JME

Journal of Mining & Environment,
Vol.8, No.4, 2017, 673-682.
DOI: 10.22044/jme.2017.966

Application of continuous restricted Boltzmann machine to detect multivariate anomalies from stream sediment geochemical data, Korit, East of Iran

A. Aryafar[1*] and H. Moeini[2]

1. Department of Mining, Faculty of Engineering, University of Birjand, Birjand, Iran
2. Department of Mining and Metallurgical Engineering, University of Yazd, Yazd, Iran

Received 2 May 2017; received in revised form 1 July 2017; accepted 5 July 2017

**Corresponding author: aaryafar@birjand.ac.ir (A. Aryafar).*

Abstract

Anomaly separation using stream sediment geochemical data has an essential role in regional exploration. Many different techniques have been proposed to distinguish anomalous from study area. In this research, a continuous restricted Boltzmann machine (CRBM), which is a generative stochastic artificial neural network, was used to recognize the mineral potential area in Korit 1:100000 sheet, located 15 km south of Tabas, South Khorasan Province (East of Iran). For this purpose, 470 geochemical stream sediment samples were collected from the study area and analyzed for 36 elements. In order to achieve the goal, in the first step, the robust factor analysis on compositional data was applied to reduce the data dimension and to limit the multivariate analysis by selecting the main components of mineralization. In this procedure, the third factor (out of 6) consisting of Cu, Pb, Zn, Sn, and Sb, related to the metallogenic properties, was considered as the input set in CRBM. In continuation, the CRBM structure with the best efficiency after trying different parameters was stabilized. High-identified error values or anomalies were exteracted using two different thresholds (ASC and ASE) after training with the whole data and reconstructing it by CRBM. The anomalies were then mapped. These indicated the promissing areas. The field studies and existing mining indices confirmly demonestrated the results obtained by CRBM.

Keywords: *Stream Sediment, CRBM, Robust Factor Analysis, Korit.*

1. Introduction

Geochemical anomaly detection using stream sediment sampling is the main and primary phase of mineral exploration that together with other methods like remote sensing and geophysics plays a major role in the regional exploration. Analysis of the stream sediment samples can reveal various geochemical anomalies, some of which can be considered as a surficial geochemical signature of the deposit-type [1]. Through recent decades, there has been developed and proposed numerous different methods to delineate anomalous and prospecting areas, from conventional parametric statistical thresholds to non-parametric methods like multi-fractal thresholds and from univariate to multivariate methods [2]. Factor analysis, one of the multivariate analysis methods, has been widely used for the interpretation of stream sediment geochemical data [3-6]. It is often applied as a tool for exploratory data analysis, dimension reduction, and to determine multi-element geochemical signatures that reflect the presence of mineralization [7]. Although in recent years, after Aitchison and some statisticians presented some theoretical solutions to deal with the problems of compositional data, new tools were developed and opened a window to correct interpretations. It spread through all the multivariate conventional methods like factor analysis, and made re-definition of the methodologies to work in simplex (i.e. the closed space of compositional data). The comparisons made showed much improved results in the studied cases [8]. With the beginning of the millennium, novel methods emerged from computer science nature-inspired models of “machine learning”. They were widely used in various areas other than computer and electronics from biotechnology and medicine to earth sciences. Neural networks are among the most expanding and applicable algorithms that have been used in mining exploration and exploitation so far [9]. The new field of “deep learning” has made a revolution in pattern recognition and anomaly identification. Deep learning algorithms transform their inputs through more layers than shallow learning algorithms. At each layer, the signal is transformed by a processing unit, like an artificial neuron, whose parameters are 'learned' through training [10]. In 2006, a publication by Geoffrey Hinton and Ruslan Salakhutdinov drew additional attention by showing how many-layered feed forward neural network could be effectively pre-trained one layer at a time, treating each layer, in turn, as an unsupervised restricted Boltzmann machine, then tuning it using supervised back propagation [11]. A restricted Boltzmann machine (RBM) is a generative stochastic artificial neural network that can learn a probability distribution over its set of inputs [12]. RBMs have found applications in dimensionality reduction [13], classification [10], collaborative filtering [14], feature learning [15], and topic modeling [14]. The most recent application of RBMs and auto-encoder networks (that are stacks of RBMs) have been presented in regional exploration of mineral resources [16, 17]. The main objectives of this research work is to delineate the geochemical anomalies of Korit 1:100,000 geological sheet sampled from stream sediments using the methodology [16] presented above. For this purpose, a continuous RBM or CRBM was designed and scripted in MATLAB. The results, after an essential step of dimension reduction and focusing on the major elements of mineralization as input to CRBM network and setting the best parameters, showed some target points as anomalies. This is considerable due to the geological field evidences and further investigations that confirmed the anomalies and made way for novel tools of exploration.

