Natural Resources Research (© 2018)
https://doi.org/10.1007/s11053-018-9421-4

Original Paper

Capturing Hidden Geochemical Anomalies in Scarce Data by Fractal Analysis and Stochastic Modeling

Nasser Madani[1,3] and Behnam Sadeghi[2]

Received 25 May 2018; accepted 13 October 2018

Fractal/multifractal modeling is a widely used geomathematical approach to capturing different populations in geochemical mapping. The rationale of this methodology is based on empirical frequency density functions attained from global or local distributions. This approach is quite popular because of its simplicity and versatility; it accounts for the frequency and spatial distribution of geochemical data considering self-similarity across a range of scales. Using this technique for detection of geochemical anomalies in scarce data, however, is problematic and can lead to systematic bias in the characterization of the underlying populations. In this paper, an innovative technique is presented that provides good results without a priori assumptions. A simulation approach is adopted for fractal analysis by generating different possible distribution scenarios for the variable under study to reveal the underlying populations that are frequently hidden due to lack of data. The proposed technique is called the global simulated size–number method, and it is validated in a case study with two synthetic datasets and another case study with real dataset from the Ushtagan gold deposit in northeast Kazakhstan.

KEY WORDS: Fractal modeling, Monte Carlo simulation, Kernel density function, Ushtagan gold deposit.

INTRODUCTION

Various mathematical and statistical methods have been applied to generate accurate geochemical or geophysical anomaly maps using the data provided and based on the frequency and spatial distribution of geochemical data. For example, traditional statistical methods of exploratory data analysis (EDA) (Tennant and White 1959; Hawkes and Webb 1962; Tukey 1977) and modern techniques such as fractal analysis (Mandelbrot 1983) have been used. Among them, fractal/multifractal modeling can be more robust in identifying significant geochemical anomalies because traditional statistical methods such as EDA are based mainly on geochemical value frequency distributions, but neglect the spatial variation of geochemical data.

Fractal/multifractal modeling accounts for both the frequency distribution and spatial variation of the geochemical values (Mandelbrot 1983; Feder 1988; Cheng 2012; Zuo and Wang 2016). In a nutshell, fractal/multifractal modeling considers both statistical and spatial distributions of geochemical data (e.g., Sadeghi et al. 2012, 2015; He et al. 2013; Luz et al. 2014). These properties are based mainly on self-similarity or self-affinity (i.e., statistical self-similarity at different scales (Mandelbrot 1983)). The most significant fractal models that have been developed to delineate geochemical anomalies can

[1]Department of Mining Engineering, School of Mining and Geosciences, Nazarbayev University, Astana, Kazakhstan.
[2]School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia.
[3]To whom correspondence should be addressed; e-mail: Nasser.Madani@nu.edu.kz

Published online: 22 October 2018

be classified into two families. The first family is based on posterior analysis or conditional distribution, such as concentration–area (C-A) and concentration–perimeter (C-P) fractal models (Cheng et al. 1994, 1996), spectrum–area (S-A) fractal models (Cheng et al. 1999, 2000), concentration–volume (C-V) fractal models (Afzal et al. 2010), spectrum–volume (S-V) fractal models (Afzal et al. 2011), simulated size–number (SS-N) fractal models (Sadeghi et al. 2015) and wavelet-number (W-N) fractal model (Chen and Cheng 2018). In these methodologies, the spatial geometry of the geochemical landscape is considered for either deterministic or stochastic mapping to show the local distribution of the attribute under study and whether it is applicable for detecting different anomalies based on further fractal analysis. The second family involves the analysis of a priori or global distributions, such as the number–size (N-S) fractal model (Mandelbrot 1983), regarding a density function with statistical parameters informed by a sample histogram. In this case, sample locations are not important; one needs to analyze only the global distribution of the samples and variables.

In general, fractal methods tend to be erratic when data are limited and result in some sawtooth-like spikes in the distribution that make it non-representative. These artifacts lead to further suspicion of the N-S fractal model (Mandelbrot 1983) because they may potentially mask the actual thresholds and populations of interest. To circumvent this problem, one idea is to smooth out the histogram, particularly to reform the shape of the fluctuations and increase the resolution of the distribution (Pyrcz and Deutsch 2014; Rossi and Deutsch 2014). Certain techniques are available for this type of smoothing. The most straightforward approach is to fit a predefined parametric density function such as a power, normal or lognormal distribution to the sample data distribution (Johnson and Kotz 1970; Scott 1992; Borradaile 2003). Although these parametric models resolve spikes in a sample histogram, earth-related data in fact rarely display parametric behavior. Other solutions include nonparametric approaches, such as simulated annealing (Journel and Xu 1994; Deutsch 1996) or quadratic programming (Xu and Journel 1995) and optimization algorithms, which are based on closeness to input quartiles, target mean and variance values (Deutsch and Journel 1998). Kernel density estimation (Fix and Hodges 1951) is an alternative nonparametric approach that fits smoothed probability and cumulative distribution functions (pdf and cdf) to empirical distribution functions (Silverman 1986; Scott 1992; Altman and Leger 1995; Sheikhpour et al. 2017). This technique requires a kernel density function, an optimal measure of bandwidth and a dataset (Wolfgang et al. 2004; Alexandre 2009; Samawi et al. 2016).

The algorithm proposed in this paper is based on the estimation of the density function model for scarce data using a kernel estimator. Using the estimated function, the possible scenarios of the underlying distribution can then be generated by applying Monte Carlo simulation. Once the simulated values are statistically qualified, the fractal analysis paradigm (SS-N), following Sadeghi et al. (2015), can be implemented to differentiate the geochemical anomalies. The resulting technique is called the global simulated size–number (GSS-N) method. The main difference between the SS-N method (Sadeghi et al. 2015) and the algorithm (GSS-N) is that the GSS-N method is based on global distribution, which is independent in terms of spatial continuity, whereas the SS-N method is based on local distribution, in which it is necessary to account for the spatial geometry of the geochemical landscape.

This paper will first show how the proposed innovative algorithm works by using two synthetic case studies, and then apply it to an actual case study from Kazakhstan to capture from just a few trench samples the alternative populations associated with a gold deposit. These case studies show that the GSS-N method can be applied to the pre-feasibility or feasibility study stages of a project that is faced with a problem of scarce data. The results of this study are intended to be useful guiding further exploration.

METHODOLOGY

Number–Size (N-S) Fractal Method

The N-S method, proposed by Benoit Mandelbrot (1983), is a model based on the relationship between the cumulative number of samples and their related size, the latter representing metal concentration in this research (Li et al. 1994; Turcotte 1996; Shi and Wang 1998; Sanderson et al. 1994; Zuo et al. 2009; Sadeghi et al. 2012, 2015). Considering Eq. 1, the number of samples with higher concentrations is less than the number of the samples with lower concentrations. Monecke et al.

(2005) used the N-S method to describe enrichment of minerals based on replacement by metasomatic processes that cause the formation of hydrothermal deposits in the Waterloo massive-sulfide deposit in Australia. Sadeghi et al. (2012) applied this model in 3D for the first time to separate mineralized zones and wall rocks and then in 2015 improved the model by combining the N-S model with simulation methods, resulting in the proposal of the SS-N model.

In the equation developed by Mandelbrot (1983), the minus sign denotes the inverse relationship between the number of samples and their associated concentrations:

$$N(\geq \rho) = F\rho^{-D} \tag{1}$$

where ρ is element concentration, $N(\geq \rho)$ is cumulative number of samples with concentration values greater than or equal to ρ, F is a constant, and D is the scaling exponent or fractal dimension of the spatial distribution of element concentrations. Log-log plots of N vs. ρ are used because if logarithms are applied to Eq. 1, the outcome is a straight-line equation. After generating log-log plots, they are fitted with straight lines to find the intersection points, which represent thresholds. The final interpolated map is classified using the obtained thresholds.

