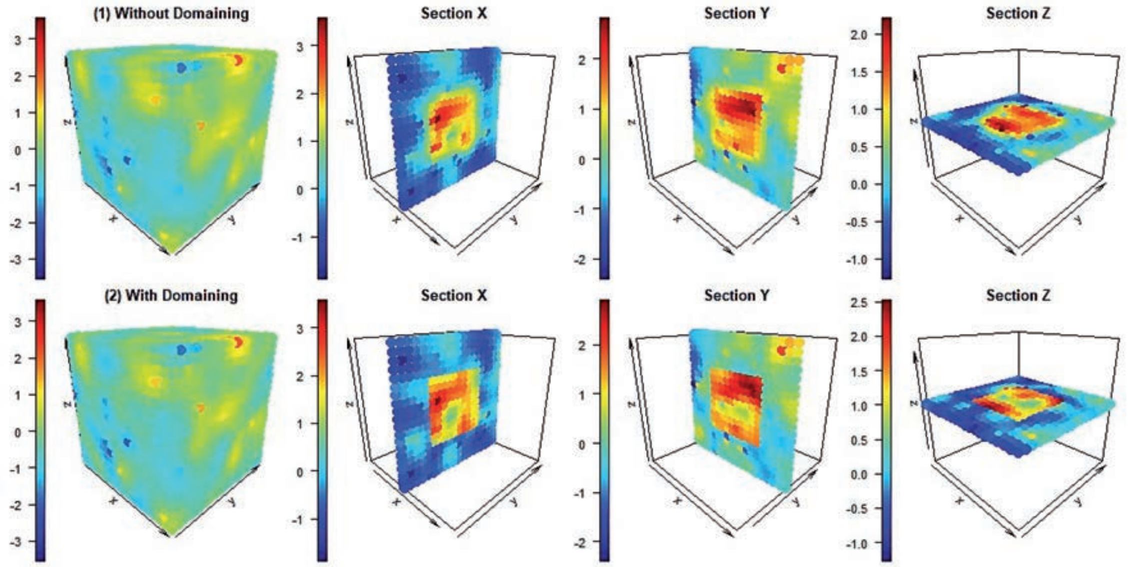


Fig. 10 - Estimation results of the synthetic data set for the second variable.

Table 6 displays the correlation between the estimation results and primary values for each variable, used as a suitable criterion for the estimation's validation. The results demonstrate the superiority of the second method over the first method, and as is clear, the application of domaining by the GHC significantly improved the estimation results, especially in the internal cluster. It is worth mentioning that the low correlation of the third variable can be justified by considering the nature of this variable, which is considered noise.

Table 6 - Correlation of estimated and primary values for each variable.

Estimation methods	**1**	**2**	**3**
Estimation without domaining (1)	0.77	0.72	0.52
Estimation with domaining (2)	0.83	0.77	0.54

3.2. Real data set (Mehdi Abad deposit)

The Mehdi Abad lead and zinc deposits data set, extracted from drilling 56 boreholes, includes 7757 cores with approximately 2 m in length. Samples consist of 6 variables, including the lead, zinc, silver, and X, Y, and Z values as the geographical coordinates.

The exploratory analysis and lithological studies of the study area constitute the basis of this selection. The following are the modifications done for the composite data set with a length of 1 m:

- the normalised value of coordinates,
- the normalised value of lead, zinc, and silver.

In this section, deposit domaining is implemented based on 3 clusters: low-grade, high-grade, and medium-grade. The results and statistics related to each cluster have been listed in Table 7. The red cluster is the low-grade domain that owns 3421 members, outlined by the low grade of lead (1.16%) and zinc (1.98%) located NW of the deposits. The green cluster is a high-grade