2. Geological setting

The Korit geological 1:100,000 quadrangle map is a part of Tabas block in the vicinity of Lut and Posht-e-Badam blocks that covers the west of South Khorasan province and the south of the city of Tabas, East of Iran. The geological formations, faults and sampling points in the region are shown in Figure 1. A part of Parvadeh Coal Company is in the SW corner of the map. This zone is the largest and most complicated geological unit in Iran as it went through numerous incidents and experienced numerous phases of metamorphic, magmatic, orogenic, and folding activities. The Lut block and the Afghan block that once had formed one single continent were separated from each other and an ocean was formed between them. The formation, evolution, closure, and orogeny that resulted in the closure of this ocean are considered as major developments in the east and SE of Iran [18]. Cretaceous played a key role in east and southeast of Iran in regard to the tectono-magmatic events, sedimentary basins, and formation of various rock units like ophiolitic complex, carbonate rocks and flysches, and later events that occurred in Tertiary were the continuation of developments in Cretaceous. Cretaceous sediments are seen in relatively large areas within the Lut and Flysch zones. Triassic deposits formed in eastern, central, and western parts of Lut block, but they are much more extensive in the Tabas because of the activity of Kalmard, Kuh–Banan and Anar faults. The Triassic sediments in Iran are mainly of shallow marine or continental shelf nature [19]. The presence of plant remains and coal beds suggests a continental or lagoon environment for the deposits [20]. The deposits of Middle and Upper Triassic were laid over the other deposits with a disconformity. Extensive lead–zinc mineralization occurred in Triassic rocks within this block and formed numerous deposits. In most parts of this area, facies such as sandstone, shale, and marl were formed during Jurassic [21]. In Early Cretaceous, marine transgression took place within the block in a large scale and gave rise to rock facies such as conglomerate, sandstone, and clastic limestone. At the end of Cretaceous (Maastrichtian–Paleocene), most part of deposits underwent severe folding accompanied with metamorphism and led to the formation of a disconformity between the deposits of Paleocene and Late Cretaceous [22]. In parts of this area, sedimentary facies belonging to Cenozoic (Paleocene) began to form with basal conglomerate and sandstone. They overlie the older rocks with disconformity [23]. During Quaternary and concurrent to final shape-up of the highlands, many sedimentary basins lost their connections with seas and turned into vast plains where evaporative sediments such as gypsum and salt along with clay and marl with desert characteristics were formed [21]. A vast playa and some wetlands cover a large area of the map that extend to the east but in the western part, there are mountains and high hills with N-S trend. The most important formations with potential mineralization are Nayband, Garedu, Ab-haji and Shotori. The thick coal layers in the SW part of the map were bedded in Nayband (Ghadir) and Ab-Haji formations. The gypsum of Magu formation has also formed economical reserves in some parts [24].

Fluorite and barite deposits occur in the Triassic Shotori Formation of eastern Central Iran in the Tabas area, which are accompanied by lead–zinc. Important lead–zinc deposits have been formed at Triassic times within dolomites and dolomitic limestones (Shotori Formation in Eastern and Central Iran). It is worth mentioning that most of the fluorite reserves of Iran are the associated gangue minerals of this phase of lead–zinc mineralization [25, 26].