Kernel Density Estimator

Kernel densities are nonparametric techniques for smoothing histograms (Scott 1992). A kernel density estimator is applicable whenever sawtooth-like spikes appear in the distribution due to data scarcity. The concept of this estimator is to ''fill in'' gaps in raw data distributions (Leuangthong and Deutsch 2003). Given this background, a kernel density estimator can predict the pdf and cdf of a set of random data. The estimator formulas for any real values of u are given in the literature (Hill 1985; Silverman 1986; Jones 1993; Bowman and Azzalini 1997).

$$\text{In pdf}: \hat{f}(u) = \frac{1}{n\Delta z}\sum_{i=1}^{n} K\left(\frac{u-u_i}{\Delta z}\right) \tag{2}$$

$$\text{In cdf}: \hat{f}(u) = \frac{1}{n}\sum_{i=1}^{n} G\left(\frac{u-u_i}{\Delta z}\right) \tag{3}$$

$$G(u) = \int_{-\infty}^{n} K(t)\cdot \mathrm{d}t$$

where $u_1, u_2,\ldots,u_n$ are random samples from an unknown distribution, n is number of data points, and Δz is bandwidth obtained by splitting the range of the data between the maximum and minimum observed values (Izenman 1991), which controls the smoothness of the density curve. Attention must be paid to defining the bandwidth because if it is large it generates a very smooth kernel function. $K(.)$ is the kernel function, which is nonnegative and linked to a subset of a particular density function such as a normal, box, triangle or Epanechnikov function (Epanechnikov 1969). $\hat{f}(u)$ is an estimate of the density function obtained by placing the underlying kernel function $K(.)$ over each random sample u_i in a dataset. Therefore, the kernel function $K(.)$ is a density function with location parameter u and the bandwidth of interest. Because the resulting estimate $\hat{f}(u)$ is obtained from summation of sub-kernel functions $K(.)$, the selection of $K(.)$ is not critical and quite often a normal function can be used for the sake of simplicity (Wand and Jones 1995). In other words, by defining each of those mentioned underlying functions, the final kernel density estimate will be closely related to the desired histograms and similarly will represent the probability distribution of the sample data with small differences. A kernel distribution results in a continuous and smooth probability curve, in contrast to a histogram, which builds discrete bins, and places each data value in the relevant bin. Through a small example, Figure 1 shows the visual comparison between a kernel fitted distribution and the related histogram from three arbitrary sample data: $u_1 = 13$, $u_2 = 29$, and $u_3 = 29$. For construction of a typical global distribution, the abscissa of the histogram is divided into subintervals or bins, which span the range of the data (Fig. 1a). As seen in Figure 1a, with only a few data points, the histogram generates a discrete probability function unsuitable for specific applications such as differentiating the alternative populations by fractal analysis. Figure 1b shows the overall fitted probability distribution function (solid line) by a normal kernel function that creates an individual probability density curve (dashed lines) for each data value and then sums the tiny smooth curves to form a unique, smooth, continuous probability density function for the entire data set. The tiny kernel functions in this

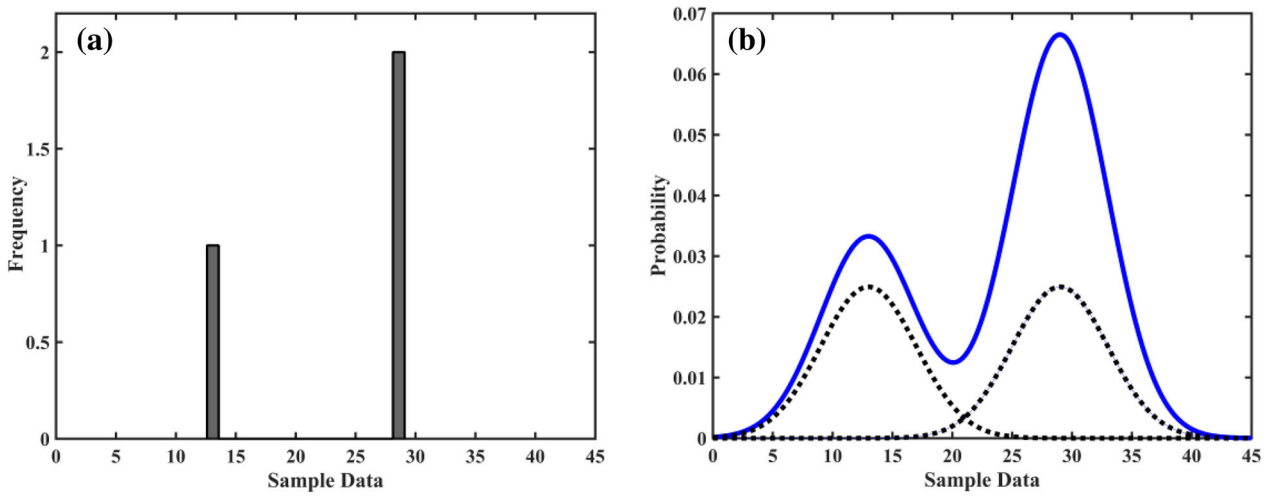


Figure 1. (**a**) Histogram and (**b**) fitted model by normal kernel density function to three sample dataset.

case are normal, but their summations according to Eq. 2 result in a bimodal distribution, which is not necessarily normal but follows the desired distribution.

Monte Carlo Simulation

Monte Carlo simulation is a broad term for computational algorithms that generates random numbers (realizations) given a specific density function (Pyrcz and Deutsch 2014). The Monte Carlo simulation of random numbers from an arbitrary probability or cumulative distribution function can be obtained by Deutsch and Journel (1998) (Fig. 2):

1) Generation of a random number q, which is uniformly distributed between zero and one, $q\epsilon[0,1]$
2) Retrieving the inverse of the cumulative probability distribution (cdf) function:

$$y = f^{-1}(q) \tag{5}$$

where y is the inverse cumulative probability distribution $f^{-1}(.)$. According to the ergodic theory (Chilès and Delfiner 2012), in order to converge the statistical parameters (such as the mean and variance) of simulated values to the original dataset, it is typically necessary to simulate a large number of values.

Global Simulated Size–Number (GSS-N) Method

As mentioned previously, scarce dataset will result in a poorly formed global distribution function, leading to general problems when using further analysis. However, fractal and multifractal techniques need a trustworthy distribution to precisely delineate the domain of geochemical populations. The innovative GSS-N method in this paper is based on applying a smoothing technique (kernel density) to circumvent the problem of discrete probability functions, which stems from scarce data. The proposed algorithm is described as follows:

a) *Generating the histogram of the data* In this step, the histogram of the available dataset is checked to determine whether the distribution is representative. To implement this, some criteria such as the coefficient of variation and kurtosis of the data are used. The former is the standard deviation normalized by a mean; it is a unitless measure of the variability (Davis 1986; Rossi and Deutsch 2014). The kurtosis is a good measure to ascertain whether the data distribution is flat or peaked. A negative kurtosis indicates that most samples are concentrated in the center of the distribution, making the shape of the histogram in the form of a sawtooth.
b) *Empirical cumulative distribution function* This step includes calculation of a primary nonparametric estimate of the true cdf of a

