Front. Earth Sci. 2019, 13(1): 229–246
https://doi.org/10.1007/s11707-018-0705-0

RESEARCH ARTICLE

Applied statistical functions and multivariate analysis of geochemical compositional data to evaluate mineralization in Glojeh polymetallic deposit, NW Iran

F DARABI-GOLESTAN (✉), A HEZARKHANI

Department of Mining and Metallurgical Engineering, Amirkabir University of Technology, Tehran, Iran

Abstract Various genesis of epithermal veins as well as host rock cause complication in the modeling process. Thus LINEST and controlling function were applied to improve the accuracy and the quality of the model. The LINEST is a model which is based on multiple linear regression and refers to a branch of applied statistics. This method concerns directly to the application of *t*-test (TINV and TDIST to analyses of variables in the model) and *F*-test (FDIST, *F*-statistic to compare different models) analysis. Backward elimination technique is applied to reduce the number of variables in the model through all the borehole data. After 18 steps, an optimized reduced model (ORM) was constructed and ranked in order of importance as Pb > Ag > P > Hg > Mn > Nb > U > Sr > Sn > As > Cu, with the lowest confidence level (CL) of 92% for Cu. According to the epigenetic vein genesis of Glojeh polymetallic deposit, determination of spatial patterns and elemental associations accompanied by anomaly separation were conducted by *K*-means cluster and robust factor analysis method based on centered log-ratio (clr) transformed data. Therefore, 12 samples (cluster 2) with the maximum distance from centroid, indicates the intensity of vein polymetallic mineralization in the deposit. In addition, an ORM for vein population was extracted for Sb > Al > As > Mg > Pb > Cu > Ag elements with the R^2 up to 0.99. On the other hand, after 23 steps of optimization process at the host rock population, an ORM was conducted by Ag > Te > Hg > Pb > Mg > Al > Sb > As represented in descending order of *t*-values. It revealed that Te and Hg can be considered as pathfinder elements for Au at the host rock. Based on the ORMs at each population Ag, Pb, and As were often associated with Au mineralization. The concentration ratio of $(t_{Sb} \times t_{Al})_{vein}/(t_{Sb} \times t_{Al})_{background}$ as an enrichment index can intensify the mineralization detection. Finally, Glojeh deposit was evaluated to be classified as a vein-style Au (Ag, Pb, As)-polymetallic mineralization.

Keywords LINEST and controlling function, optimized reduced model, log-ratio (clr) transformation, *K*-means cluster, robust factor analysis, pathfinder elements, vein-style Glojeh polymetallic mineralization

1 Introduction

An assemblage of minerals that forms during hydrothermal alteration can reflect the geochemical composition of ore-forming fluids (Zhu et al., 2011; Dora and Randive, 2015). But such assemblages are relatively complex due to the alteration for a polymetallic hydrothermal ore deposit (Stanciu, 1973; Grancea et al., 2002). It can be influenced by numerous factors such as geology, pressure, temperature, hydrology, oxygen, and sulfur fugacity (Oyman et al., 2003; Hezarkhani, 2008; Zhao et al., 2014; Radosavljević et al., 2015; Yang et al., 2015). Also, ore zonation in polymetallic deposits may be complicated due to repeated activity and telescoping (Grancea et al., 2002; Martínez-Abad et al., 2015). Due to the fast variation of some elements in polymetallic veins, it seems essential to conduct different modeling procedures in the vein and host rock population. Based on these considerations, the objective of this study is to characterize mineralization and provide an exploration model in polymetallic epithermal veins and host rock at Glojeh deposit. Regressions (according to the LINEST function) and statistical distribution functions (*F*-test and student's *t*-test) with different confidence levels (CL%) have been used by Garson (2012), Briand and Hill (2013), Borah et al. (2015), and Myers et al. (2016) in various surveys. For computa-

Received March 17, 2017; accepted February 25, 2018

E-mail: pooyan@aut.ac.ir

tion of the LINEST function, Microsoft Office (or MS Office) 2010 and its newer version is enabled (or is able) to calculate more than 43 independent variables and 138 samples (indicated as a 138×43 matrix), simultaneously. Yalta (2008) and Hargreaves and McWilliams (2010) reported a lot of problems with Excel 2007 and older versions of Microsoft Office. To generate polynomial trend line equations, they used a specified (forced) intercept, which was presented as a threshold value in mineral exploration.

The outputs of the LINEST and supplementary function (TDIST, TINV, and FDIST) could be easily simulated in geochemical modeling and mining science (Blythe and Lea, 2008; Zhang et al., 2016). The process was started by creating a full model (FM; which contain all elements) to predict the target element (Cordell and Clayton, 2002). The optimization was conducted according to the step-by-step backward elimination process. Then, the insignificant elements were eliminated by logical interpretation of the presence of elements in the model (Samal et al., 2008). When the maximum p-value (of contributed elements in the related step) was lower than 10% ($P_{\max} < 0.1$), the model was optimized. Accordingly, an optimal reduced model (ORM) has been constructed when all elements contributed with CL% (CL% = 1−TDIST function) higher than 90% (Samal et al., 2008). Majority of the variables at the ORM may be taken into account as pathfinder elements for Au. Several numbers of Sb, As, Bi, Cu, Pb, Se, Ag, Hg, Te, and Zn elements have been determined as common pathfinder elements for Au by Bierlein and McKnight (2005), Nude et al. (2012), and Reith et al. (2005). In addition, possible pathfinder elements for Au were defined by accurate interpretation and comparison of Au models for each vein and host rock population in such deposits (Nude et al., 2012; Hamilton et al., 2017). The geochemical pattern of anomalous vein samples is significantly different from background (Cheng, 2007; Darabi-Golestan et al., 2013). These differences may be reflected in element spatial associations, scale of features, and concentration values (Cheng, 2007; Jovic et al., 2011; Ziaii et al., 2011). Therefore, in order to achieve these goals the K-means cluster and factor analysis have been used.

The K-means algorithm is an efficiency method for multi-elemental sample clustering in large data sets (Mihai and Mocanu, 2015). A set of objects in a K-means cluster are more similar to each other according to some similarity or dissimilarity measures (Huang, 1998; Zarandi and Yazdi, 2008; Mohammadi et al., 2018). In addition, factor analysis based on log-ratio transformed data is applied to reveal underlying patterns in a more vivid manner for compositional data (Filzmoser and Hron, 2008). The centered log-ratio (clr), isometric log-ratio (ilr), and the additive log-ratio (alr) are different transformation tools for interpreting the compositional data (Egozcue et al., 2003; Filzmoser et al., 2009). The aim of the present study is to delineate the spatial patterns and association of total elements and separated clusters of vein and background population using different models.

2 Geological setting and sampling

The Glojeh district is located at approximately 30 km north of Zanjan Province, in the northern part of 1:100,000 geological map of Zanjan, in the Tarom-Hashtjin metallogenic zone in NW of Iran (Darabi-Golestan and Hezarkhani, 2017b). The Cenozoic volcanic rocks are extensively distributed throughout the Tarom-Hashtjin metallogenic zone which comprising trachy-basalt, trachy-andesite, andesite, andesitic-basalt, rhyodacite, and rhyolite lava flows and tuff (Mehrabi et al., 2014; Fig. 1). The Glojeh mineralized veins are hosted by the rhyodacite, lithic tuff units and other Cenozoic volcanic rocks (Darabi-Golestan and Hezarkhani, 2016). The deposit consists of a set of sub-vertical polymetallic Pb-Zn-Cu-Ag-Au sulphide-bearing veins, and considerable amounts of Bi and Cd. The deposition mechanism of gold bearing veins is generally high-grade, and fault hosted. In addition, ore minerals textures are disseminated, replacement, banded (open space filling), and stockwork.

In the first exploration phase, seven trenches (named here TR0 to TR6) were excavated and a total of 387 composite samples were collected from totally 878 meters long of all trenches. Afterward, during the second phase, several trenches (TR1-2 to TR5-6) were opened between the pervious trenches (Fig. 1). The number of 460 composite samples were prepared to determine changes of concentrations at a total of 482.8 m long in these trenches. In addition, a total number of 11 boreholes were designed and drilled with dip of 45° and azimuth direction of 180° (same as the trenches) in the area. All boreholes were drilled about 1503.24 m, and 806 composite samples were collected and analyzed. Only the number of 152 core samples (consisting of 14 duplicated and replicate samples) from BH2N1 borehole were analyzed for 43 elements using ICP-MS at Earth Sciences Development Company Lab of Iran. All the other samples were analyzed for Au, Ag, Cu, Pb, and Zn, which have been recognized as mineralized elements at the Glojeh deposit. The fire assay (for Au analysis) and atomic absorption spectrophotometry (for associated Ag, Cu, Pb, and Zn elements) techniques have been conducted to analyze the samples at Zarkavan and Earth Sciences Development Company Lab of Iran, respectively. The total number of 73 duplicated and replicate samples were taken in order to assess the accuracy and precision of the analyses.

In this study, ICP-MS analyses from BH2N1 were carried out to create a comprehensive model in the area for 43 elements. The total depth is 140 m, where about 127.69 m of this borehole was discovered with rhyolitic-tuff rocks and 12.31 m hosted in the brecciated veins. A and B major veins were intersected by this borehole at