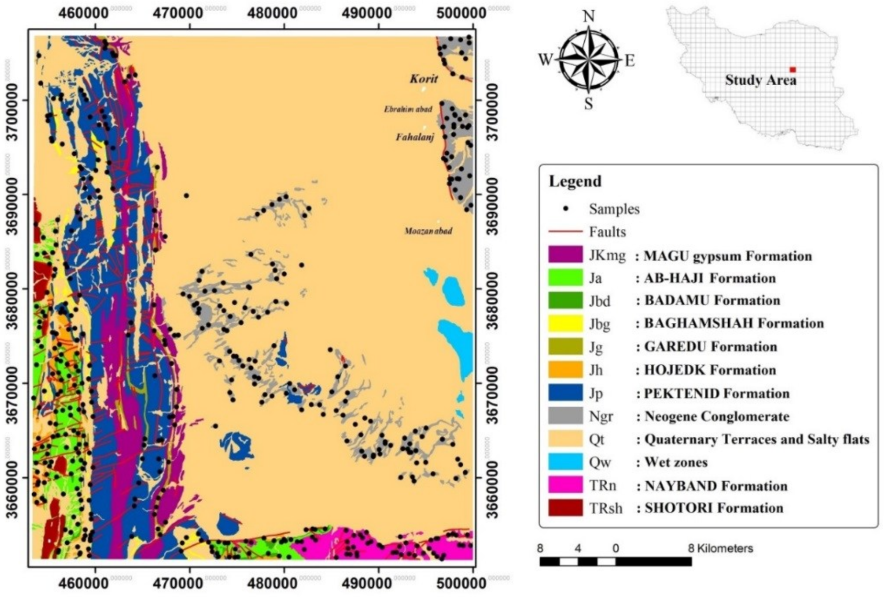

Figure 1. Geological map of Korit showing sampling locations, formations and faults.

3. Methodology

3.1. Robust factor analysis on compositional data

Application of robust factor analysis to the compositional data to reduce its dimension (and removing minor elements regarding their high uniqueness and less importance) and limiting the multivariate analysis to the main factors of mineralization is an essential step in preparing an exploratory geochemical input set for CRBM. Factor analysis is a popular multivariate technique used to approximate the *p* original variables of a dataset by linear combinations of a smaller number *k* of laten *t* variables, called factors. This must be done in such a way that the covariance matrix (or the correlation matrix) of the *p* original variables is fitted well [27]. When applying factor analysis to the compositional data, it is crucial to apply an appropriate transformation. A log-transformation will often reduce data skewness but does not accommodate the compositional nature of the data. Robust factor analysis can be obtained via a robust estimation of the covariance matrix in the *ilr*[1]-transformed space. The results obtained then have to be back-transformed to the *clr*[2]-transformed space that allows for an interpretation [8]. For the random vector *y*, the factor analysis model is defined as $y = \Lambda f + e$ with the factors *f* of dimension $k < D$, the error term *e*, and the loadings matrix Λ . Using the usual model assumptions, the factor analysis model can be written as $Cov(y) = \Lambda\Lambda^T + \Psi$, where $\Psi = Cov(e)$ has a diagonal form. The diagonal elements are called uniquenesses (or unique variances), and they include the part of the variance of the components of *y* that is not explained by the factors. In the case of compositional data, the vector *y* is the *clr* transformed random vector $(y = clr(x))$. The problem of singularity of $Cov(y)$ can be solved

[1] Isometric log-ratio
[2] Centered log-ratio

by projecting the diagonal matrix Ψ onto the hyper plane $y_1 + \cdots \quad y_D \quad 0$ formed by the *clr* transformation. The new model is then $Cov(y) = \Lambda\Lambda^T + H\Psi H^T$, where H comes from the definition of the clr transformation. Since $H^T = H$, then $\Psi^* = H\Psi H$. The matrix Ψ^* has no longer a diagonal form. Then in an iterative procedure, the eigenvalues $\Lambda = \left(\lambda_{ij}\right)$ are estimated from the relation $C(y) - \Psi^* = \Lambda\Lambda^T$. The diagonal elements ψ_i of Ψ are updated by $\psi_i = \{C(y)\}_{i.j} - \sum_{j=1}^{k} \lambda_{ij}^2 \quad \{C(y)\}_{i.j}$, which denotes the $\mathrm{i_{th}}$ diagonal element of $C(y)$. The iteration continues until the elements in Ψ^* stabilize [8]. For a better interpretation of the estimated loadings matrix Λ, an orthogonal or oblique rotation can be performed. The estimation of loadings and scores is based on the estimation of the covariance matrix $Cov(y)$. Traditionally, the estimation is done with the sample covariance matrix. However, it is well known that in case of outlying observations in the data set, this estimation may lead to very unreliable results. In this case, a robust estimation is required, and a popular choice is the MCD (minimum covariance determinant) estimator, for which also a fast algorithm is available [28]. The MCD estimator looks for a subset *h* out of *n* observations with the smallest determinant of their sample covariance matrix. The robust estimator of covariance is the sample covariance matrix of the *h* observations, multiplied by a factor for consistency at normal distribution. The subset size *h* determines the robustness of the estimator, and it can be varied between half the sample size and *n*. In order to deal with singularity of robust procedures in case of *clr*-transformed data, a way out is to use the *ilr* transformation. The *ilr* transformation can then be utilized to obtain a robust estimation of the covariance matrix of the random vector $z = ilr(x) = V^T y$ which V is a $D \times (D-1)$ matrix with orthonormal basis vectors in its columns (i.e. $V^T V = I_{D-1}$) [8]:

$$V = (v_1, \cdots v_D \quad ,\quad v_i = \sqrt{\frac{i}{i+1}} \left(\frac{1}{i}, \cdots \frac{1}{i} \qquad \cdots \right)^T \quad i \qquad \cdots D \tag{1}$$

Using Equation (1), the obtained robust covariance matrix $Cov(z)$ is then back-transformed to the *clr* space by $Cov(y) = VCov(z)V^T$. The resulting robust version of $Cov(y)$ can now be used for the parameter estimation in factor analysis [8]. We applied this procedure with a varimax rotation of the factors in order to achieve a better interpretation of the resulting factors. The number of factors was 6, which resulted in a reasonable percentage of explained variability and lower uniquenesses (as shown in Table 1) demonstrating an acceptable separation of elements in factors. Figure 2 shows the resulting loading plot of the robust factor analysis for the clr transformed data. This Reimann representation of the loadings shows the loading value of the elements on the different factors by the position of the element names in the plot. In addition, the percentages at the top of the plot display the cumulative explained percentage of total variability [29].

It is clear that the 3rd factor consisting of Cu, Pb, Zn, Sn, and Sb is related to the metallogenic properties. They have a strong relationship together as well as high loadings. This strong auto-correlation is also clear in the ternary plot (Figure 3 left), drawn by the CoDaPack software that is a suitable tool to explore the situation of the compositional data in simplex space [31]. Their ilr-transformed components of the 3rd factor also show this linear relationship in euclidean space (Figure 3 right).

Table 1. Uniquenesses of the robust factor analysis model for each element.

Al	As	Ba	Be	Bi	Ca	Cd	Ce	Co	Cr	Cu	Fe	Ga	K	La
0.1	0.515	0.375	0.144	0.149	0.299	0.212	0.227	0.095	0.055	0.01	0.046	0.095	0.159	0.132
Li	**Mg**	**Mn**	**Mo**	**Ni**	**P**	**Pb**	**Rb**	**Sb**	**Sn**	**Ti**	**U**	**V**	**Zn**	**Zr**
0.321	0.364	0.203	0.218	0.201	0.321	0.055	0.101	0.091	0.068	0.034	0.207	0.036	0.033	0.133

All the calculations were done using *robCompositions* package [30].