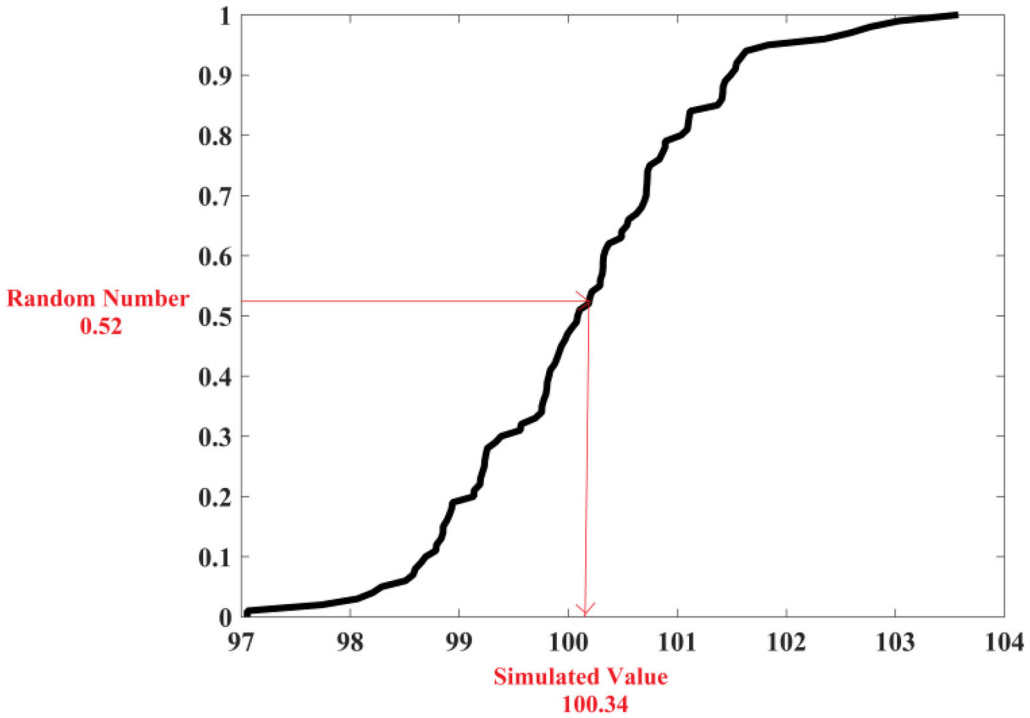


Figure 2. Graphical illustration of Monte Carlo simulation algorithm.

variable. In this graph, the abscissa shows data values ordered from smallest to largest and the ordinate represents the cumulative probability assigned to each data point (Davis 1986). When the sample size of the data is small, the curve gives a sawtooth-like shape and has a very spiky stair-step appearance.

c) *Kernel density function modeling* This step applies one of the kernel functions to fit a cdf model to the obtained experimental cdf from the previous step. In this context, three items are significant: type of function, bandwidth and range of the data. The commonly used functions are the normal, box, triangle and Epanechnikov functions. For instance, the box kernel produces a density curve that is less smooth than the others. For the selection of the optimum function and the bandwidth, one can apply the cross-validation technique (Liu et al. 2014), plug-in methods (Tenreiro 2017) or even visual inspection, which provides a qualitative benchmark and can be accomplished through trial-and-error. The tentative range of the data in a model can be either identical to the original dataset or be chosen arbitrarily according to geological constraints. All these parameters should be considered cautiously to prevent over- and under-smoothing.

d) *Monte Carlo simulation* In this step, random numbers uniformly distributed between 0 and 1 are generated and then their inverse values are derived over the kernel fitted model as simulated values (samples). The realizations are now prepared to calculate the empirical cumulative distribution function of the simulation results. However, one audit should be made at this point to evaluate the average of the computed empirical simulated cdf. This curve should converge to the fitted kernel density function according to the ergodic theory (Chilès and Delfiner 2012). For fast, effective convergence, as large as possible a sample size is recommended, with as many realizations as possible.

e) *Fractal analysis* The rationale of this step is to implement fractal analysis on the simulated values. Sadeghi et al. (2015) showed

that the SS-N fractal model (a combination of geostatistical simulation and fractal/multifractal analysis) is capable of delineating mineralized zones better than deterministic paradigms. This approach employs different realizations of the spatial variability beyond geostatistical simulation of the attribute under study. The main difference of the algorithm proposed in this paper and the SS-N algorithm is that the GSS-N is implemented on the realizations obtained from the global distribution of the underlying variable but not on the local distribution as in the SS-N method. In the GSS-N method, each realization results in one fractal curve and once the average of those curves is known, one is able to differentiate the possible thresholds on the averaged curve.