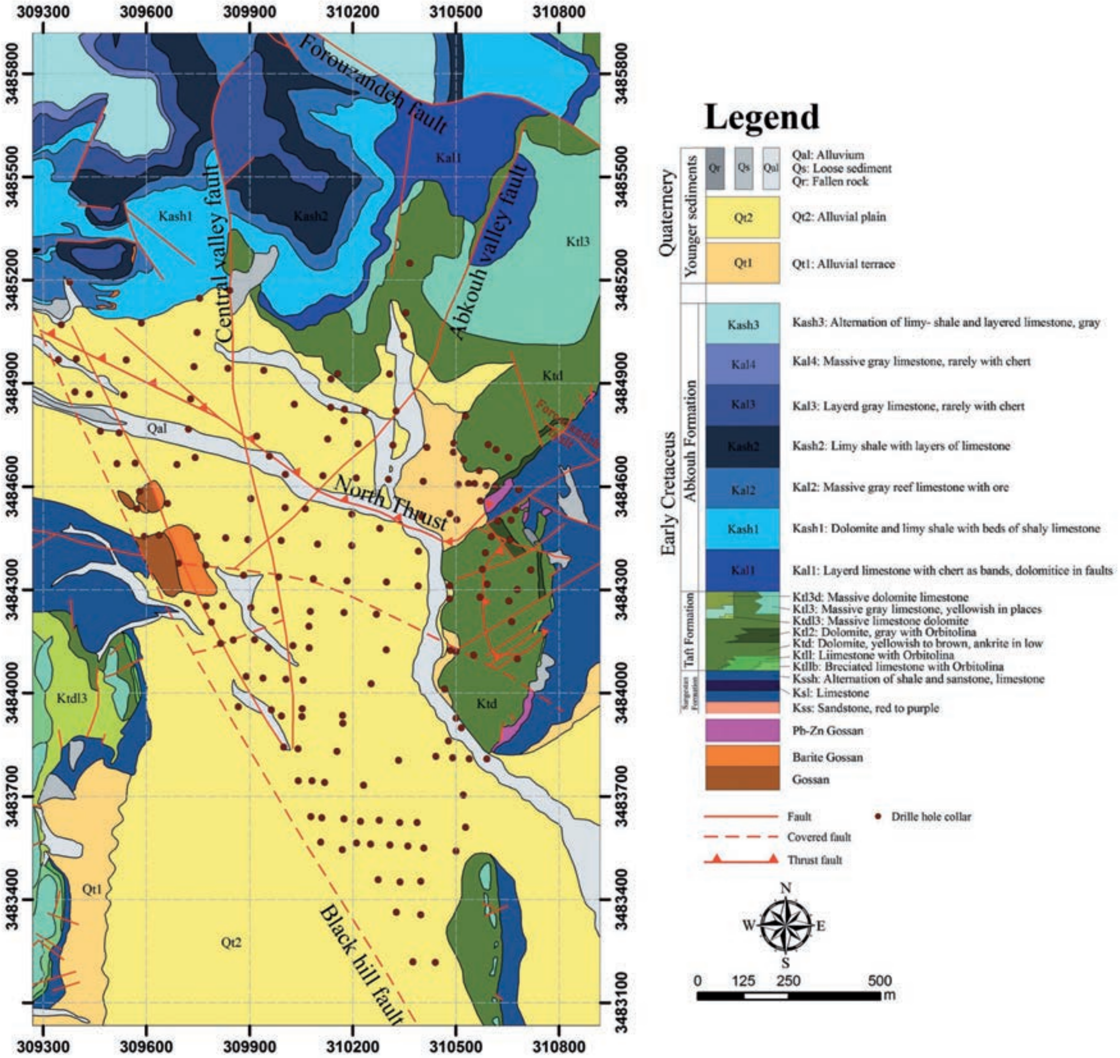

309300
309600
309900
310200
310500
310800
3485800
3485500
3485200
3484900
3484600
3484300
3484000
3483700
3483400
3483100
Forouzandeh fault
Kal1
Kash1
Kash2
Central valley fault
Abkouh valley fault
Ktl3
Ktd
Qal
North Thrust
Ktdl3
Ktd
Qt1
Qt2
Black hill fault
Legend
Quaternery
Younger sediments
Qr
Qs
Qal
Qal: Alluvium
Qs: Loose sediment
Qr: Fallen rock
Qt2
Qt2: Alluvial plain
Qt1
Qt1: Alluvial terrace
Early Cretaceus
Abkouh Formation
Kash3
Kash3: Alternation of limy- shale and layered limestone, gray
Kal4
Kal4: Massive gray limestone, rarely with chert
Kal3
Kal3: Layerd gray limestone, rarely with chert
Kash2
Kash2: Limy shale with layers of limestone
Kal2
Kal2: Massive gray reef limestone with ore
Kash1
Kash1: Dolomite and limy shale with beds of shaly limestone
Kal1
Kal1: Layerd limestone with chert as bands, dolomitice in faults
Taft Formation
Ktl3d: Massive dolomite limestone
Ktl3: Massive gray limestone, yellowish in places
Ktdl3: Massive limestone dolomite
Ktl2: Dolomite, gray with Orbitolina
Ktd: Dolomite, yellowish to brown, ankrite in low
Ktll: Liimestone with Orbitolina
Ktllb: Breciated limestone with Orbitolina
Kssh: Alternation of shale and sanstone, limestone
Ksl: Limestone
Kss: Sandstone, red to purple
Pb-Zn Gossan
Barite Gossan
Gossan
Fault
Covered fault
Thrust fault
Drille hole collar
N
W
E
S
0
125
250
500
m