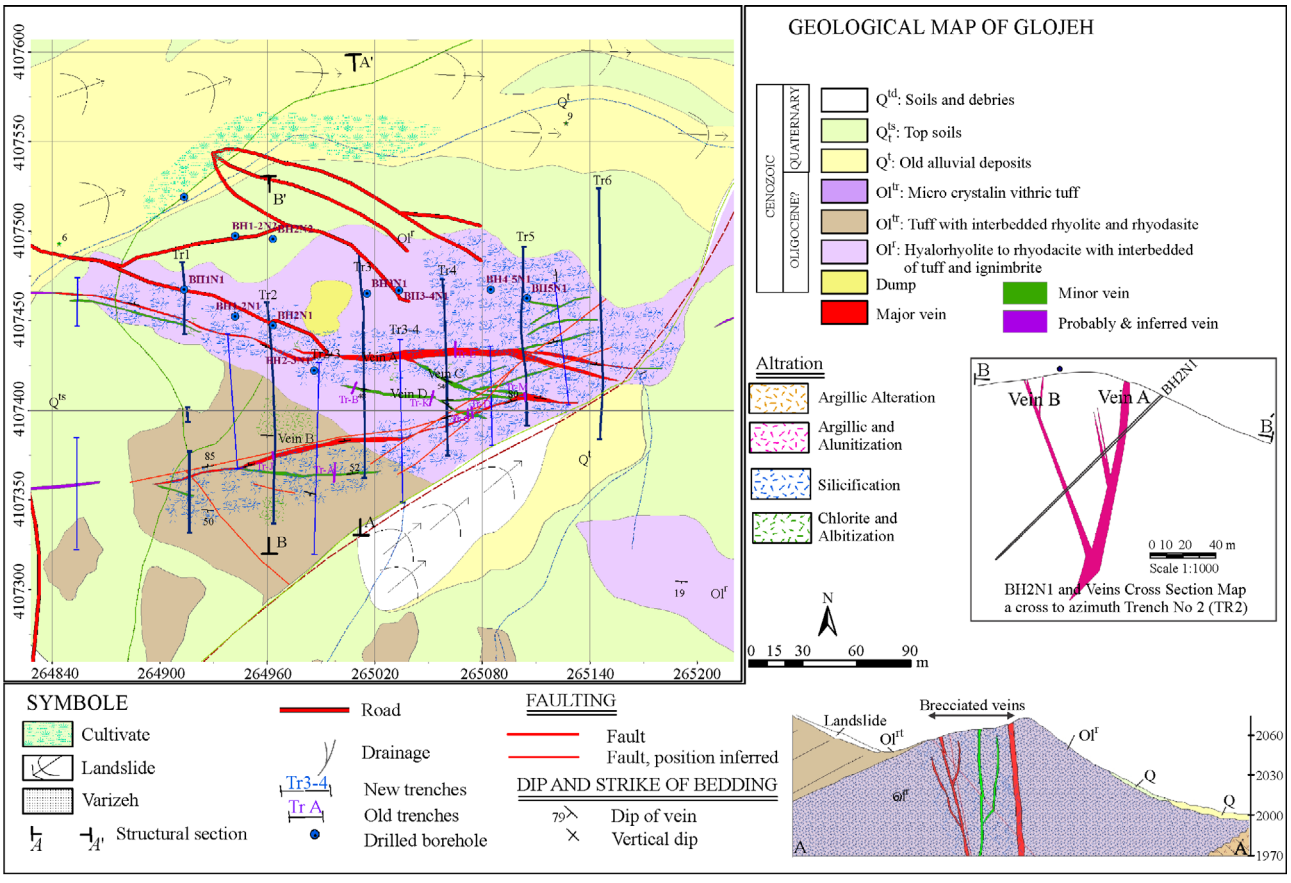

Fig. 1 Geological map of Glojeh polymetallic veins deposit.

26.2–33.9 m and 79.68–81.51 m, respectively. The general schema of exploratory trenches and a cross section map of the wall and the bottom of TR4-5 are depicted in Figs. 1 and 2, respectively. Total length of trench TR4-5 is 75 m, and the deep is 1.8 m, and 70 cm wide. The average of Au analyzed in the dataset equals to 2836 ppb (parts per billion). Besides, several silicified veinlets including chalcopyrite mineralization and strongly hematitic veinlets are observed in this trench. Tr4-5 is covered with silicic tuff, quartz brecciated zone (1.4 m includes 7450 ppb Au) and vein (9 m contains 13,750 ppb Au).

3 Methodology

In mining science and geological knowledge, the variables could be considered as dependent or/and independent, related to the object of the investigation and data processing (Savazzi and Reyment, 1999; Bise, 2013). With the aim of modeling, it is essential that the relationship between these variables are well defined (Briand and Hill, 2013; Darabi-Golestan and Hezarkhani, 2016). Implementation and identification of an optimized reduced model (ORM) between elements is conducted based on the LINEST function accompanied by TDIST and FDIST functions in staged backward modeling for borehole number BH2N1. An epigenetic vein polymetallic deposit must show different specifications between elements in veins and background populations. Therefore, two distinct considerations given to the vein and background population may lead to better acknowledgments of elemental association and variation. In order to reach this aim, *K*-means cluster analysis and factor analysis are applied to the clr transformed data. The combination of these methods describes an advanced explanation to separate polymetallic veins from background and, accordingly, spatial association between elements and samples.

3.1 LINEST function

The easiest equation for the line by one independent variable is $y = \beta_1 x + \beta_0$, but for multiple variables the equation changes to $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \beta_0$. The LINEST function calculates the statistics for a line with best fit to the data by using the least squares method. The LINEST function syntax runs as LINEST (known_y's, [known_x's], [const.], [stats]) and leads to calculation of threshold or intercept value (constant value or β_0), different coefficients (β_1,β_2,...β_{k-1},β_k), standard errors (SE_0,SE_1,

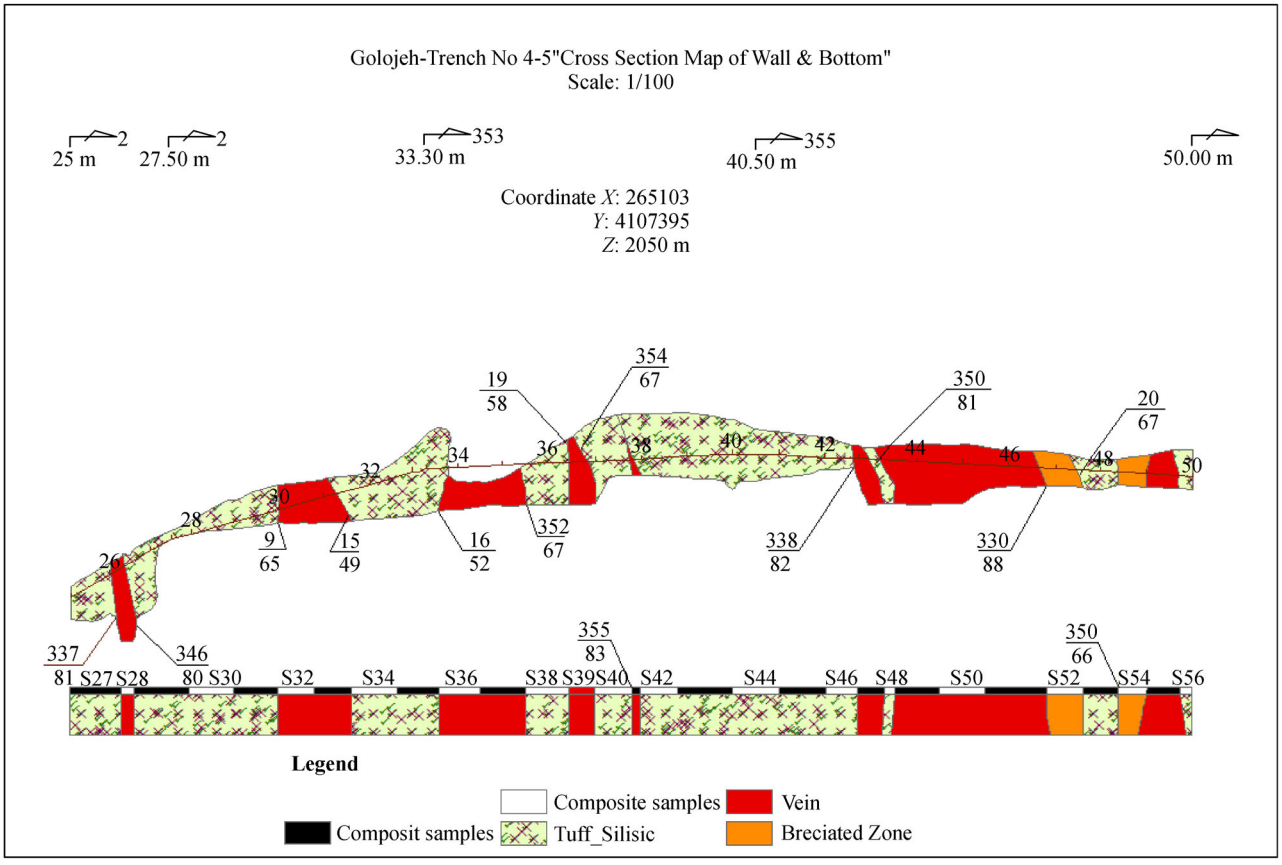

Fig. 2 Cross section map of wall and bottom of TR4-5.

...,SE_k), standard error for actual y-value (SE_y), squared correlation coefficient between estimated y-values and real y-values (R^2; ranges from 0 to 1 where values closer to 1 indicate better fit), F-statistic or the F-observed value (F-value), degrees of freedom (df2), sum of squares due to regression (SSR) and residual sum of squares (SSE) which are presented in Table 1 (Kutner et al., 2005; Larose, 2006; Garson, 2012; Briand and Hill, 2013; Borah et al., 2015).

The β_0 value is considered as a constant option. If constant option is TRUE, β_0 is calculated normally in accordance with the Au variation and if constant option is FALSE, then β_0 equals to 0. The authors agree with Hargreaves and McWilliams (2010) and recommend that by considering the threshold value, the LINEST function modeling is more useful, especially in geochemical exploration. Therefore, in mineral and geochemical surveys, the constant option value must be considered. The average of the estimated parameter should be close to the true parameter values, because the least squares estimator is unbiased (Myers et al., 2016). If stats is TRUE, then the LINEST can create additional regression statistics (R^2, SE_{β_0}, F, df2, SSR, SSE) and if stats option is FALSE or omitted, then LINEST returns only the β_k - coefficients and the β_0 constant (Briand and Hill, 2013; Borah et al., 2015). The F-value is determined to evaluate modeling and observe relationship between the dependent and independent variables (DeCoursey, 2003; Borah et al., 2015). A confidence level is determined by F-value

Table 1 The matrix of LINEST function results (regression statistics, after Borah et al., 2015)

Intercept	X_1	X_2	…	X_{k-1}	X_k
Intercept β_0	Slope β_1	Slope β_2	…	Slope β_{k-1}	Slope β_k
Standard error for $\beta_0 = SE_0$	Standard error for $\beta_1 = SE_1$	Standard error for $\beta_2 = SE_2$	…	Standard error for $\beta_{k-1} = SE_{k-1}$	Standard error for $\beta_k = SE_k$
R^2	SE_y				
F-statistic = F	Degrees of freedom = df2				
Regression SS = SSR[a)]	Residual SS = SSE[b)]				

a) Sum of Squares Regression; b) Sum of Squared Errors of prediction.

(Breidenbach et al., 2016), and the df factor also helps to find critical *F*-values (Briand and Hill, 2013).