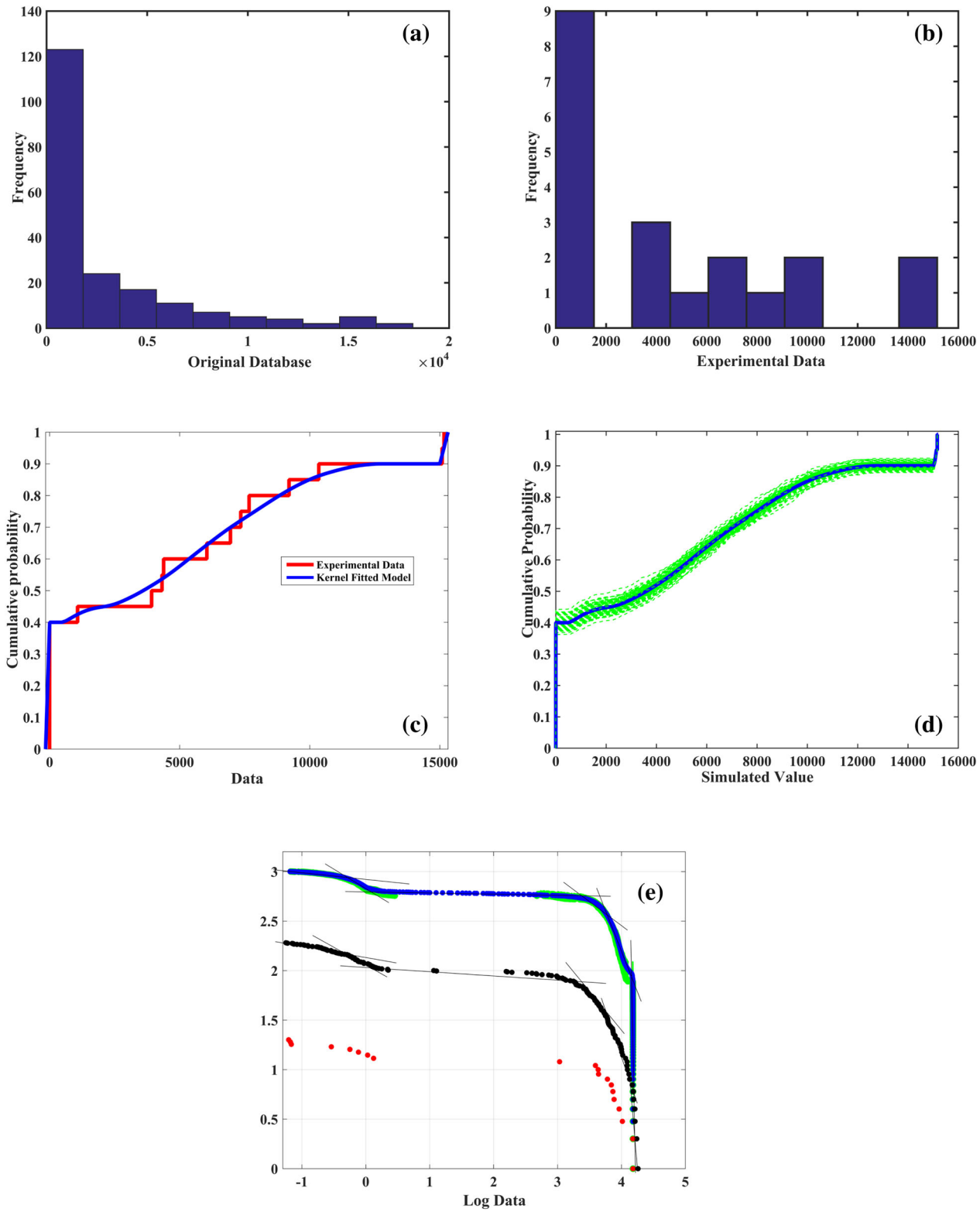
CASE STUDY

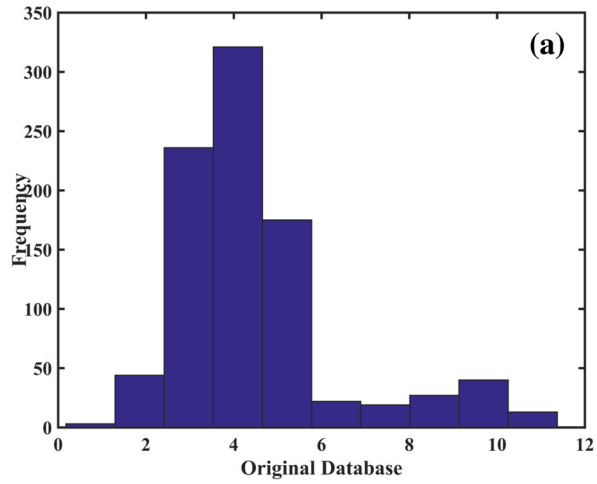
To demonstrate the capability of the proposed method, it is tested in two case studies. The first is based on two synthetic datasets, and the second employs an actual case study. In both case studies, relative discussions are provided.

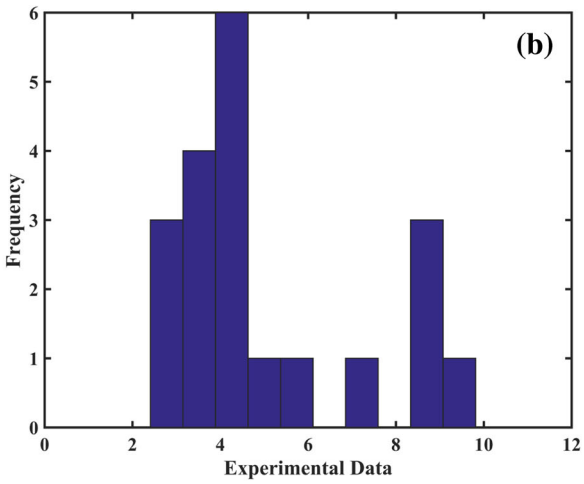
Case Study with Synthetic Datasets

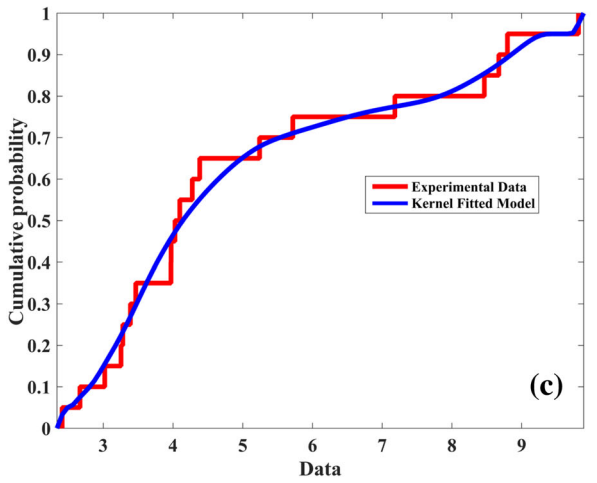
Two different distributions were considered to generate synthetic datasets suitable for applying the proposed algorithm. The first distribution (case I) consists of 200 random numbers that follow an exponential distribution, composed of two major populations with different mean (i.e., 0.5 and 5000). The second distribution (case II) consists of 900 normal random values composed of two major populations with different mean (i.e., 4 and 9). Aside from having different distribution types, the two major populations were deemed to be hidden in cases I and II but are distinguishable through visual inspection of the histogram (Figs. 3a and 4a). These two distributions were selected as references, and then 20 samples were drawn randomly from each distribution to produce two series of scarce data called ''experimental data''. Using this method, one can ensure that the histogram of either set of experimental data is not continuous; besides, scarce data dramatically impact the shapes of the distribution (Figs. 3b and 4b). As a consequence, the empirical cdf is expected to display a sawtooth shape. Figures 3c and 4c (red solid lines) show that the resulting probability distribution functions computed from those scarce data are not trustworthy and justify the need to use some smoothing technique for fitting a theoretical cdf model (Figs. 3c and 4c: blue solid lines). The Epanechnikov kernel function (a nonparametric approach) was then considered to fit a model to an empirical cdf. As already discussed, this function does not have a distinct difference compared with other functions. Following the algorithm steps described earlier, 1000 numbers between 0 and 1 uniformly generated, and then equivalently, 1000 simulated values were inversely drawn from the kernel fitted model by Monte Carlo simulation. For fast, reliable convergence of the simulated values to the model, 100 total realizations were considered a sufficient number for such an examination. Figures 3d and 4d illustrate the empirical cumulative distribution functions for all the realizations (dashed green lines) and their respective average (blue solid line) that fairly converges to the fitted model (Figs. 3c and 4c: blue solid line). This convergence ensures that the realizations are statistically sound and they can be taken into account for further processing. Fractal analysis was implemented in each realization separately to obtain their relevant curve. Those curves were then averaged to obtain one unique curve, so as to account for defining the thresholds in each population (Figs. 3e and 4e). As shown in this figure, the shape of the fractal curve in the simulated results mimics the reference models and, consequently, the thresholds retrieved from the average realization curve are similar to those from the reference models. However, fractal analysis of scarce experimental data does not provide adequate knowledge for differentiating the populations (note the red points in the fractal sheet). Two synthetic case studies verify that the proposed algorithm comprises a safe paradigm to tackle the problem of data scarcity for fractal/multifractal analysis considering two different common global distributions (exponential and normal).

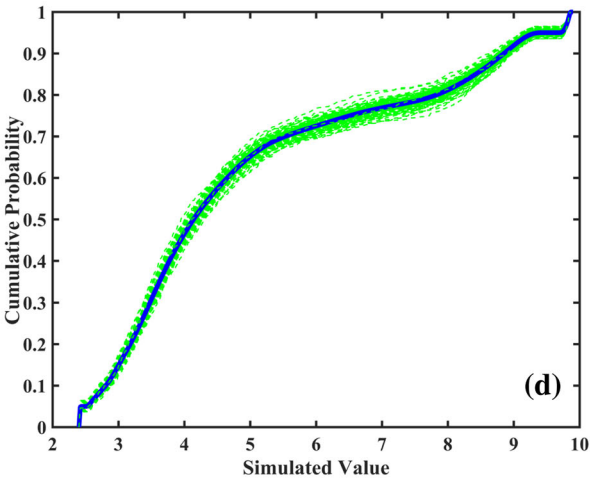
Figure 3. Synthetic case study I, exponential distribution; (**a**) ► Reference distribution, (**b**) experimental data, (**c**) cumulative distribution function, (**d**) cumulative distribution function of the realizations and their checking, (**e**) fractal analysis on the simulated values (blue points), reference model (black points) and the scarce data (red points).