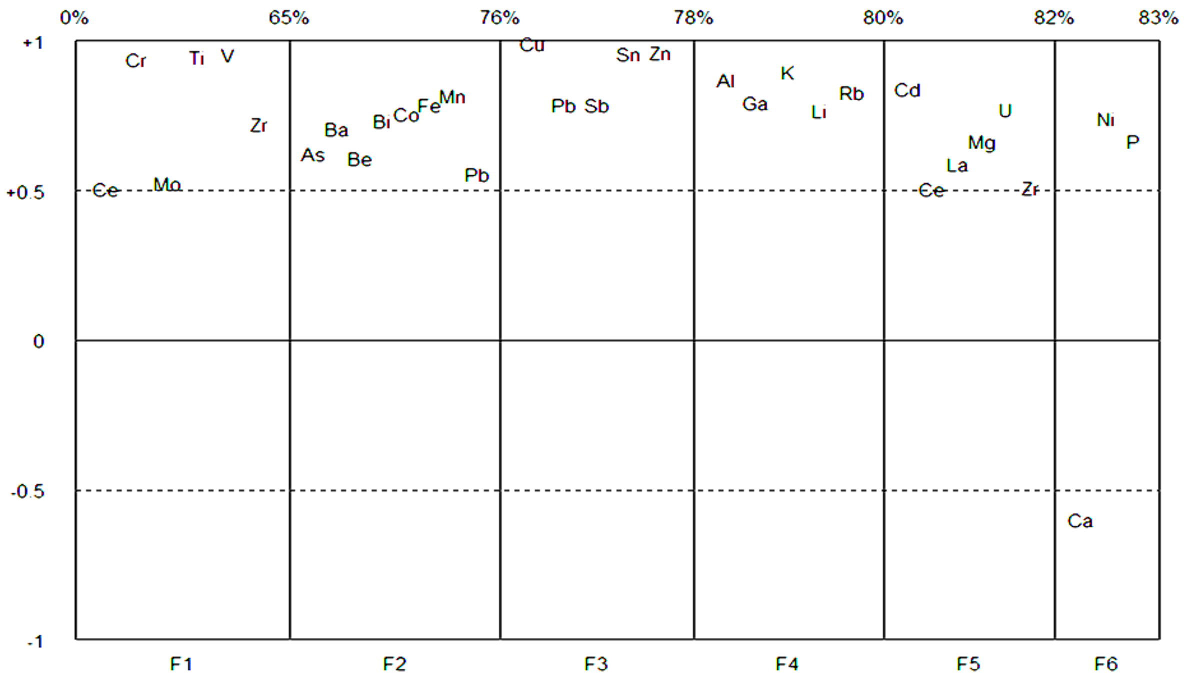

Figure 2. Loading plot of the factors.

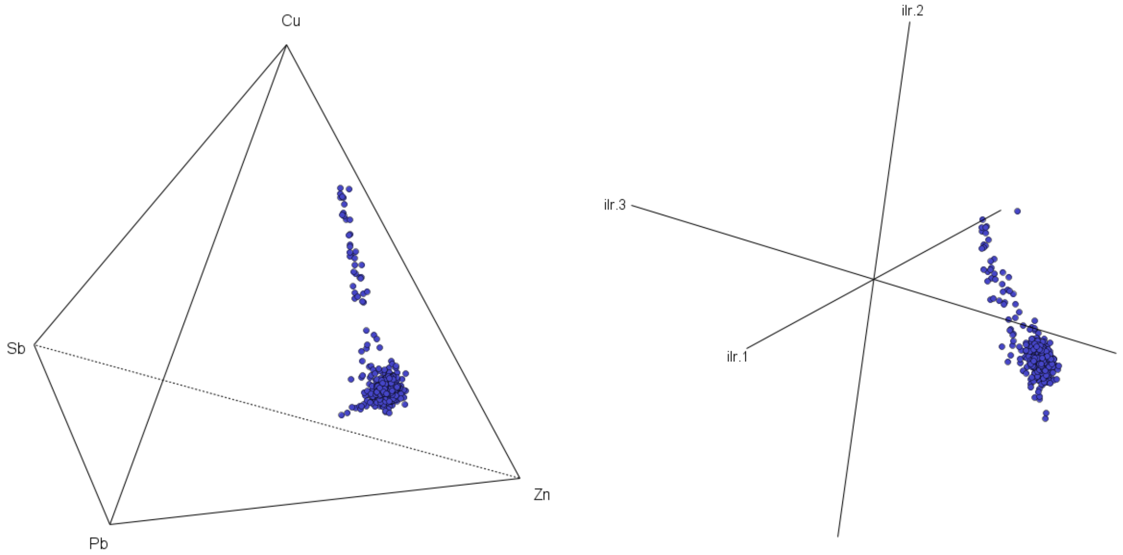

Figure 3. Ternary plot (left) and ilr-plot (right) of the 3rd factor.