3.2 TDIST and TINV functions

LINEST function was applied accompanied by TINV [TINV(*p*-value, df) = *t*-value] and TDIST [TDIST (*t*-value, df, 2 tails) = *p*-value] controller functions. Based on these definitions, TINV function calculates the inverse of the two-tailed student's *t*-distribution and TDIST calculates the student's *t*-distribution. Therefore, *p*-values were calculated with the TDIST function using calculated *t*-values and the df for probability function. If x (as a variable) has a distribution with mean of μ and standard deviation of σ, and n (number of samples) is sufficiently large, then *t*-value shows approximately the standard normal distribution (DeCoursey, 2003; Remenyi et al., 2011). The *t*-distribution is known by Eq. (1) or it can be directly calculated from the first and second rows of Table 1:

$$t = \frac{\overline{x} - \mu}{(\sigma/\sqrt{n})} = \frac{\mu_k(\beta_k)}{SE_k}, \quad (1)$$

where μ is the population mean, $\overline{x}$ is the sample mean, and σ is the estimator for population standard deviation with n independent samples (DeCoursey, 2003). The distribution of the t statistic is called *t*-distribution or the student's *t*-distribution (Remenyi et al., 2011). The *t*-distribution with k-degrees of freedom is symmetric and bell-shaped (with mean 0 and variance 1), like the normal distribution but has heavier tails, meaning that it is more prone to producing values that fall far from its mean (DeCoursey, 2003; Larose, 2006; Røislien and Omre, 2006). The *t*-distribution is useful for understanding the statistical behavior of variables, in which variation in the denominator is amplified and may lead to large outlying values when the denominator of the ratio falls close to zero (Székely and Rizzo, 2013). As df2 keeps growing, the *t*-distribution approaches to the standard normal distribution, and in fact the approximation is quite close to $k \geqslant 30$. The value of *t*-distribution for two-tails probability is shown in Table 2. Probability of A and P area (CL% from 80% to 99.9%) with specification df (from 1 to ∞) is defined as reference *t*-value (Larose, 2006). For example, the *t*-value at CL% of 95% at an infinite df is equal 1.96. For a constant df, when A increases (or P decreases) then *t*-values increased. Accordingly, the regression coefficients tend to be significant. Therefore, *t*-value is used to test the significance of regression coefficients.

3.3 *F*-distribution function

The *F*-distribution is right skewed and described as FDIST (x, df1, df2). It is equal to the probability of the *F*-distribution for value $\geqslant x$ (x = the area of the right tail

Table 2 The *t*-distribution values (two-tailed)

df	A / P	0.80 0.20	0.90 0.10	0.95 0.05	0.98 0.02	0.99 0.01	0.995 0.005	0.998 0.002	0.999 0.001
1		3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2		1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3		1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4		1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5		1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6		1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7		1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8		1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9		1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10		1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
20		1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
30		1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40		1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50		1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60		1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80		1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100		1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120		1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞		1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

beyond) with df1 and df2 degree of freedom (DeCoursey, 2003; Larose, 2003; Hill and Lewicki, 2006; Zaiontz, 2014; Fig. 3). If X has a F-distribution then Pr:($X \leqslant x$) = P. The critical values for the F-distribution are tabulated by different significance level or P-values (Jiang et al., 2015). P-value or α is defined by inverse cumulative distribution function of Fisher's F-distribution (Borah et al., 2015). The value of FDIST function was used to test the significance of linearity performance. If it is lower than 0.05, then a relationship between Au and other element is approved and as long as this differences is lower than 0.05 (FDIST $\ll$ 0.05), the result shows that it has been achieved by no chance (DeCoursey, 2003).

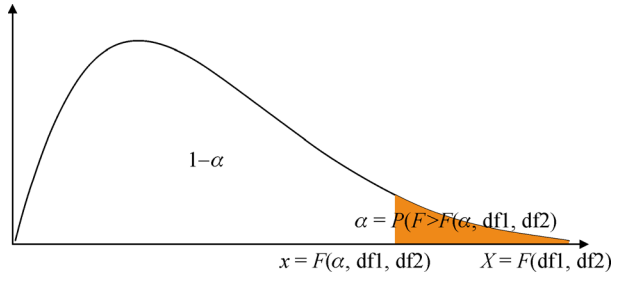

Fig. 3 Probability plot of F-distribution function.

3.4 K-means cluster analysis

K-means cluster is an informative multi-component statistical technique that has been used in many geochemical studies to classify samples (observations) into groups (Liu et al., 2016; Namhata et al., 2017). In the geochemical data analysis, which there are two or more numeric columns of elements that represents different sample concentration measurements, this method can be used successfully (Vriend et al., 1988; Templ et al., 2008). This method is done based on MacQueen's algorithm (MacQueen, 1967), and it applied according to the nonhierarchical clustering of observations (Huang, 1998; Guha and Mishra, 2016; Soheily-Khah et al., 2016).

3.5 Factor analysis

Factor analysis has been applied to classify variables into factors, based on their similarities to correlation analysis or covariance matrix (Karamanis et al., 2009; Suresh et al., 2012; Ramasamy et al., 2013). It was used as a feature extractor method to identify unobserved latent variables among the data (Fox, 1983; Dehak et al., 2011; Darabi-Golestan et al., 2017). The reduced factors indicate the associations between variables (Abdi et al., 2013; Golestan et al., 2014; Tokatli et al., 2014) based on the largest amount of variability in the data (Fávaro et al., 2007; Darabi-Golestan and Hezarkhani, 2017a). It can be applied to extract the factor loading coefficients and score factors by different methods such as principal components, maximum likelihood, and principal axis factoring (Almasi et al., 2014; Yousefi et al., 2014). Prior to applying robust factor analysis, the geochemical compositional data were opened using CoDaPack software v 2.01. Then, they were transformed by centered (clr) log-ratio. A log-ratio transformation could overcome the elemental closure effect for compositional data analyses where it is clearly visible for the analysis based on the log-transformed data (Pawlowsky-Glahn and Egozcue, 2006; Buccianti and Grunsky, 2014; Wang et al., 2014).

4 Results

4.1 Numerical borehole model

Numerical LINEST modeling of the geochemical data in Glojeh deposit was conducted based on a total of 138 samples of BH2N1 borehole. Each sample was analyzed for 43 elements where Au was considered as a dependent variable and other 42 elements were the independent variables (k = 42) for Au modeling. With the implementation of the LINEST function, all 42 elements and the threshold value (β_0) were contributed in the FM, at the first step. At this step, df1 (depends on variables which included in the FM) and df2 (depends on samples) can be calculated with the following formula (Table 3):

$$\begin{aligned} \mathrm{df2} &= \mathrm{samples} - \mathrm{variables} = 138 \\ &- (42\ \mathrm{element} + 1\ \mathrm{threshold\ limit\ value}) = 95, \\ \mathrm{df1} &= k - \mathrm{df2} - 1 = 138 - 95 - 1 = 42. \end{aligned} \quad (1)$$

The majority of variables which contribute to modeling may have the least effect in modeling and must be removed from FM by using different controlling functions. At each step, the F-statistic, t-value (student's t-value or TINV function), and FDIST function for all variables were determined. In order to model optimization, the element (or elements) with the least t-value (insignificant element which contribute with the least CL%) is removed from the modeling process (Larose, 2006; Borah et al., 2015). Finally, after several steps of elimination the model could be considered as ORM, that all of the present variables contribute with at least 95% CL (or p-value [α] equal or lower than 0.05). According to the sensitivity of the simulated model and low interaction effects of elements in the early steps of the optimization process, the first two elements, and after 15 steps only the first element with the lowest t-value were eliminated. The Se and Cr show the lowest t-value (equal to 0.02 and 0.11, respectively) and the highest error among all variables in the first step. In order to optimize the model Se and Cr were eliminated. Accordingly, at the second, third, and eventually at the 17th step, simultaneously Ce-Bi, La-Zn, and Tl were

Table 3 Model optimization in all data of Glojeh deposit by average CL of 90%

Step	Cor.	df1	df2	FDIST	TINV	TDIST of 90%	Elements by minimum *t*-value		Elements by minimum *t*-value	
							Ele. 1	Value	Ele. 2	Value
1	0.77	42	95	1.49E−16	1.661	0.99	Se	0.02	Cr	0.11
2	0.77	40	97	1.82E−17	1.661	0.90	Ce	0.12	Bi	0.16
3	0.77	38	99	2.08E−18	1.660	0.88	La	0.15	Zn	0.16
4	0.77	36	101	2.23E−19	1.660	0.88	Zr	0.15	Rb	0.17
5	0.77	34	103	2.21E−20	1.660	0.79	Threshold limit	0.26	Te	0.30
6	0.77	33	105	3.92E−21	1.659	0.72	Ni	0.35	Sc	0.46
7	0.77	31	107	3.95E−22	1.659	0.66	Y	0.45	V	0.51
8	0.77	29	109	3.72E−23	1.659	0.65	Mo	0.45	Li	0.51
9	0.77	27	111	3.45E−24	1.659	0.42	Be	0.80	Al	0.82
10	0.76	25	113	4.9E−25	1.658	0.58	Th	0.56	W	0.98
11	0.76	23	115	5.91E−26	1.658	0.51	Ti	0.65	Fe	1.04
12	0.76	21	117	6.16E−27	1.658	0.29	Sb	1.05	S	1.18
13	0.75	19	119	1.02E−27	1.658	0.56	Ca	0.59	Ba	0.88
14	0.75	17	121	8.06E−29	1.658	0.47	K	0.73	Cd	0.99
15	0.75	15	123	6.94E−30	1.657	0.31	Mg	1.01	Co	1.23
16	0.74	13	125	7.04E−31	1.657	0.23	Na	1.21		
17	0.74	12	126	2.54E−31	1.657	0.10	Tl	1.64		
18	0.74	11	127	1.58E−31	1.657	0.08	Cu	1.75		
19	0.73	10	128	1.14E−31	1.657	0.06	As	1.87		

eliminated, respectively (Table 3). The threshold value (β_0) has the lowest *t*-value (equal to 1.66) on the 5th step, while by removing β_0, the df1 doesn't change (Table 3; changes from 34 to 33 at this stage is due to the removing Te). The process is similar for all the 19 steps in the research. As illustrated in Tables 3 and 4 for the optimal model, $\beta_0 = 0$ so df1 = 138−127 = 11 and df2 = 127. In addition, at the 18th step of optimization, the R^2, Se(Au), *F*, df2, SSR, and SSE criteria are calculated equal to 0.736, 0.529, 32.124, 127, 98.956, and 35.565, respectively. Accordingly, the optimized model was determined as:

$$\begin{aligned}&\text{Optimized model after 18 steps} \rightarrow \text{Au}\\ &= 0.326 \times \text{U} - 0.158 \times \text{Sr} + 0.27 \times \text{Sn} + 0.379 \times \text{Pb}\\ &\quad - 0.285 \times \text{P} + 0.196 \times \text{Nb} - 0.198 \times \text{Mn} - 0.235\\ &\quad \times \text{Hg} - 0.1211 \times \text{Cu} + 0.16 \times \text{As} + 0.293 \times \text{Ag},\\ &\text{df1} = 138(\text{all the samples}) - \text{df2} = 138 - 127\\ &\quad = 11 \text{ relates to the elements where exist in model,}\\ &\text{FDIST}(F,\ \text{df1},\ \text{df2}) = 1.58\text{E}{-}31 <<< 0.05,\\ &\text{TINV}(p\text{-value index, df2}) = \ \text{TINV}(0.05,\ 127) = 1.978.\end{aligned}$$