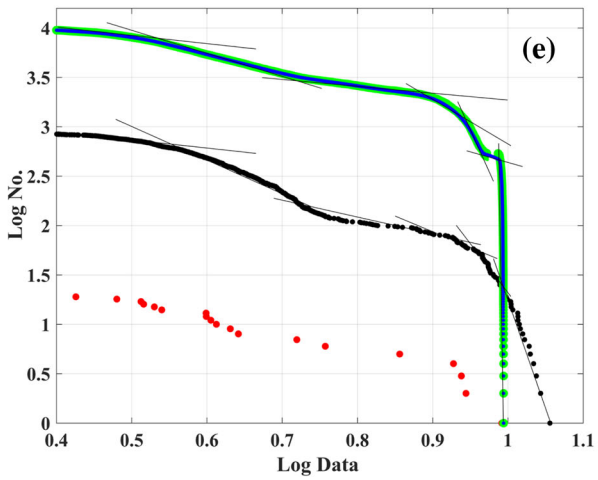

(a)
Frequency
Original Database
×10^4
(b)
Frequency
Experimental Data
(c)
Cumulative probability
Experimental Data
Kernel Fitted Model
Data
(d)
Cumulative Probability
Simulated Value
(e)
Log Data


(a)
Frequency
Original Database


(b)
Frequency
Experimental Data


(c)
Cumulative probability
Experimental Data
Kernel Fitted Model
Data


(d)
Cumulative Probability
Simulated Value


(e)
Log No.
Log Data

◀**Figure 4.** Synthetic case II, normal distribution; (**a**) Reference distribution, (**b**) experimental data, (**c**) cumulative distribution function, (**d**) cumulative distribution function of the realizations and their checking, (**e**) fractal analysis on the simulated values (blue points), reference model (black points) and the scarce data (red points).

Case Study with Real Dataset (Ushtagan Gold Deposit)

A real dataset from Ushtagan gold deposit in Kazakhstan is used here. This case study is presented as an interesting example of scarce data because this phenomenon (lack of data) is a very common characteristic of the majority of gold deposits, particularly in the feasibility study phase of a project.

The Ushtagan gold deposit is located in northeast Central Kazakhstan and administratively is situated in the Bayanaul district of the Pavlodar region (Fig. 5). The Pavlodar region is one of the main industrialized regions of Kazakhstan with developed mining, fuel and energy sectors, and a multisector industrial complex. From a geological setting standpoint, the Ushtagan deposit is located in the middle-Devonian volcanic system. The central part of the volcanic system is represented by a stock of plagiogranite–porphyrites (Fig. 6). Different fragmental tuffogenic rocks with rare interlayers of effusive formations of acid and intermediate composition are common in the volcanogenic strata (Fig. 6), which are sometimes strongly silicified, tourmalinitized and pyritized. In present-day terrain, this structure is represented by small bald peaks formed by quartz–sericite–tourmaline and quartz–tourmaline rocks. Practically, the rock-type boundaries are not obvious macroscopically but are determined only by microscopic study. The fault tectonics are rather distinct. The following types of faults were recognized: (a) curved (semi-circular); (b) faults and diagonal slip faults; (c) accompanying echelon fractures. A curved fault restricts the neck of the plagiogranite–porphyrite from the northeast. This fault is rather steep, but with a weak incline to the periphery of the volcanic tectonic structure. The fault is accompanied by intensive milling of rocks and quartz–gold–sulfide mineralization. Faults and diagonal slip faults most commonly have submeridian and northeast directions; more rarely they have sub-latitudinal direction. One fault (in the

Figure 5. Geographical situation of the Ushtagan gold deposit.

northeast) cuts the volcanic structure and restricts ore-bearing metasomatites from northwest in the region (Fig. 6). Mineralized rocks are characterized by intensive pyritization and vein–veinlet silicification with tourmaline. They are accompanied by intensive jarositization and limonitization from the surface. Gold-bearing bodies are represented by tourmaline–quartz metasomatites with breccia texture developed after igneous-sedimentary rocks. There is less developed mineralization in metasomatites of breccia plagiogranite–porphyrite. The mineralized zone represents a linear stockwork with uneven, complex internal structure. The highest gold and silver grades are limited to quartz–sulfide and quartz–chalcedonic veinlet zones which differ in their intense rust color compared to the rocks on the surface.

Gold-hosting tourmalinized rocks of the Ushtagan area were discovered in 1953. In 1955, during prospecting, two lenses of secondary quartzite were mapped. The database in this study was obtained from eight trenches made by Goldbelt Resources LTD in 1966 on the deposit with a total volume of 3746 m^3. The level of trench deepening is 0.3–0.4 m (Fig. 6). The majority of trenches (due to the complexity of the visual identification of ore intervals) were tested along their entire length by channel sampling and using gold spectral or atomic absorption analyses followed by fire assay test. According to the observations in the trenches, the tuffs, breccias and mineralized bodies with quartz and quartz–tourmaline veins dip sub-vertically (70–80°) in the northeastern part. The initial statistical analysis on 26 gold accumulation samples $\left(\text{meter} \times \frac{\text{gram}}{\text{ton}} \text{ abreviated as } \frac{\text{mg}}{\text{t}}\right)$ taken from the eight trenches revealed that the lack of data potentially results in an erratic global distribution, a characteristic which makes it a suitable target for implementing the proposed algorithm (Fig. 7). The histogram in Figure 7 displays some discontinuity from approximately 40 to 90 mg/t. A moderately high coefficient of variation and negative kurtosis can also be another set of clues for this scarce dataset (Table 1).

To capture the populations in this deposit, taking the proposed algorithm into account, we recall the previously mentioned instructions from Sect. 3.3:

a) *Generating the histogram of the data* As shown in Figure 7 and Table 1, the scarcity of the data is clearly apparent and motivates one to use the proposed algorithm.
b) *Empirical cumulative distribution function* Once the probability distribution function is known, the empirical cdf can then be computed. This function, as explained before, is a step function that for these data increases by 1/26 for each of the 26 samples.
c) *Kernel density function modeling* the tentative theoretical function model for fitting to the empirical cdf has been considered ''normal'' with bandwidth ''0.4'' (Fig. 8a). To check the accuracy of the fitted model, cross-validation was performed by omitting one actual data point and re-estimating it using the kernel fitted model obtained from the remaining data (Fig. 8b). Two values for the underlying datum were scanned, which all are uniformly distributed between 0 and 1. The first value gives the original empirical cdf (abscissa in Fig. 8b), the second value gives the empirical cdf (ordinate, Fig. 8b), and they were measured from the raw dataset and the kernel fitted model, respectively. The cross-validation results illustrate that the scattered points, composed of first and second values, are well distributed along the diagonal and are in agreement with a small error.
d) *Monte Carlo simulation* In this step, 1000 samples within 100 realizations were drawn from the ''normal'' kernel fitted model. To check whether the simulated results were statistically valid, the empirical cdf curve over each realization was computed (green solid line) and then their related average (solid blue line) (Fig. 9) was generated (Fig. 8a). This comparison shows that the simulated results on average converge to the model and can be employed for use in further analysis.
e) *Fractal analysis* The posterior simulated values were obtained by the process of N-S fractal modeling to capture the concealed populations for both sets of fractal analysis obtained from the raw dataset (Fig. 10a) and the average of the results from the realizations (Fig. 10b). Detecting the thresholds through fractal analysis of the raw data is somewhat complex, tricky and needs particular interpretation. Nevertheless, through the GSS-N method, besides the previous