3.2. Continuous restricted Boltzmann machine (CRBM)

Product of experts combines many individual experts by multiplying the probabilities together and then renormalizing. Each expert in the model can constrain different dimensions in a high-dimensional space, and their product will then constrain all the dimensions. Product of experts can effectively model the high-dimensional data and produce much sharper distributions than the individual experts or mixture models of the experts [32]. As an extended product of experts, a CRBM [33] can be trained iteratively using the minimizing contrastive divergence [34] and used to model complex high-dimensional continuous data. During an iterative training, the large probability samples have more chance to contribute to the architecture of a CRBM, so the trained model can better encode and reconstruct the large probability training samples. In geochemical exploration, the geochemical background and anomaly samples take, respectively, the large probability and small probability. If the model is trained on all the multivariate geochemical samples in a study area, the trained model will be able to identify the multivariate geochemical anomaly samples from the training geochemical sample population [16]. CRBM is a stochastic neural network, where each unit has some random behaviors when activated. It has one visible layer and one hidden layer with only inter-layer connections [33]. Figure 4 shows a CRBM with two visible, three hidden, and two permanently-on bias units. The visible and hidden units have continuous states generated by adding a zero-mean Gaussian noise to the input of a sampled sigmoid unit, and are connected by weight matrix W [33].

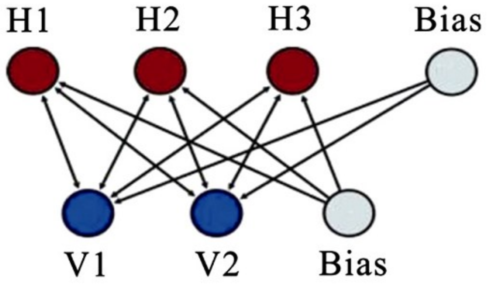

Figure 4. CRBM topology with 3 hidden and 2 visible units [16].

Let v_i and h_j represent the states of visible unit *i* and hidden unit *j*, respectively, and $w_{ij} = w_{ji}$ be the bidirectional weights [16]. Given the states of hidden units, the states of visible units can be expressed by:

$$v_i = \varphi_i\left(\sum_j w_{ij} h_j + \sigma . N_i\left(0.1\right)\right) \tag{2}$$

and given the states of visible units, the state of hidden units can be stated as Equation 3.

$$h_j = \varphi_j\left(\sum_i w_{ij} v_i + \sigma . N_j\left(0.1\right)\right) \tag{3}$$

where function $\varphi(x)$ is a sigmoid function as Equation 4.

$$\varphi(x) = \theta_L + (\theta_H - \theta_L) . \frac{1}{1 + exp(-ax)} \tag{4}$$

with lower and higher asymptotes at θ_L and θ_H, respectively. Parameter *a* is the noise-control parameter that controls the slope of the sigmoid function, and thus the nature and extent of the unit's stochastic behavior [16]. Function $N_i\left(0.1\right)$ represents a Gaussian random variable with zero mean and unit variance that with σ constitute a noise input component according to the following probability distribution [33]:

$$p\left(n_i\right) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-n_i^2}{2\sigma^2}\right) \tag{5}$$

In each training cycle, the model encodes and reconstructs the training samples population that is sequentially presented to it, while the inter-layer connection weights are modified. For each training sample, a one-step reconstruction executes the following: (a) samples are used to set the continuous states of visible units $\{v_i\}$; (b) using Equations (3) and (4), $\{v_i\}$ are transformed into the continuous states of hidden units $\{h_j\}$; (c) using Equations (2) and (4), $\{h_j\}$ are transformed into the one-step reconstructed continuous states of visible units $\{\hat{v_i}\}$; and (d) using Equations (3) and (4), $\{\hat{v_i}\}$ are transformed into the one-step reconstructed continuous states of hidden units $\{\hat{h}_j\}$ [16]. The contrastive divergence update equation for weights is:

$$\Delta\hat{w}_{ij} = \eta_w\left(v_i h_j - \hat{v_i}\hat{h}_j\right) \tag{6}$$

Where $v_i h_j$ is the mean over the training data and η_w is the learning rate for W. The noise-control parameter *a* is also updated while the training is running. Let s_i express v_i for the visible units and h_i for the hidden units. Then parameter a_i is updated as follows [16]:

$$\Delta\hat{a}_i = \frac{\eta_a}{a_i^2}\left(s_i^2 - \hat{s}_i^2\right) \tag{7}$$