The syntax of the TDIST function is the inverse of the TINV function for a two-tailed case (Angelo et al., 2007; Briand and Hill, 2013). It is introduced as TINV (*p*-value, df2) = *t*-value and TDIST (*t*-value, df2, 2 tails) = *p*-value; or if it considers for one-tailed case, then TDIST (*t*-value, df2, 2) = 2×TDIST (*t*-value, df2, 1). The *t*-value = $|\frac{\beta_k}{SE_k}|$ for two-tail probability is calculated in all steps of the modeling process. Also note that *x* must be non-negative, but since the *t*-distribution is symmetric for *x* = 0 (Table 2), we can use ABS (*t*) as seen for Sr, P, Mn, Hg, and Cu (Table 4). At the 17th step, *t*-value (Tl) = 1.64 = $A_{0.896}$ < $A_{0.90}$ and accordingly Tl may be eliminated for optimization. At the 18th step, the minimum *t*-value (*t*-value(Cu) = 1.75 = $A_{0.917} \approx A_{0.92}$ or TDIST_{Cu} (1.75, 127, 2) = *p*-value = 0.083) for the remaining elements (U, Sr, Sn, Pb, P, Nb, Mn, Hg, Cu, As, and Ag) is covered by the relevant significance test. Therefore, optimization is completed after 18 steps (Table 3). From the beginning of the optimization process, at each step the TINV function related to 90% CL and df was calculated and presented in Table 3. It indicates the lowest acceptable *t*-values for elements in the optimized model as:

$$\begin{aligned}&\text{Cu } t\text{-value } (1.75)\\ &> \text{TINV function related to 90\% CL } (1.657),\end{aligned}$$

where TINV function was varied from 1.661 to 1.657, df2 changes from 95 to 127 within 18 steps, simultaneously (Table 3). According to the reference values of Table 2, a

90% CL reflects a TINV function value close to 1.658 for a df2 = 120. Eventually, all elements in the optimal model have the least 90% CL or $\alpha < 0.10$, revealing *t*-value $\geqslant 1.657$. The *t*-value for a variable in the model is calculated in association with the contribution of other variables (related to df2). For example, U is another character in the optimized model that revealed TINV (0.008, 127)= 2.7 or TDIST (2.7, 127, 2) = *p*-value = 0.008 (Table 4).

The FDIST function and R^2 are the other two criteria for optimization. The FDIST (*F*-value, df1, df2) function must have a value much lower than α. When the *F*-value increases sequentially from 4.653 to 21.720, the FDIST function decreases from 1.49E–16 to 1.58E–31 during 18 steps (Table 3). The calculated value of FDIST (32.124, 11, 127) is equals to 1.58E–31 ($<< 0.05$) at the optimized model. It implies a strong relationship between variables in order to approve the model constructed for Au after 18 steps (Tables 3 and 4). The reduction trend in the FDIST function implies a significant improvement in the accuracy of the model. The *F*-value (Eq. (2)) is determined by the LINEST function according to the mean squares of regression (MSR; Eq. (3)) and mean squares of error (MSE; Eq. (4)). MSR and MSE are obtained by SSR (Eq. (5)) and SSE (Eq. (6)) (Larose, 2006; Darabi-Golestan and Hezarkhani, 2016). The SSR value indicates the amount of variations in data are explained by the regression. However, a smaller value of SSE leads to a better fit on regression model (Larose, 2006). The total sum of square (SST) value is calculated by using the expression SST = SSR + SSE. Accordingly, the correlation coefficient is also determined (R^2= 73.6%) (Eq. (7) and Fig. 4). These statistics are computed by the following formula:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{8.995}{0.28} = 32.124, \tag{2}$$

$$\text{MSR} = \frac{\text{SSR}}{\text{df}} = \frac{98.95}{11} = 8.995, \tag{3}$$

$$\text{MSE} = \frac{\text{SSE}}{(k - \text{df} - 1)} = \frac{35.56}{127} = 0.28, \tag{4}$$

explained variation:

$$\text{SSR} = \sum_{i=1}^{k} (\hat{y}_i - \overline{y})^2 = \sum_{i=1}^{126} (\hat{y}_i - (0))^2 = 98.95, \tag{5}$$

unexplained variation:

$$\text{SSE} = \sum_{i=1}^{126} (y_i - \hat{y}_i)^2 = 35.56, \tag{6}$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSR}}{\text{SSR} + \text{SSE}} = \frac{98.95}{98.95 + 35.56} \approx 73.6\%. \tag{7}$$

Figure 5 illustrates the differences between the *t*-test (TDIST) and the *F*-test (FDIST) in optimization process in Glojeh. It reveals, selecting different elements and models are so complicated through backward process, while it can be facilitated by the *t*-test and the *F*-test statistics. At each step, a separate *t*-test was performed for all the predictors that participate in modeling. On the other hand, the *F*-test considers the linear relationship between Au and a set of predictors (e.g., Pb, Ag, P, Hg, Mn, Nb, U, Sr, Sn, As, and Cu at the optimized model) which participate in modeling. Hence, after 18 steps and by application of TINV, TDIST, FDIST, and other controlling functions, the optimal model is determined (Fig. 5). During the elimination process of elements, df1 have decreased from 42 elements in the FM to 11 elements (Pb, Ag, P, Hg, Mn, Nb, U, Sr, Sn, As, and Cu) in the optimal model (18th step). β_k coefficients are made by the sample covariance. These coefficients can be calculated via the COVAR function, while the LINEST function creates these coefficients directly. The ORM indicates the correlation of 73.6% after 18th step (Tables 3 and 4, Fig. 4). The maximum and minimum *t*-values in the ORM are related to Pb and Cu equal to 5.198 and 1.75, respectively (Table 4). When *t*-values are sorted in descending order, it is observed that:

$$t_{\text{Pb}} > t_{\text{Ag}} > t_{\text{P}} > t_{\text{Hg}} > t_{\text{Mn}} > t_{\text{Nb}} > t_{\text{U}} > t_{\text{Sr}} > t_{\text{Sn}} > t_{\text{As}} > t_{\text{Cu}}.$$

4.2 *K*-means cluster analysis

K-means cluster analysis was conducted for 43 elements form 138 samples, which obtained from ICP-MS analysis. For a better understanding of the models, the geological data from dataset were classified into three major groups. *K*-means non-hierarchical clustering indicates that 82, 12, and 44 samples observed at first, second and third clusters, respectively (Table 5). The aim is that the within cluster sum of squares is minimized. The average and maximum distance from centroid (0, 0) are the highest values for the second cluster, where the maximum distance at first and

Table 4 The matrix of LINEST function results of BH2N1 at 18th step

Element (*x*)	U	Sr	Sn	Pb	P	Nb	Mn	Hg	Cu	As	Ag
$\beta(x)$	0.326	−0.158	0.270	0.379	−0.285	0.196	−0.198	−0.235	−0.121	0.160	0.293
Se(*x*)	0.121	0.062	0.116	0.073	0.084	0.063	0.063	0.073	0.069	0.074	0.085
t-value = $\beta(x)$/Se(*x*)	2.700	2.551	2.322	Max *t*-value = 5.198	3.375	3.099	3.146	3.233	Min *t*-value = 1.750	2.172	3.470
TDIST = *p*-value	0.008	0.012	0.022	0.000	0.001	0.002	0.002	0.002	0.083	0.032	0.001

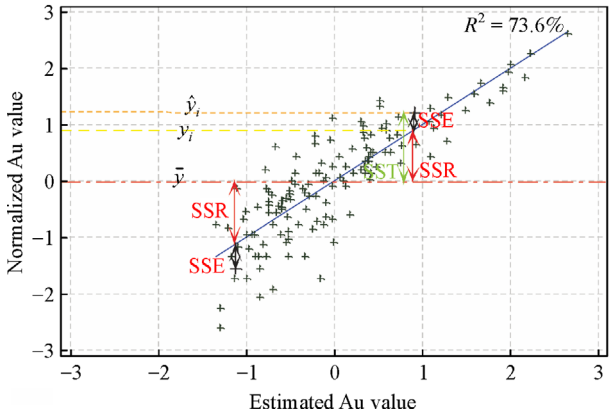

Fig. 4 Total (SST), explained (SSR), and residual (SSE) variation of BH2N1 from optimized regression line and actual Au values.

third cluster are related to the sample ID's of 52 and 100, respectively (Table 5 and Fig. 6(b)). Cluster 1 mainly corresponds to the rhyodacite (and interbedded tuff and rhyolite) volcanic section; cluster 2 is dominantly related to polymetallic vein, veinlet, and brecciated zones of Eocene and Oligocene ages (Mehrabi et al., 2016); and cluster 3 is predominantly associated with silicified lithic tuff host rocks that are genetically linked to sulfide-bearing quartz stockworks (Fig. 1). The third cluster shows sensitive responses to increases in Cu and Mo concentrations.

4.3 Robust factor analysis

A factor analysis method based on centered log-ratio (clr) transformed data was applied to geochemical data in the Glojeh deposit. Factor loadings and variance estimations were performed according to the maximum likelihood factor analysis of the correlation matrix, whereas the data were clustered in three groups using *K*-means clustering method. The first four factors explained 62.3% of the total variance, while the first factor accounted for 42.5% of the variation in the data. The second, third, and fourth factors

Table 5 *K*-means cluster analysis results

Item	Cluster 1	Cluster 2	Cluster 3
Number of observations	82	12	44
Within cluster sum of squares	1270.338	493.912	516.369
Average distance from centroid	3.856	6.253	3.283
Maximum distance from centroid	5.677	9.281	6.403

explained 10.6%, 4.9%, and 4.3% of the total variation, respectively. The factor loadings that indicate the characteristic of variables at the first and second factor are shown as a vector in Fig. 6(a). One of the advantages of applying the clr transformation is to overcome the closure effect of vectors in the classic factor analysis. Therefore, the spatial distribution of elements and samples are improved and eigenvectors (loadings) are spread in a full circle. A biplot of factor loadings revealed that Au, Ag, Pb, As, Zn, Te, Be, Cu, S, Li, W, and Sb are more correlated than other elements (Fig. 6(a)), in order to the direction of anomalous samples (Fig. 6(b)). These elements are concentrated at the polymetallic vein samples (cluster 2) consisting of 12 samples (see Fig. 6(b) for sample ID's 35, 33, 31, 28, 32, 29, 30, 47, 80, 36, 34, and 81). These results have been achieved by Darabi-Golestan and Hezarkhani (2017b) using *R*- and *Q*-mode analysis and correspondence analysis. These 12 samples have the maximum distance from centroid (Fig. 6(b)). This high eccentricity indicates the intensity of vein polymetallic mineralization in the Glojeh deposit. Cu sulfide mineralization (as pyrite chalcopyrite and galena) in this deposit leads to high association and overlapping of Cu and S eigenvectors (Fig. 6(a)).