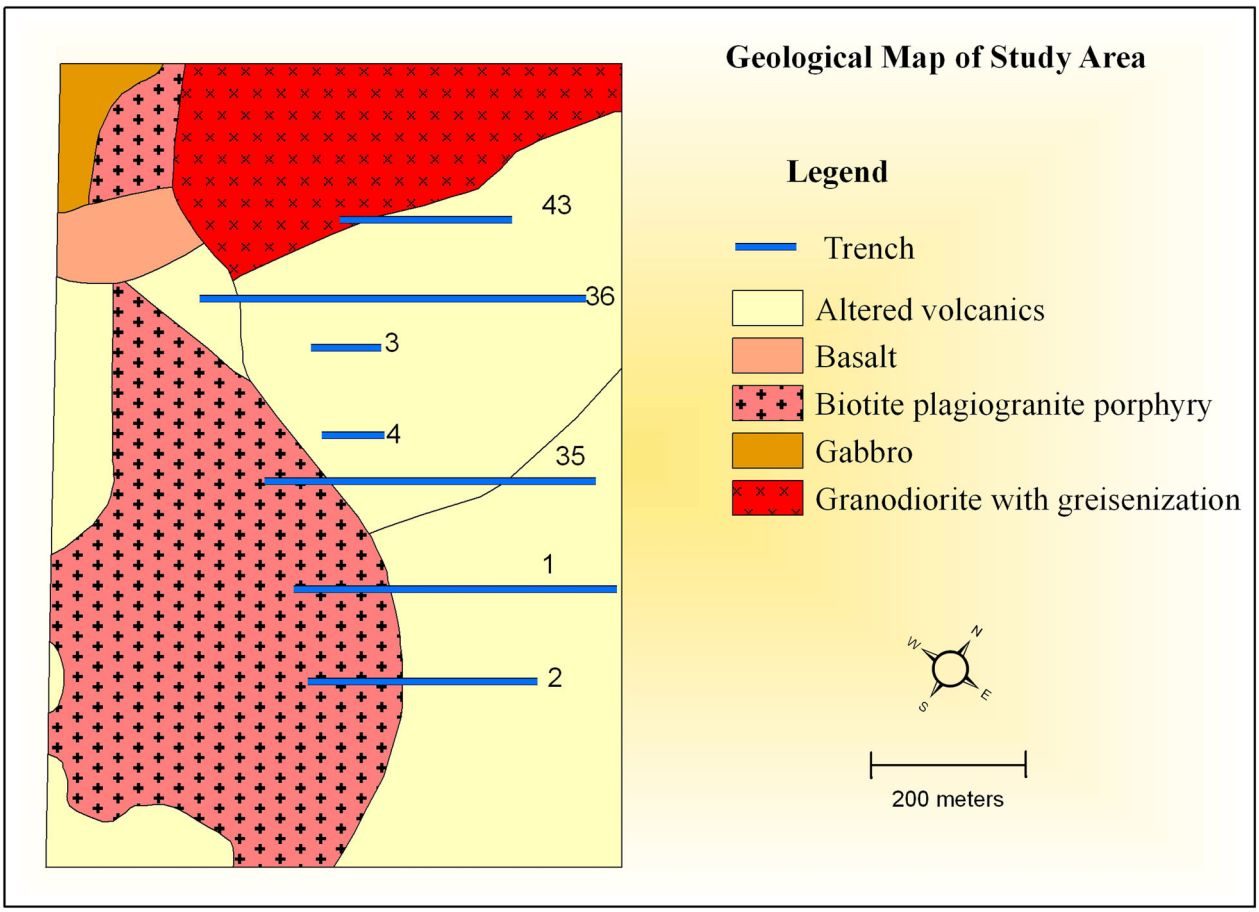


Figure 6. Simplified geological map of the Ushtagan region.

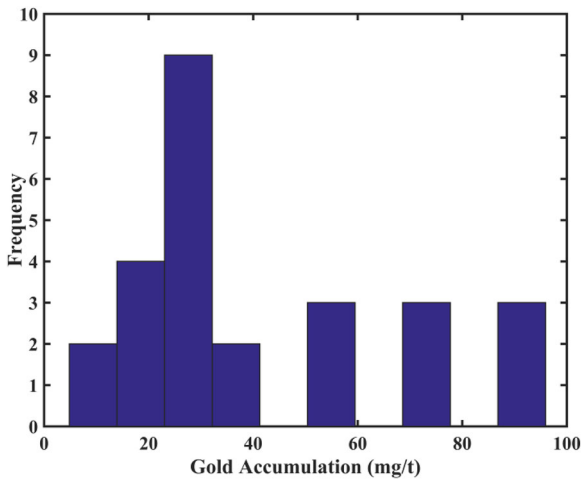


Figure 7. Histogram of gold accumulation obtained from 26 samples.

Table 1. Statistical parameters of gold accumulation data

Parameter	Value
Number of data	26
Mean	41.69077
Median	31.6
Mode	50.4
Standard deviation	26.24942
Sample variance	689.0319
Kurtosis	− 0.34803
Skewness	0.887586
Range	91.1
Minimum	4.84
Maximum	95.94
Coefficient of variation	0.629622

steps and all the statistical checking during the simulation, it was verified that the simulated values on average follow the model. Therefore, one is able to define the thresh-

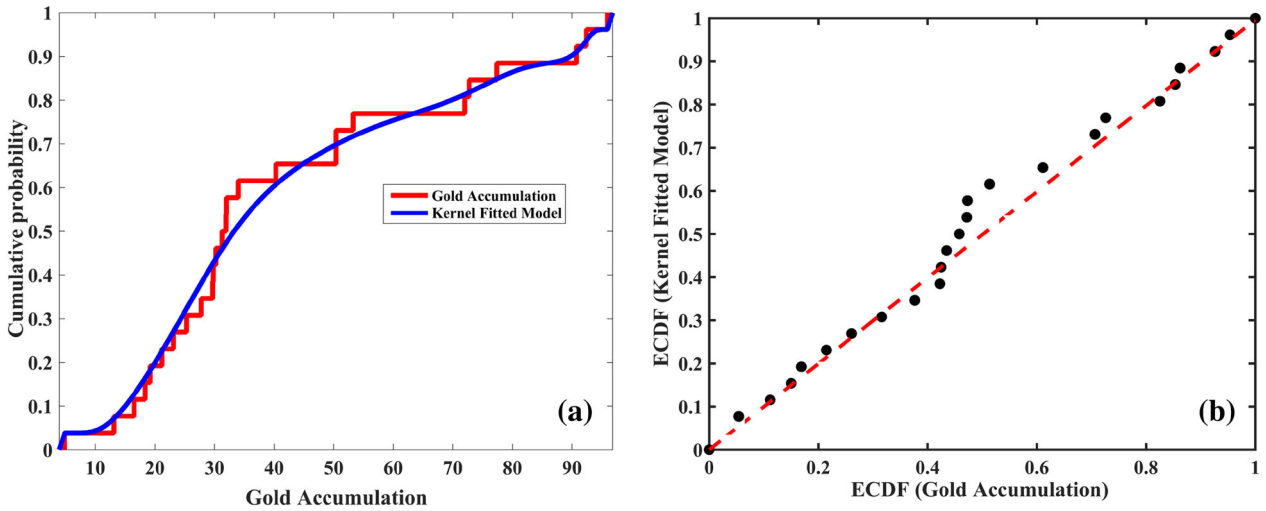


Figure 8. Fitting the model to the empirical cdf of gold accumulation and its evaluation by cross-validation; (**a**) kernel fitted model, (**b**) cross-validation for the fitted model.