Fig. 11 - Geological map of the Mehdi Abad deposit.

domain with 308 members distinguished by high-grade lead (4.31%), zinc (11.38%), and silver (166.22 g/t) located somewhere deep in the central part of the deposit. Finally, the blue one covers the medium-grade domain with 1.52% lead, 4.03% zinc, and 37.99 g/t silver.

Table 7 - Statistic parameters of variables for each cluster.

GHC	Red cluster (n = 3421)		Green cluster (n = 308)		Blue cluster (n = 4028)	
	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Pb (%)	1.16	1.25	4.31	2.92	1.52	1.37
Zn (%)	1.98	1.94	11.38	8.55	4.03	3.39
Ag (g/t)	33.75	31.30	166.22	121.58	37.99	37.64

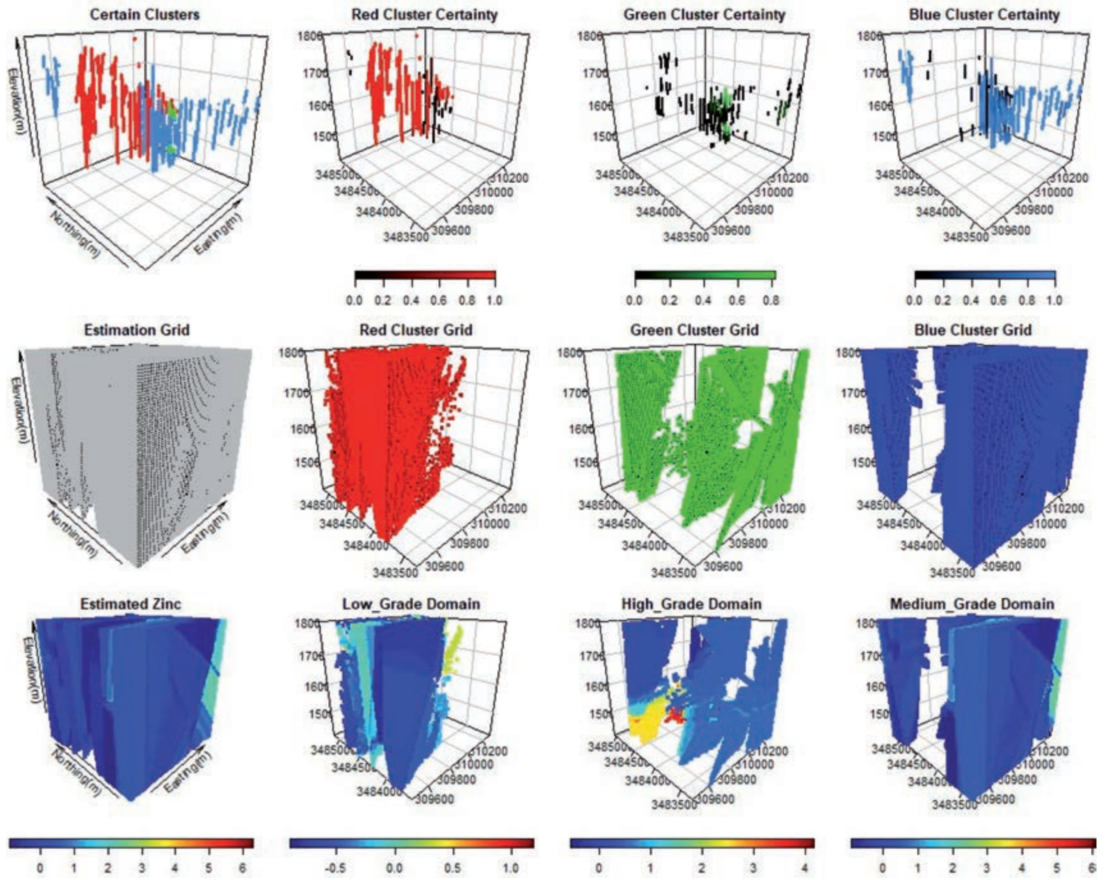


Fig. 12 - GHC resulting clusters and their certainty (1st row), cluster grids and their certainty (2nd row), estimation results of zinc (3rd row).