4.4 Gold evaluation within host rock

A total number of 126 samples from the first and third clusters, obtained from *K*-means cluster analysis, were used as the background population. The LINEST modeling for Au starts with 42 independent variables (df1). The FM

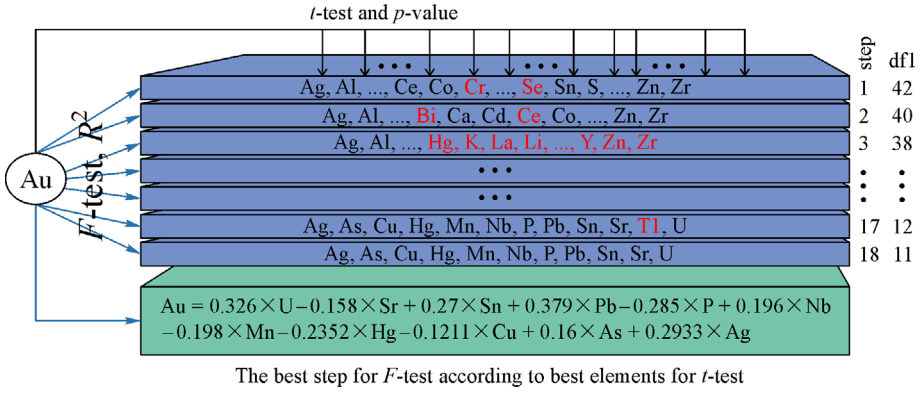

Fig. 5 The differences between the *t*-test and *F*-test for optimizing model in host rock of Glojeh deposit.

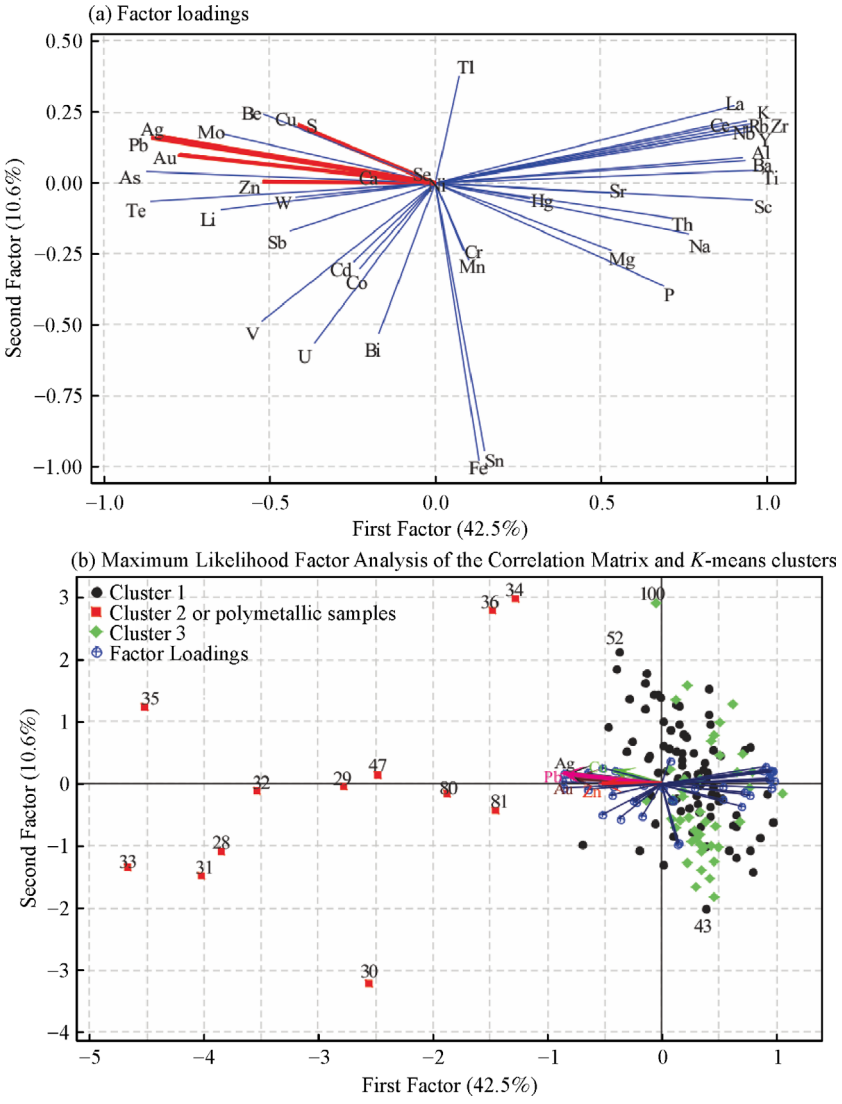

Fig. 6 Results of robust factor analysis according to clr-transformed data: (a) biplot of the first vs. second loading factors; (b) biplot of the first vs. second loading factors accompanied to scores of samples which categorized in three cluster by *K*-means clustering method. 12 samples (cluster 2) that have the maximum distance from centroid, indicates the intensity of vein polymetallic mineralization in the Glojeh deposit.

which contains all 42 elements and threshold value (β_0) were determined at the first step. The elimination process of insignificant elements of the FM led to decrease of df1 from 42 (in the FM) to 8 elements (Ag, Al, As, Hg, Mg, Te, Pb, and Sb), after 23 steps. Accordingly, Mn and Se at the first step and Nb at the 22nd step were eliminated and the ORM was obtained (Table 6) for host rock population while all the elements contributed by 95% CL (*p*-value lower than 0.05; Table 7). The process continues until the lowest *t*-value of the elements are equal to or greater than the TINV function. After 23 steps and by applying TINV, TDIST, FDIST, and other controlling functions, the ORM model was determined that it is given by Eq. (8).

$$\begin{aligned}\text{Au value by regeression} &= 0.285 \times \text{Ag} - 0.1448 \times \text{Al} + 0.152 \times \text{As} - 0.227 \times \text{Hg} \\ &\quad - 0.238 \times \text{Mg} + 0.22 \times \text{Pb} + 0.14 \times \text{Sb} + 0.266 \times \text{Te}.\end{aligned} \tag{8}$$

In the optimization process modeling, after 23 steps and elimination of 38 elements, the correlation was decreased gradually from 70.1% to 59.6% (Table 6). The result of the LINEST function at the 23rd step (Table 7) shows that the maximum TDIST function for As is equal to 0.05, whereas Ag has the minimum error ($t_{\text{Ag}} > t_{\text{Te}} > t_{\text{Hg}} > t_{\text{Pb}} > t_{\text{Mg}} > t_{\text{Al}} > t_{\text{Sb}} > t_{\text{As}}$). As concluded, Ag, Te, and Hg are fundamental

Table 6 Model optimization in host rock of Glojeh deposit by CL of 95%

Step	R^2	df1	df2	F value	FDIST	TINV of 95%	Elements by minimum t-value		Elements by minimum t-value	
							Ele. 1	Value	Ele. 2	Value
1	0.701	42	83	4.653	1.20E−09	1.989	Mn	0.09	Se	0.07
2	0.701	40	85	5.002	2.40E−10	1.988	Na	0.13	Bi	0.11
3	0.701	38	87	5.386	4.50E−11	1.988	Li	0.14	Zr	0.13
4	0.701	36	89	5.811	7.70E−12	1.987	Ce	0.31	Ni	0.17
5	0.701	34	91	6.279	1.30E−12	1.986	Rb	0.4	V	0.34
6	0.7	32	93	6.788	2.10E−13	1.986	La	0.54	Mo	0.42
7	0.698	30	95	7.347	3.40E−14	1.985	Zn	0.74	Y	0.44
8	0.696	28	97	7.953	5.70E−15	1.985	Tl	0.79	Sc	0.78
9	0.693	26	99	8.597	1.10E−15	1.984	U	0.99	Ti	0.53
10	0.689	24	101	9.326	2.00E−16	1.984	K	1.28	Fe	1.18
11	0.681	22	103	10.007	6.20E−17	1.983	Cr	1.24	Sn	1.1
12	0.67	20	105	10.69	2.70E−17	1.983	Cd	1.33	P	0.9
13	0.661	18	107	11.615	8.50E−18	1.982	Be	1.55	Ba	1.35
14	0.649	16	109	12.649	3.50E−18	1.982	Co	1.38		
15	0.643	15	110	13.254	2.30E−18	1.982	Cu	1.41		
16	0.637	14	111	13.934	1.50E−18	1.982	Sr	1.56		
17	0.629	13	112	14.63	1.20E−18	1.981	Th	1.4		
18	0.622	12	113	15.554	7.20E−19	1.981	Ca	1.29		
19	0.617	11	114	16.721	3.70E−19	1.981	W	1.64		
20	0.608	10	115	17.859	3.00E−19	1.981	Threshold limit	1.69		
21	0.613	10	116	18.44	8.90E−20	1.981	S	1.54		
22	0.605	9	117	19.992	5.90E−20	1.98	Nb	1.76		
23	0.596	8	118	21.72	5.30E−20	1.98	As	1.954		
24	0.582	7	119	23.716	6.40E−20	1.98	Al	2.15		

Table 7 The matrix of LINEST function results of BH2N1 after 23 stages

Step	Parameter	Te	Sb	Pb	Mg	Hg	As	Al	Ag
1	β_k	0.266	0.143	0.220	−0.238	−0.227	0.152	−0.145	0.285
2	SE_k	0.091	0.072	0.085	0.099	0.081	0.078	0.068	0.091
3	R^2 and SE_y	0.596	0.570						
4	F and df	21.720	118						
5	SSR and SSE	56.452	38.336						
6	t-value = ABS(β_k/SE_k)	2.93	1.98	2.58	2.40	2.79	1.954	2.12	3.12
7	P value = TDIST function	0.004	0.050	0.011	0.013	0.006	0.05	0.036	0.002

elements of a productive model for the host rock population. The association of Te, Hg, and Au is typically encountered in magmatic, metamorphic, and hydrothermal deposits (Cook and Ciobanu, 2005; Ciobanu et al., 2006), and placer deposits could be derived from these rocks (Naumov and Osovetsky, 2013; Parnell et al., 2016).

4.5 Gold evaluation within veins

A total number of 12 samples from a second cluster of K-means analysis was taken for polymetallic vein modeling. With consideration of β_0 as a threshold value, we were allowed that only use 11 elements in modeling which satisfy the df value in the modeling process. This restriction is caused by the low number of samples and the type of deposit (polymetallic vein ore deposit). Ag, Al, As, Cu, Hg, Mg, Pb, Sb, Te, and Zn which were the most important elements for modeling Au in the host rock, were selected to further analysis and Au modeling at the veins and brecciated zones. These elements shows genetic

relationship with this type of mineralization. The $R^2 = 99\%$ is obtained in 5th step of modeling, whereas in the first step it was equal to 88%. The majority of this difference is caused by β_0, since after elimination of β_0, a significant increase in R^2 happened from 88% to 99% (Table 8, 3rd step). As already predicted, according to epigenetic vein genesis of Glojeh deposit, all the enriched elements have a similar genesis and they are related to each other during mineralization. So assigning a β_0 in the veins modeling could create huge errors in the model.