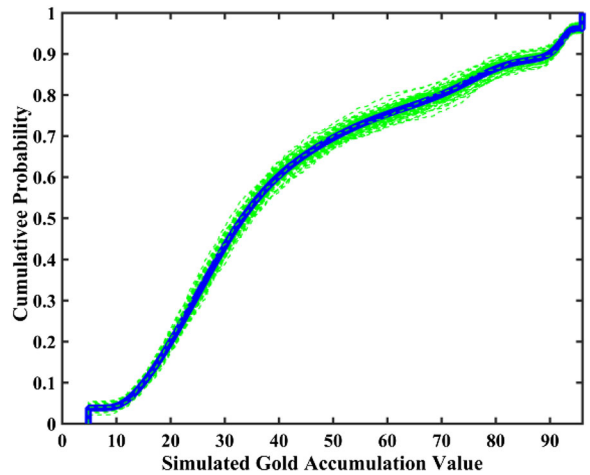


Figure 9. Convergence of the simulated values to the tentative kernel model in terms of empirical cdf.

olds on this graph (Fig. 10b) without any loss of information. For each population, Table 2 shows the ranges for gold accumulation variability and further, the statistical parameters obtained from fitting the linear functions (solid lines) by least squares technique to each population in both cases (original data and GSS-N method) over the fractal points. According to Table 2, the R^2, adjusted R^2 and root-mean-squared error (RMSE) for the GSS-N method indicate valid results. High values of R^2, adjusted R^2 and the closeness of RMSE to zero imply that more points incorporated to construct such an acceptable fit lead to producing a more reliable range of geochemical populations. However, care must be taken to examine the populations derived from the original data (26 samples) according to the statistical validation parameters, as the population ranges seem to be suspicious.

DISCUSSION AND CONCLUSION

We presented an innovative algorithm for geochemical anomaly identification for variables of interest in scarce data. The difference between this technique and the methodology applied by Sadeghi et al. (2015) for anomaly identification is explained in this section. Sadeghi et al. (2015) presented the simulated size–number (SS-N) approach considering a Monte Carlo simulation over the local distribution. For instance, turning band simulation (Emery and Lantuejoul 2006) or sequential Gaussian simulation (Almeida and Journel 1996) has been recommended to use for construction of such a stochastic local distribution that can be used subsequently for differentiating deposit anomalies. However, this approach is robust when sufficient data are available and can be considered subject to the proper spatial continuity analysis (variogram) and further proba-

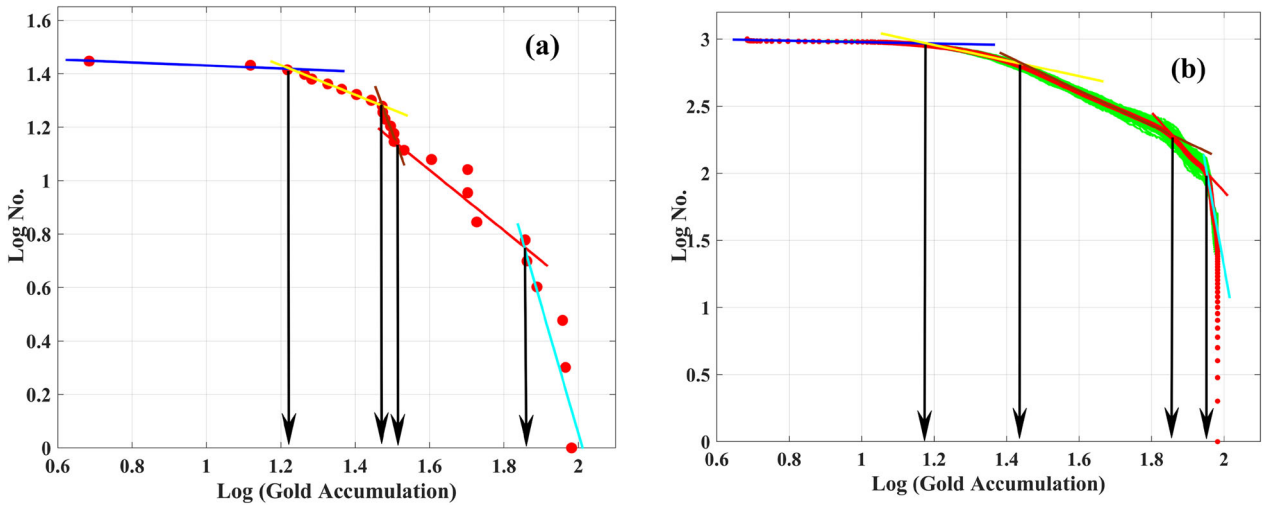


Figure 10. Comparing the population capturing by fractal analysis obtained from the raw data and GSS-N method; (**a**) fractal analysis of original data (26 samples), (**b**) fractal analysis of simulated values.

Table 2. Properties of the geochemical populations captured by the N-S method in original data and captured by the GSS-N method in simulated data

Populations	Original data (N-S)				Simulated values (GSS-N)			
	Range	R^2	Adjusted R^2	RMSE	Range	R^2	Adjusted R^2	RMSE
Background	4.83–16.5	0.87	0.75	0.00	4.83–15.13	0.89	0.89	0.00
Weak anomaly	16.51–30.19	0.90	0.89	0.02	15.14–26.30	0.97	0.97	0.00
Moderate anomaly	30.20–31.62	0.93	0.91	0.01	26.31–66.06	0.99	0.99	0.00
Strong anomaly	31.63–75.85	0.86	0.83	0.05	66.07–89.12	0.99	0.99	0.00
Very strong anomaly	75.86–95.49	0.76	0.69	0.15	89.13–95.49	0.92	0.92	0.03

bilistic modeling (i.e., generating the different realization maps). Reliable maps produced from a large dataset are predictable; the analysis showed that the anomalies acquired from these maps by fractal analysis are more trustworthy than those obtained by applying deterministic solutions (such as kriging). However, when data are scarce, fractal analysis does not give correct local distributions because non-robust variogram models often lead to biased thresholds. The GSS-N method as offshoot of the SS-N method is proposed as an innovative algorithm for dealing with geochemical mapping with scarce data, and it does not require consideration of local distributions. The case studies using synthetic and real dataset confirm the capability of the proposed algorithm.

Fractal/multifractal methods of analysis have been widely used to delineate anomalies in geochemical and ore deposit modeling because of their practical benefit and versatility. These methods are prone to potential biases due to the non-representative distribution that results from scarce data. A global Monte Carlo simulation algorithm is proposed to address this problem of non-representativeness of the distribution due to scarce data. Fitting straight lines on N-S multifractal log-log plots would be problematic and uncertain in case of scarce data; however, using the GSS-N method overcomes the problem of scarce data and thus facilitates straight-line fitting on log-log plots. Therefore, in case of scarce data, thresholds can be easily recognized with higher certainty through the proposed GSS-N method compared to the N-S fractal model. The proposed algorithm employs a simulation approach to calculate possible distribution scenarios of the variable under study based on fitting a kernel density function to the empirical cdf. The theoretical results from using two synthetic datasets and real

dataset from a gold deposit demonstrate that new algorithm removes possible biases, insures reproduction of the actual hidden distribution of scarce data and improves the capture of the underlying populations even when the data are scarce.

ACKNOWLEDGMENT

The first author acknowledges Nazarbayev University for funding this work via Social Policy Grant. The authors also appreciate Prof. Priscilla P. Nelson, head of the Department of Mining Engineering in Colorado School of Mines, for providing the data for real case study for this paper. The authors also appreciate Dr. Masoumeh Khalajmasoumi for her kind supports. The authors are appreciated the constructive comments from two anonymous reviewers, and also we are grateful to Dr. John Carranza for the valuable comments which substantially helped improving the final version of the manuscript.

REFERENCES

Afzal, P., Fadakar Alghalandis, Y., Khakzad, A., Moarefvand, P., & Rashidnejad Omran, N. (2011). Delineation of mineralization zones in porphyry Cu deposits by fractal concentration–volume modeling. *Journal of Geochemical Exploration*, *108*, 220–232.

Afzal, P., Khakzad, A., Moarefvand, P., Rashidnejad Omran, N., Esfandiari, B., & Fadakar Alghalandis, Y. (2010). Geochemical anomaly separation by multifractal modeling in Kahang (Gor Gor) porphyry system. Central Iran. *Journal of Geochemical Exploration*, *104*, 34–46.