As illustrated, Fig. 12 indicates the observations of each three clusters attained from the GHC algorithm and shows the position of the drilling boreholes; moreover, the visual view of each observation and cluster's certainty is shown. The second row of this figure designates the designed grids and their certainty, which presents a complete overlap in the study area so that the sum of designed grids for each domain is equal to the grid of the whole mentioned area. It should be noted that the missing parts are not estimated, and the sides of each block are equal to 10, 10, and 5 m in the direction E-W, N-S, and the depth, respectively. Finally, the third row indicates the results derived from the estimation process in two steps:

1. without domaining,
2. with domaining.

Table 8 includes statistical parameters such as statistics of primary standardised values and estimated values, the estimation's *SS* of the variance, and cross-validation results. As shown from this table, the estimation method utilising domaining in its procedure has better performance in estimating the maximum, minimum, and standard deviation. Even so, the outcomes of the cross-validation highlight that the second method is superior compared to the estimation method without domaining. Note that the cross-validation has been calculated according to the whole data set. Indeed, all primary observations are removed, according to the attained parameters of estimating the study area; thereupon, the mentioned parameters estimate the removed observations again in corresponding points. Ultimately, the correlation of estimation outcomes and the primary values are evaluated.

Table 8 - Validation of estimation results according to normalised variables of Zn, Pb, and Ag.

Data set	Max.	Min.	Mean	St. dev.	SS of estimation variance	Cross-validation of red cluster	Cross-validation of green cluster	Cross-validation of blue cluster
Original (Zn)	7.74	-0.9	0	1	-	-	-	-
Estimated_1 (Zn)	6.30	-0.9	-0.01	0.66	$3.11e^{12}$	0.87	0.90	0.86
Estimated_2 (Zn)	6.8	-0.9	0.07	0.73	$3.11e^{12}$	0.90	0.93	0.89
Original (Pb)	14.3	-0.96	0	1	-	-	-	-
Estimated_1 (Pb)	3.73	-0.95	-0.02	0.68	$3.12e^{12}$	0.70	0.66	0.70
Estimated_2 (Pb)	9.32	-0.96	-0.25	1.03	$3.07e^{12}$	0.79	0.70	0.76
Original (Ag)	16	-0.83	0	1	-	-	-	-
Estimated_1 (Ag)	6.66	-0.83	-0.1	0.63	$3.12e^{12}$	0.81	0.80	0.87
Estimated_2 (Ag)	7.2	-1.13	0.13	0.99	$3.08e^{12}$	0.84	0.82	0.85

4. Conclusions

The proposed model-free GHC algorithm provides the spatial continuity of resulting clusters by considering the spatial structure built from the spatial similarity of observations. Coherent, meaningful, and spatially connected clusters extracted by the algorithm reveal the underlying and hidden structures of the data sets. The performance of the proposed approach on synthetic data sets compared to the K-Means, Hierarchical, and Spectral clustering algorithms confirm the superiority of the GHC. Moreover, resulting clusters of real data sets represent significantly meaningful domains. Besides, due to the algorithm's hierarchical nature, it is possible to examine the clustering steps at different levels and provide multiple interpretations at each level. Also, with adequate knowledge of the nature of variables, it is possible to indicate their importance in classifying domains. Therefore, more effective clusters are derived from applying appropriate weighting for each variable. Hence, the Attribute Kriging algorithm provides acceptable performance in both synthetic and real data sets with the help of the GHC as an appropriate tool for automatic domaining of the study area. Attribute Kriging due to the basic needs of mineral resource modelling projects for domaining provides a smoother path to estimating mineral resources and brings more favourable results.