The low sample numbers may cause an error in the numerical modeling which should be controlled by specific factors. It is observed that in the 5th step, the TINV value is equal to 2.571 at CL = 95%. It must be clear at the present step, Ag is presented with the lowest *t*-value of 1.09 (67.3% CL in model). But elimination of this element may not be reasonable (Table 8 and Table 9). An exploration expert must assess the results with existing realities and limitations. To realize these goals, elemental percentage error of ORM in step 5 was calculated and the total average CL% was considered. It was observed that Ag participated in the model by CL% of 67.3% and an average CL% of all elements is about 83% (Table 8). The ORM is introduced by Eq. (9):

$$\begin{aligned} \mathrm{Au} = {} & 0.538 \times \mathrm{Sb} - 0.888 \times \mathrm{Pb} + 0.226 \\ & \times \mathrm{Mg} + 0.141 \times \mathrm{Cu} \\ & + 0.665 \times \mathrm{As} - 0.266 \times \mathrm{Al} + 0.552 \times \mathrm{Ag}. \end{aligned} \quad (9)$$

The *t*-values of Sb, Pb, Mg, Cu, As, Al, and Ag were calculated by the ratio of the related coefficients in the ORM model (Eq. (9)) to the standard errors (equals 0.54, 0.15, 0.38, 0.13, 0.66, 0.11, and 0.48, respectively). In descending order, *t*-values are sorted as:

$$t_{\mathrm{Sb}} > t_{\mathrm{Al}} > t_{\mathrm{As}} > t_{\mathrm{Mg}} > t_{\mathrm{Pb}} > t_{\mathrm{Cu}} > t_{\mathrm{Ag}}.$$

5 Discussion

A proposed algorithm for the analysis of geochemical data in the Glojeh polymetallic vein deposit was provided, as depicted in Fig. 7(a). Through all the data (138 samples), an optimized model consisting of Ag, As, Cu, Hg, Mn, Nb, P, Pb, Sn, Sr, and U was obtained, where Cu has the lowest CL% value of 92%. Based on the different genetic properties of vein and host rocks, a separation was conducted by *K*-means cluster and robust factor analysis of centered log-ratio (clr) transformed data. Accordingly, two distinct process carried out for different vein (12 samples) and background (126 samples) populations. According to the schematic flowchart in Fig. 7(b), the optimization process was applied using controlling functions in each of the models. In the first step, the FM was constructed by LINEST function and *p*- and *t*-values for each predictor and *F*-value of the model was calculated. At the FM all the *K* elements participated in model, and accordingly at each step the number of *L* predictor was eliminated and a new RM was obtained. After several steps and eliminating insignificant elements, the optimal reduced model (ORM) was determined. The ORM is constructed where all the remaining elements in the model had *p*-values lower than 10% and *F*-values showed gradual increases throughout the process. The importance and priority of each parameter in ORM of veins, host rock and the entire data has been introduced in Fig. 8. The CL% was determined from 1- TDIST(*t*-value, df, 2 tail) function of each elements. Hence, the average CL% values were calculated about 98.5%, 97.7%, and 83% for all the data, background, and vein population, respectively (Fig. 8). Maximum and minimum *t*-values of the ORM from all the data, background, and vein population are Pb(5.2)-Cu (1.75), Ag(3.2)-As(1.95), and Sb(4.17)-Cu(1.32), respectively. Accordingly, minimum CL% of elements in ORMs are about 92%, 95%, and 68% relating to Cu, As, and Cu.

Table 9 TDIST value (error) and CL of 95% for elements which presented in vein model

Element	*t*-value	ABS (*t*-value)	TDIST	CL%
Sb	4.173	4.173	0.009	99.1
Pb	−1.350	1.350	0.235	76.5
Mg	1.740	1.740	0.142	85.8
Cu	1.322	1.322	0.243	75.7
As	1.741	1.741	0.142	85.8
Al	−1.752	1.752	0.140	86.0
Ag	1.087	1.086	0.327	67.3

Table 8 Model optimization in vein of Glojeh deposit by average CL of 83% at 5th step

Step	R^2	df1	df2	FDIST	TINV of 95%	Element by minimum *t*-value	
						Element	Value
1	0.877	10	1	7.38E–01	12.706	Te	0.072
2	0.877	9	2	4.49E–01	4.303	Zn	0.050
3	0.876	8	3	2.27E–01	3.182	Threshold limit	0.045
4	0.995	8	4	2.71E–04	2.776	Hg	0.445
5	0.995	7	5	2.41E–05	2.571	Ag	1.090

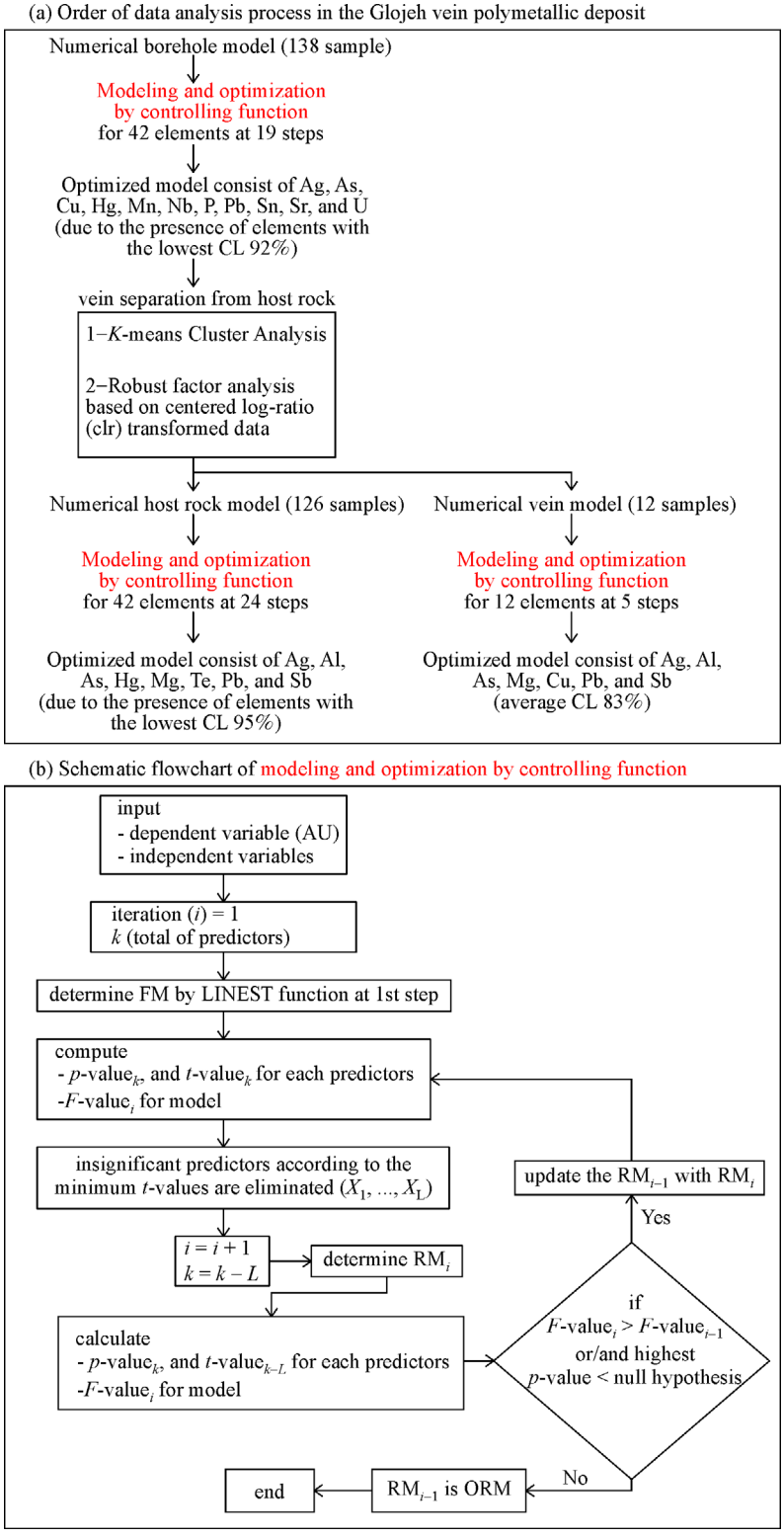

Fig. 7 The procedure of modeling the geochemical data in the Glojeh deposit using the proposed algorithm, order of data analysis (a), and a schematic flowchart of modeling and optimization by controlling function (b).

By considering CL% of 90% for elements in the entire data model, processing continues until the lowest *t*-value of elements is equal to or greater than the TINV function, as happened in the 18th step of optimization. The ORM for the entire data includes Ag, As, Cu, Hg, Mn, Nb, P, Pb, Sn, Sr, and U elements. By comparing the actual values and the predicted values from the model, R^2 is equal to 0.74. This seems reasonable according to the wide differences in the veins and background values in the all data sets. According to the Zhu et al. (2011) theories, the associations of Au-As or Au-Sb is more common in different gold deposit models. It was revealed that Au-Sb associations is strong in vein and brecciated zone, while there is a good correlation of Au-As in all the data at the Glojeh deposit (in the prospecting model).

With the implementation of the LINEST function in 24 steps and by application of different controlling functions (TINV, TDIST, FDIST, and other functions) in the host rock, the FM is optimized after 23 steps. It consists of Ag, Al, As, Hg, Mg, Te, Pb, and Sb elements with minimum CL% of 0.95% (Fig. 8). At first, the statistical correlation in the FM with 42 variables was found to be 70.1%. After gradual elimination of 35 insignificant variables, the ORM (R^2= 0.596) was created while only about 10% of this correlation was reduced (Table 6). Due to the weak genetic association of polymetallic elements in the host rock, the obtained R^2 value for Au would be satisfactory.

According to epigenetic vein genesis of mineralization, the threshold value could create huge errors in the vein model. Where there are 12 vein samples, the LINEST function could start only with 11 elements to create the FM. Finally, an ORM for vein population was obtained for Ag, Al, As, Cu, Mg, Pb, and Sb elements with the R^2 value up to 99%. The maximum and minimum CL% was determined for Sb and Ag as 0.991 and 0.673, respectively; whereas the average CL% for ORM was about 83% (Fig. 8).