Alexandre, B. T. (2009). *Introduction to nonparametric estimation*. New York: Springer.

Almeida, A. S., & Journel, A. G. (1996). Joint simulation of multiple variables with a Markov-type coregionalization model. *Mathematical Geology, 26*(5), 565–588.

Altman, N., & Leger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference, 46,* 195–214.

Borradaile, G. J. (2003). *Statistics of earth science data: Their distribution in time, space and orientation*. Berlin: Springer.

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis* (p. 1997). New York: Oxford University Press Inc.

Chen, G., & Cheng, Q. (2018). Fractal-based wavelet filter for separating geophysical or geochemical anomalies from background. *Mathematical Geosciences, 50*(3), 249–272.

Cheng, Q. (2012). Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas. *Journal of Geochemical Exploration, 122,* 55–70.

Cheng, Q., Agterberg, F. P., & Ballantyne, S. B. (1994). The separation of geochemical anomalies from background by fractal methods. *Journal of Geochemical Exploration, 51,* 109–130.

Cheng, Q., Agterberg, F. P., & Bonham-Carter, G. F. (1996). A spatial analysis method for geochemical anomaly separation. *Journal of Geochemical Exploration, 56*(3), 183–195.

Cheng, Q., Xu, Y. & Grunsky, E. (1999). Integrated spatial and spectral analysis for geochemical anomaly separation. In: Lippard, S.J., Naess, A. & Sinding-Larsen, R. (Eds.), Proceedings of the Conference of the International Association for Mathematical Geology, Vol. 1, (pp. 87–92). Trondheim, Norway.

Cheng, Q., Xu, Y., & Grunsky, E. C. (2000). Integrated spatial and spectrum method for geochemical anomaly separation. *Natural Resources Research, 9,* 43–52.

Chilès, J. P., & Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty. Wiley series in probability and statistics* (2nd ed.). Hoboken: John Wiley & Sons.

Davis, J. C. (1986). *Statistics and data analysis in geology* (2nd ed.). New York: Wiley.

Deutsch, C. V. (1996). Constrained modeling of histograms and cross plots with simulated annealing. *Technometrics, 38*(3), 266–274.

Deutsch, C. V., & Journel, A. G. (1998). *GSLIB: Geostatistical software library and user's guide*. New York: Oxford University Press.

Emery, X., & Lantuejoul, C. (2006). TBSIM: A computer program for conditional simulation of three-dimensional Gaussian random fields via the turning bands method. *Computer and Geosciences, 32*(10), 1615–1628.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications, 14,* 153–158.

Feder, J. (1988). *Fractals* (p. 283). New York: Plenum Press.

Fix, E., & Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report No. 4. Project No. 21–29-004. Randolph Field, T-X: USAF School of Aviation Medicine.

Hawkes, H. E., & Webb, J. S. (1962). *Geochemistry in mineral exploration*. New York: Harper and Row.

He, J., Yao, S., Zhang, Z., & You, G. (2013). Complexity and productivity differentiation models of metallogenic indicator elements in rocks and supergene media around Daijiazhuang Pb–Zn deposit in Dangchang County, Gansu Province. *Natural Resources Research, 22,* 19–36.

Hill, P. D. (1985). Kernel estimation of a distribution function. *Communications in Statistics—Theory and Methods, 14*(3), 605–620.

Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association, 86*(413), 205–224.

Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions*. New York: John Wiley & Sons.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing, 3*(3), 135–146.

Journel, A. G., & Xu, W. (1994). Posterior identification of histograms conditional to local data. *Mathematical Geology, 26,* 323–359.

Leuangthong, O., & Deutsch, C. V. (2003). Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology, 35*(2), 155–173.

Li, C., Xu, Y., & Jiang, X. (1994). The fractal model of mineral deposits. *Geology of Zhejiang*, *10*, 25–32 **(in Chinese with English abstract)**.

Liu, Y., Jiang, Sh., & Liao, Sh. (2014). Efficient approximation of cross-validation for Kernel methods using Bouligand influence function. In Proceedings of the 31st International

Conference on Machine Learning, PMLR 32(1), (pp. 324–332).

Luz, F., Mateus, A., Matos, J. X., & Gonçalves, M. A. (2014). Cu- and Zn-soil anomalies in the NE border of the South Portuguese Zone (Iberian Variscides, Portugal) identified by multifractal and geostatistical analyses. *Natural Resources Research, 23,* 195–215.

Mandelbrot, B. B. (1983). *The fractal geometry of nature. Updated and augmented edition*. San Francisco: W.H. Freeman.

Monecke, T., Monecke, J., Herzig, P. M., Gemmell, J. B., & Monch, W. (2005). Truncated fractal frequency distribution of element abundance data: A dynamic model for the metasomatic enrichment of base and precious metals. *Earth and Planetary Science Letters*, *232*, 363–378.

Pyrcz, M. J., & Deutsch, C. V. (2014). *Geostatistical reservoir modeling*. USA: OUP.

Rossi, M. E., & Deutsch, C. V. (2014). *Mineral resource estimation*. Berlin: Springer.

Sadeghi, B., Madani, N., & Carranza, E. J. M. (2015). Combination of geostatistical simulation and fractal modeling for mineral resource classification. *Journal of Geochemical Exploration, 149,* 59–73.

Sadeghi, B., Moarefvand, P., Afzal, P., Yasrebi, A. B., & Saein, L. D. (2012). Application of fractal models to outline mineralized zones in the Zaghia iron ore deposit, Central Iran. *Journal of Geochemical Exploration, 122,* 9–19.

Samawi, H., Chatterjee, A., Yin, J., & Rochani, H. (2016). On kernel density estimation based on different stratified sampling with optimal allocation. *Communication in Statistics—Theory and Methods, 46,* 10973–10990.

Sanderson, D. J., Roberts, S., & Gumiel, P. (1994). A Fractal relationship between vein thickness and gold grade in drill core from La Codosera, Spain. *Economic Geology*, *89*, 168–173.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley.

Sheikhpour, R., Agha Sarram, M., Zere, M. A., & Sheikhpour, R. (2017). A kernelized non-parametric classifier based on feature ranking in anisotropic Gaussian Kernel. *Neurocomputing, 267,* 545–555.

Shi, J., & Wang, C. (1998). Fractal analysis of gold deposits in China: Implication for giant deposit exploration. *Earth Sciences Journal of China University of Geosciences*, *23*, 616–618 **(in Chinese with English abstract)**.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Tennant, C. B., & White, M. L. (1959). Study of the distribution of some geochemical data. *Economic Geology, 54*(7), 1281–1290.

Tenreiro, C. (2017). A weighted least-squares cross-validation bandwidth selector for kernel density estimation. *Communications in Statistics—Theory and Methods, 46*(7), 3438–3458.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Turcotte, D. L. (1996). *Fractals and Chaos in Geophysics* (2nd ed., pp. 81–99). Cambridge, UK: Cambridge University Press.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall/CRC. ISBN 0-412-55270-1.

Wolfgang, H., Marlene, M., Stefan, S., & Axel, W. (2004). *Nonparametric and semiparametric models*. Berlin: Springer.

Xu, W., & Journel, A. G. (1995). Histogram and scattergram smoothing using convex quadratic programming. *Mathematical Geology, 27,* 83–103.

Zuo, R., Cheng, Q., & Xia, Q. (2009). Application of fractal models to characterization of vertical distribution of geochemical element concentration. *Journal of Geochemical Exploration*, *102*, 37–43.

Zuo, R., & Wang, J. (2016). Fractal/multifractal modeling of geochemical data: A review. *Journal of Geochemical Exploration, 164,* 33–41.