The GHC's accuracy depends on selecting the optimal number of adjacent neighbours to achieve the data set's most accurate spatial structure. Therefore, improving the spatial similarity calculation leads to determine the optimal number of adjacent neighbours to achieve the best possible results. The proposed clustering algorithm assigns observations to one and only one cluster. Thus, it cannot provide the uncertainty of each observation to the clusters. However, the certainty application presented leads to determine the certainty of boundary observations, which is also a proper tool for correcting misclassified observation. To enhance the certainty of clusters, one should investigate fuzzy clustering and model-based geostatistical algorithms, although the latter usually fall short of properly establishing spatial continuity of the resulting clusters according to the literature.

Acknowledgments. The authors are grateful for the data sets provided by the Mehdi Abad mining complex.

REFERENCES

Allard D. and Guillot G.; 2000: *Clustering geostatistical data*. In: Proc. 6th Geostatistics Congress, Cape Town, South Africa, pp. 49-63.

Ambroise C., Dang M. and Govaert G.; 1995: *Clustering of spatial data by the EM algorithm*. In: Soares A., Gómez-Hernandez J. and Froidevaux R. (eds), geoENV I - Geostatistics for Environmental Applications, Quantitative Geology and Geostatistics, Springer, Dordrecht, Germany, vol. 9, pp. 493-504, doi: 10.1007/978-94-017-1675-8_40.

Caliński T. and Harabasz J.; 1974: *A Dendrite method for cluster analysis*. Commun. Stat., 3, 1-27, doi: 10.1080/03610927408827101.

Chavent M., Kuentz-Simonet V., Labenne A. and Saracco J.; 2018: *ClustGeo: an R package for hierarchical clustering with spatial constraints*. Comput. Stat., 33, 1799-1822, doi: 10.1007/s00180-018-0791-1.

Emery X. and Ortiz J.M.; 2004: *Defining geological units by grade domaining*. Department of Mining Engineering, University of Chile, Santiago del Chile, Chile, Technical report, pp. 1-13.

Fouedjio F.; 2016a: *A clustering approach for discovering intrinsic clusters in multivariate geostatistical data*. In: Lecture Notes in Computer Science, Perner P. (ed), Machine Learning and Data Mining in Pattern Recognition, New York, NY, USA, vol. 9729, pp. 491-500, doi: 10.1007/978-3-319-41920-6_39.

Fouedjio F.; 2016b: *A hierarchical clustering method for multivariate geostatistical data*. Spatial Stat., 18, 333-351, doi: 10.1016/j.spasta.2016.07.003.

Madenova Y. and Madani N.; 2021: *Application of Gaussian mixture model and geostatistical co-simulation for resource modeling of geometallurgical variables*. Nat. Resour. Res., 30, 1199-1228, doi: 10.1007/S11053-020-09802-4.

Oliver M.A. and Webster R.; 1989: *A geostatistical basis for spatial weighting in multivariate classification*. Math. Geol., 21, 15-35, doi: 10.1007/BF00897238.

Pawitan Y. and Huang J.; 2003: *Constrained clustering of irregularly sampled spatial data*. J. Stat. Comput. Simul., 73, 853-865, doi: 10.1080/0094965031000099131.

Romary T., Rivoirard J., Deraisme J., Quinones C. and Freulon X.; 2012: *Domaining by clustering multivariate geostatistical data*. In: Abrahamsen P., Hauge R. and Kolbjørnsen O. (eds), Quantitative Geology and Geostatistics, Oslo, Norway, vol. 17, pp. 455-466, doi: 10.1007/978-94-007-4153-9_37.

Romary T., Rivoirard J. and Deraisme J.; 2015: *Unsupervised classification of multivariate geostatistical data: two algorithms*. Comput. Geosci., 85, 96-103, doi: 10.1016/j.cageo.2015.05.019.

Stegman C.L.; 2001: *How domain envelopes impact on the resource estimate-case studies from the Cobar gold field, NSW, Australia*. In: Edwards A.C. (ed), Mineral Resource and Ore Reserve Estimation - The AusIMM Guide to Good Practice, Australasian Institute of Mining and Metallurgy, Melbourne, Australia, Monograph Series 23, pp. 221-236.

von Luxburg U.; 2007: *A tutorial on spectral clustering*. Stat. Comput., 17, 395-416, doi: 10.1007/s11222-007-9033-z.

Corresponding author: Omid Asghari
Simulation and Data Processing Laboratory, School of Mining Engineering, College of Engineering, University of Tehran
North Kargar St., Tehran, Iran
Phone: +98 912 2183503; e-mail: o.asghari@ut.ac.ir