The variation trend from maximum to minimum *t*-values in the ORM (Fig. 8) for each population are shown in Fig. 9. The indicated veins are depicted based on the boreholes BH2N1, BH2N2, and trench TR2 intersections (with vein). Therefore, the Au variation is actually based on the differences in distribution and the population of the certain elements which participate throughout the borehole, background, and vein models. Following the modeling process, the important elements have been recognized according to *t*-values as:

$$\text{For host rock}: t_{Ag}>t_{Te}>t_{Hg}>t_{Pb}>t_{Mg}>t_{Al}>t_{Sb}>t_{As},$$

$$\text{For vein}: t_{Sb}>t_{Al}>t_{As}>t_{Mg}>t_{Pb}>t_{Cu}>t_{Ag},$$

$$\text{For all data}: t_{Pb}>t_{Ag}>t_{P}>t_{Hg}>t_{Mn}>t_{Nb}>t_{U}>t_{Sr}>t_{Sn}>t_{As}>t_{Cu}.$$

6 Conclusions

The results of the LINEST and supplementary function are easily simulated in geochemical modeling and mineral exploration. The backward elimination procedure using LINEST and supplementary function has been applied in

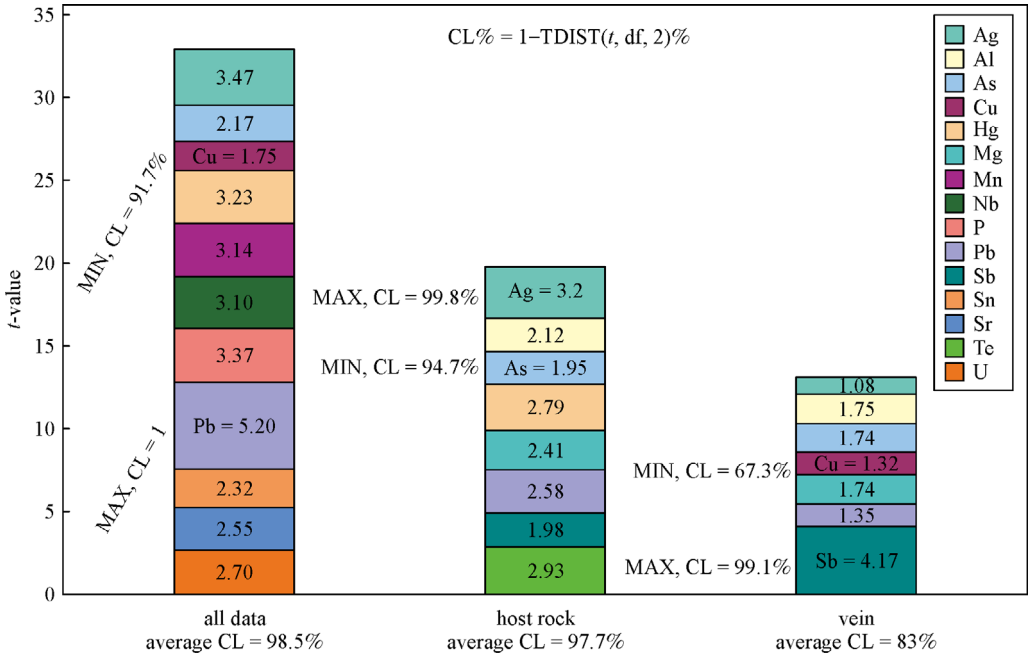

Fig. 8 Comparison the *t*-value of elements in modeling from vein and brecciated zone, host rock and all the data throughout the BH2N1 borehole.

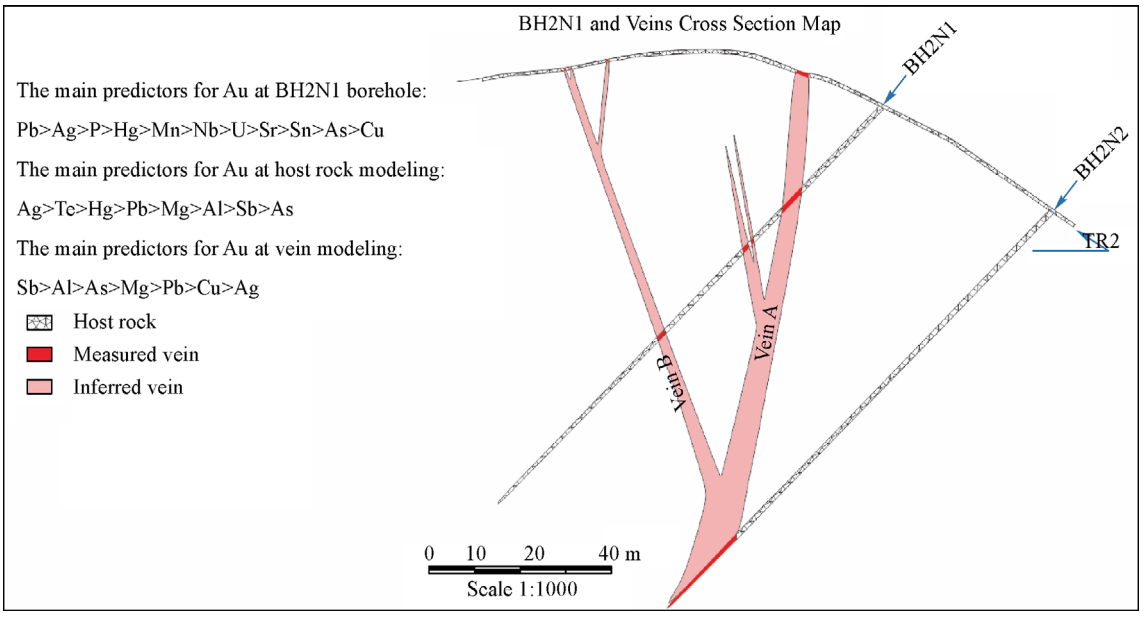

Fig. 9 The variation trend of *t*-values for elements which participate for distinct background, vein and throughout all the BH2N1 borehole models.

geochemical exploration, and it could be expanded for mineral modeling process. Using LINEST and controlling functions for the entire data, the most prominent elements for Au modeling were determined as Pb > Ag > P > Hg > Mn > Nb > U > Sr > Sn > As > Cu, after eighteen steps. In order to determine the spatial distribution of elements and samples, accompanied by separation anomalous vein from the host rock, the multivariate analysis can be applied. The *K*-means cluster and factor analysis based on clr-transformed data can represent the polymetallic samples from the host rock. A total number of 12 samples (cluster 2) that had the maximum distance from centroid will indicate the intensity of vein polymetallic mineralization in the Glojeh deposit. In order to avoid closure effects between 2D loadings factors and scores of the geochemical compositional data set, factor analysis of clr-transformed data led to better results. This finding confirms the results from *K*-means clustering to recognize anomalous samples. Accordingly, the anomalous samples are outstanding for Pb, Ag, As, Te, Au, Mo, Zn, Be, Cu elements. In host rock modeling, Ag, Te, and Hg are the most important elements which could be outlined based on their similar geochemical mobility and characteristics. Therefore, Te and Hg have been considered as pathfinder elements for Glojeh polymetallic vein deposit. According to the different Au dispersion pattern (related to zonation of Sb > Al > As > Mg > Pb > Cu > Ag in vein and Ag > Te > Hg > Pb > Mg > Al > Sb > As in host rock), the concentration ratio of $(t_{Sb} \times t_{Al})_{vein}/(t_{Sb} \times t_{Al})_{background}$ can be suggested as an enrichment index or geochemical criteria within the Glojeh polymetallic deposit. Accordingly, the most common elements in the presented models are Pb, Ag, and As. The geochemical behavior of Pb, Ag, and As together with Au are well correlated under the different physicochemical conditions, which could be due to their geochemical characteristics in the hydrothermal ore deposits. Therefore, Glojeh may be classified as a vein-style Au (Ag, As)-polymetallic deposit.

Acknowledgements The authors are grateful to the Iranian Mines and Mining Industries Development & Renovation Organization (IMIDRO) for their permission to have access to Glojeh deposit dataset.

References

Abdi H, Williams L J, Valentin D (2013). Multiple factor analysis: principal component analysis for multi-table and multi-block data sets. Comput Stat, 5(2): 149–179

Almasi A, Jafarirad A, Kheyrollahi H, Rahimi M, Afzal P (2014). Evaluation of structural and geological factors in orogenic gold type mineralisation in the Kervian area, north-west Iran, using airborne geophysical data. Explor Geophys, 45(4): 261–270

Angelo R T, Cringan M S, Chamberlain D L, Stahl A J, Haslouer S G, Goodrich C A (2007). Residual effects of lead and zinc mining on freshwater mussels in the Spring River Basin (Kansas, Missouri, and Oklahoma, USA). Sci Total Environ, 384(1–3): 467–496

Bierlein F P, McKnight S (2005). Possible intrusion-related gold systems in the western Lachlan Orogen, southeast Australia. Econ Geol, 100: 385–398

Bise C J (2013). Modern American Coal Mining: Methods and Applications. Society for Mining, Metallurgy and Exploration

Blythe J N, Lea D W (2008). Functions of height and width dimensions in the intertidal mussel, *Mytilus californianus.* J Shellfish Res, 27(2):

385–392

Borah P, Singh M K, Mahapatra S (2015). Estimation of degree-days for different climatic zones of North-East India. Sustainable Cities and Society, 14: 70–81

Breidenbach J, McRoberts R E, Astrup R (2016). Empirical coverage of model-based variance estimators for remote sensing assisted estimation of stand-level timber volume. Remote Sens Environ, 173: 274–281

Briand G, Hill R C (2013). Teaching basic econometric concepts using Monte Carlo simulations in Excel. Int Rev Econ Educ, 12: 60–79

Buccianti A, Grunsky E (2014). Compositional data analysis in geochemistry: are we sure to see what really occurs during natural processes? Elsevier

Cheng Q (2007). Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol Rev, 32(1–2): 314–324

Ciobanu C L, Cook N J, Spry P G (2006). Telluride and selenide minerals in gold deposits—How and why? Mineral Petrol, 87(3−4): 163–169

Cook N J, Ciobanu C L (2005). Tellurides in Au deposits: implications for modelling. In: Mao J W, Bierlein F P, eds. Mineral Deposit Research: Meeting the Global Challenge. Proceedings of the 8th Biennial SGA Meeting, Beijing, China, 1387–1390

Cordell H J, Clayton D G (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet, 70(1): 124–141

Darabi-Golestan F, Ghavami-Riabi R, Asadi-Harooni H (2013). Alteration, zoning model, and mineralogical structure considering lithogeochemical investigation in Northern Dalli Cu–Au porphyry. Arab J Geosci, 6(12): 4821–4831

Darabi-Golestan F, Hezarkhani A (2016). High precision analysis modeling by backward elimination with attitude on interaction effects on Au (Ag)-polymetallic mineralization of Glojeh, Iran. J Afr Earth Sci, 124: 505–516

Darabi-Golestan F, Hezarkhani A (2017a). Evaluation of elemental mineralization rank using fractal and multivariate techniques and improving the performance by log-ratio transformation. J Geochem Explor, doi: 10.1016/j.gexplo.2017.09.011

Darabi-Golestan F, Hezarkhani A (2017b). R- and Q-mode multivariate analysis to sense spatial mineralization rather than uni-elemental fractal modeling in polymetallic vein deposits. Geosystem Engineering, doi: 10.1080/12269328.2017.1407266

Darabi-Golestan F, Hezarkhani A, Zare M (2017). Assessment of ^{226}Ra, ^{238}U, ^{232}Th, ^{137}Cs and ^{40}K activities from the northern coastline of Oman Sea (water and sediments). Mar Pollut Bull, 118(1−2): 197–205

Darabi-Golestan F, Hezarkhani A, Zare M (2014). Interpretation of the sources of radioactive elements and relationship between them by using multivariate analyses in Anzali Wetland Area. Geoinformatics & Geostatistics. An Overview, 1: 4

DeCoursey W (2003). Statistics and Probability for Engineering Applications. New York: Elsevier

Dehak N, Kenny P J, Dehak R, Dumouchel P, Ouellet P (2011). Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process, 19(4): 788–798

Dora M, Randive K (2015). Chloritisation along the Thanewasna shear zone, Western Bastar Craton, Central India: its genetic linkage to Cu–Au mineralisation. Ore Geol Rev, 70: 151–172

Egozcue J J, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C (2003). Isometric logratio transformations for compositional data analysis. Math Geol, 35(3): 279–300

Fávaro D, Damatto S, Moreira E, Mazzilli B, Campagnoli F (2007). Chemical characterization and recent sedimentation rates in sediment cores from Rio Grande reservoir, SP, Brazil. J Radioanal Nucl Chem, 273(2): 451–463

Filzmoser P, Hron K (2008). Outlier detection for compositional data using robust methods. Math Geosci, 40(3): 233–248

Filzmoser P, Hron K, Reimann C (2009). Principal component analysis for compositional data with outliers. Environmetrics, 20(6): 621–632

Fox R J (1983). Confirmatory Factor Analysis. Wiley Online Library

Garson G (2012). Multiple regression (statistical associates blue book series). Asheboro, NC: Statistical Associates Publishers

Grancea L, Bailly L, Leroy J, Banks D, Marcoux E, Milési J, Cuney M, André A, Istvan D, Fabre C (2002). Fluid evolution in the Baia Mare epithermal gold/polymetallic district, Inner Carpathians, Romania. Miner Depos, 37(6–7): 630–647

Guha S, Mishra N (2016). Clustering data streams. In: Garofalakis M, Gehrke J, Rastogi R. Data Stream Management: Processing High-Speed Data Streams. Springer Berlin Heidelberg, 169–187

Hamilton A, Campbell K, Rowland J, Browne P (2017). The Kohuamuri siliceous sinter as a vector for epithermal mineralisation, Coromandel Volcanic Zone, New Zealand. Miner Depos, 52(2): 181–196

Hargreaves B R, McWilliams T P (2010). Polynomial trendline function flaws in Microsoft Excel. Comput Stat Data Anal, 54(4): 1190–1196

Hezarkhani A (2008). Hydrothermal evolution of the Miduk porphyry copper system, Kerman, Iran: a fluid inclusion investigation. Int Geol Rev, 50(7): 665–684

Hill T, Lewicki P (2006). Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. Tulsa: StatSoft, Inc.

Huang Z (1998). Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov, 2(3): 283–304

Jiang S, Liu M, Hao J, Qian W (2015). A bi-layer optimization approach for a hybrid flow shop scheduling problem involving controllable processing times in the steelmaking industry. Comput Ind Eng, 87: 518–531

Jovic S M, Guido D M, Ruiz R, Páez G N, Schalamuk I B (2011). Indium distribution and correlations in polymetallic veins from Pingüino deposit, Deseado Massif, Patagonia, Argentina. Geochem Explor Environ Anal, 11(2): 107–115

Karamanis D, Ioannides K, Stamoulis K (2009). Environmental assessment of natural radionuclides and heavy metals in waters discharged from a lignite-fired power plant. Fuel, 88(10): 2046–2052

Kutner M H, Nachtsheim C J, Neter J, Li W (2005). Applied Linear Statistical Models. New York: McGraw-Hill Irwin

Larose D T (2006). Data Mining Methods & Models. New York: John Wiley & Sons: 1–223

Larose D T (2003). Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken: John Wiley & Sous., 1–223

Liu Y, Cheng Q, Zhou K, Xia Q, Wang X (2016). Multivariate analysis

for geochemical process identification using stream sediment geochemical data: a perspective from compositional data. Geochem J, 50(4): 293–314

MacQueen J (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, 281–297

Martínez-Abad I, Cepedal A, Arias D, Fuertes-Fuente M (2015). The Au–As (Ag–Pb–Zn–Cu–Sb) vein-disseminated deposit of Arcos (Lugo, NW Spain): mineral paragenesis, hydrothermal alteration and implications in invisible gold deposition. J Geochem Explor, 151: 1–16

Mehrabi B, Siani M G, Azizi H (2014). The genesis of the epithermal gold mineralization at North Glojeh veins. NW Iran. IJSAR, 15: 479–497

Mehrabi B, Siani M G, Goldfarb R, Azizi H, Ganerod M, Marsh E E (2016). Mineral assemblages, fluid evolution, and genesis of polymetallic epithermal veins, Glojeh district. NW Iran. Ore Geol Rev, 78: 41–57

Mihai D, Mocanu M (2015). Statistical considerations on the *k*-means algorithm. Annals of the University of Craiova-Mathematics and Computer Science Series, 42: 365–373

Mohammadi N M, Hezarkhani A, Maghsoudi A (2018). Application of *K*-means and PCA approaches to estimation of gold grade in Khooni district (central Iran). Acta Geochimica, 37(1): 102–112

Myers R H, Montgomery D C, Anderson-Cook C M (2016). Response Surface Methodology: Process and Product Optimization Using Designed Experiments. John Wiley & Sons

Namhata A, Zhang L, Dilmore R M, Oladyshkin S, Nakles D V (2017). Modeling changes in pressure due to migration of fluids into the above zone monitoring interval of a geologic carbon storage site. Int J Greenh Gas Control, 56: 30–42

Naumov V, Osovetsky B (2013). Mercuriferous gold and amalgams in Mesozoic-Cenozoic rocks of the Vyatka-Kama Depression. Lithol Miner Resour, 48(3): 237–253

Nude P M, Asigri J M, Yidana S M, Arhin E, Foli G, Kutu J M (2012). Identifying pathfinder elements for gold in multi-element soil geochemical data from the Wa-Lawra belt, northwest Ghana: a multivariate statistical approach. Int J Geosci, 3(01): 62–70

Oyman T, Minareci F, Pişkin Ö (2003). Efemcukuru B-rich epithermal gold deposit (Izmir, Turkey). Ore Geol Rev, 23(1–2): 35–53

Parnell J, Spinks S, Bellis D (2016). Low-temperature concentration of tellurium and gold in continental red bed successions. Terra Nova, 28 (3): 221–227

Pawlowsky-Glahn V, Egozcue J (2006). Compositional data and their analysis: an introduction. Geol Soc Lond Spec Publ, 264(1): 1–10

Radosavljević S A, Stojanović J N, Vuković N S, Radosavljević-Mihajlović A S, Kašić V D (2015). Low-temperature Ni-As-Sb-S mineralization of the Pb (Ag)-Zn deposits within the Rogozna ore field, Serbo-Macedonian Metallogenic Province: ore mineralogy, crystal chemistry and paragenetic relationships. Ore Geol Rev, 65: 213–227

Ramasamy V, Sundarrajan M, Paramasivam K, Meenakshisundaram V, Suresh G (2013). Assessment of spatial distribution and radiological hazardous nature of radionuclides in high background radiation area, Kerala, India. Appl Radiat Isot, 73: 21–31

Reith F, McPhail D, Christy A (2005). Bacillus cereus, gold and associated elements in soil and other regolith samples from Tomakin Park Gold Mine in southeastern New South Wales, Australia. J Geochem Explor, 85(2): 81–98

Remenyi D, Onofrei G, English J (2011). An introduction to statistics using Microsoft Excel. Academic Conferences and Publishing International Ltd.

Røislien J, Omre H (2006). T-distributed random fields: a parametric model for heavy-tailed well-log data. Math Geol, 38(7): 821–849

Samal A R, Mohanty M K, Fifarek R H (2008). Backward elimination procedure for a predictive model of gold concentration. J Geochem Explor, 97(2–3): 69–82

Savazzi E, Reyment R (1999). Aspects of Multivariate Statistical Analysis in Geology. Elsevier

Soheily-Khah S, Douzal-Chouakria A, Gaussier E (2016). Generalized *k*-means-based clustering for temporal data under weighted and kernel time warp. Pattern Recognit Lett, 75: 63–69

Stanciu C (1973). Hydrothermal alteration of Neogene volcanics rocks from ore deposits in Gutai Mountains (East Carpathians). Rev Roum Geol Geophys Geogr Ser Geol, 17: 43–62

Suresh G, Sutharsan P, Ramasamy V, Venkatachalapathy R (2012). Assessment of spatial distribution and potential ecological risk of the heavy metals in relation to granulometric contents of Veeranam lake sediments, India. Ecotoxicol Environ Saf, 84: 117–124

Székely G J, Rizzo M L (2013). The distance correlation *t*-test of independence in high dimension. J Multivariate Anal, 117: 193–213

Templ M, Filzmoser P, Reimann C (2008). Cluster analysis applied to regional geochemical data: problems and possibilities. Appl Geochem, 23(8): 2198–2213

Tokatli C, Köse E, Çiçek A (2014). Assessment of the effects of large borate deposits on surface water quality by multi statistical approaches: a case study of Seydisuyu Stream (Turkey). Pol J Environ Stud, 23: 1741–1751

Vriend S, Van Gaans P, Middelburg J, De Nijs A (1988). The application of fuzzy *c*-means cluster analysis and non-linear mapping to geochemical datasets: examples from Portugal. Appl Geochem, 3 (2): 213–224

Wang W, Zhao J, Cheng Q (2014). Mapping of Fe mineralization-associated geochemical signatures using logratio transformed stream sediment geochemical data in eastern Tianshan, China. J Geochem Explor, 141: 6–14

Yalta A T (2008). The accuracy of statistical distributions in Microsoft® Excel 2007. Comput Stat Data Anal, 52(10): 4579–4586

Yang L, Wang Q, Liu X (2015). Correlation between mineralization intensity and fluid–rock reaction in the Xinli gold deposit, Jiaodong Peninsula, China: constraints from petrographic and statistical approaches. Ore Geol Rev, 71: 29–39

Yousefi M, Kamkar-Rouhani A, Carranza E J M (2014). Application of staged factor analysis and logistic function to create a fuzzy stream sediment geochemical evidence layer for mineral prospectivity mapping. Geochem Explor Environ Anal, 14(1): 45–58

Zaiontz C (2014). Real statistics using Excel. http://www.real-statistics.com/regression/power-regression/

Zarandi M F, Yazdi E H (2008). A type-2 fuzzy rule-based expert system model for portfolio selection. Proceeding of The 11th Joint Conference On Information Sciences. Atlantis Press

Zhang D, Cheng Q, Agterberg F, Chen Z (2016). An improved solution of local window parameters setting for local singularity analysis based on Excel VBA batch processing technology. Comput Geosci, 88: 54–66

Zhao X, Xue C, Symons D T, Zhang Z, Wang H (2014). Microgranular enclaves in island-arc andesites: a possible link between known epithermal Au and potential porphyry Cu–Au deposits in the Tulasu ore cluster, western Tianshan, Xinjiang, China. J Asian Earth Sci, 85: 210–223

Zhu Y, An F, Tan J (2011). Geochemistry of hydrothermal gold deposits: a review. Geoscience Frontiers, 2(3): 367–374

Ziaii M, Carranza E J M, Ziaei M (2011). Application of geochemical zonality coefficients in mineral prospectivity mapping. Comput Geosci, 37(12): 1935–1945