In Equation 7, η_a is the learning rate for the noise-control parameter *a*. In each training iteration, the online algorithm feeds the training samples to the model. In CRBM, the visible and hidden units are used for pre-processing the input data and capturing the data structure and probability distribution, respectively. The number of the visible units is equal to the dimension of the input data plus one permanently-on unit but the number of the hidden units depends on the complexity of the input data [16]. CRBM [33] can be viewed as an associative memory that can be trained to encode the complex non-linear non-Gaussian training data distribution [3]. Therefore, it is appropriate for modeling the complex multivariate probability distribution of the training geochemical sample population taken from a complex geological setting. In geochemical exploration, the geochemical anomaly samples take much less probability than the geochemical background samples. Therefore, they can be identified by the trained CRBM. The number of hidden units defines the architecture of the model. If they are too few, the model does not have enough resources to learn the general features of the training sample population. Too many hidden units also increase the training time and can cause over-fitting. A good method to configure this is to start with few hidden units. If the model performs poorly, then the number of hidden units is increased. In case the model's performance does not improve, the network is 'stable' or its architecture reaches optimality. For the multivariate geochemical anomaly identification, two anomaly criteria, average square contribution (ASC) and average square error (ASE), are defined based on the stabilized CRBM. The average square contribution is quite similar to the "novelty signal" defined by [16]. In the stabilized model, the average training signal contribution $\Delta\hat{w}_{ij}$ in Equation (6) is small for well-encoded background samples and large for poorly-encoded anomaly samples [16]. Thus ASC can be stated based on the training signal contribution of an input sample, and a threshold value can be used to recognize the anomaly samples presented to the model. Equation 8 is used to compute ASC [16]:

$$ASC = \frac{1}{pq}\sum_{ij}\left(v_i h_j - \hat{v}_i \hat{h}_j\right)^2 \qquad (8)$$

In Equation 8, p is the number of visible units except for the visible bias unit and q is the number of hidden units except for the hidden bias unit. The reconstructed errors are small for the background samples and large for the anomaly samples. Thus ASE can be defined based on the reconstructed errors of an input sample, and a threshold value can be applied to recognize the anomaly samples presented to the model. The ASE value can be computed as [16]:

$$ASE = \frac{1}{p}\sum_{i=1}^{p}\left(v_i - \hat{v}_i\right)^2 \qquad (9)$$

The threshold values for ASCs and ASEs can be chosen using some statistical methods. If ASC (or ASE) of a sample is less than the threshold value, it belongs to the background; otherwise, the anomaly.

4. Data preparation, process, and results

The dataset used in this study consists of 470 geochemical stream sediment samples analyzed for 36 elements in order to explore the prospected anomalous areas in the regional scale. The compositional descriptive statistics of the data is presented in Table 2. It shows that some elements are barycentered in the 36-dimensional simplex. Radius test of the compositional dataset with 3 statistics (Table 3) confirms the multivariate normality of the simplex. This test is based on the property that, under normality, the Mahalanobis distances from the samples to the mean are chi-squared distributed [35]. Variation in simplex is displayed by three statistic: logratio(lr) variances, clr-variances, and total variance. Lr-variances are not displayed here but the other two important variances show that some elements such as S, Cu, Ca, Ba, Sr, Sn, Cs, Mg, Na, Sn, and Ti have higher clr-variances and a more essential role in the variability of the simplex. This may be due to more presence of them in the lithology of the area and their mobility in the stream sediments. The total variance is relatively low.

In the first step, the dimension of the data was reduced using the robust factor analysis, and as stated earlier, elements of the 3rd factor were selected as the input layer of CRBM. The stabilized CRBM structure with the best performance (MSE = 0.00354) after trying different parameters was: 100 hidden neurons, learning rate 0.5, learning momentum 0.7, noise control parameter for visible units 0.2, noise control parameter for hidden units 0.9, sigmoid activation function, standard deviation of normal noise 0.4, and max iteration 100. The plots of two different thresholds (ASC & ASE) to identify anomalies after training with the whole data and reconstructing it by CRBM are shown in Figure 5. The horizontal axis in both of them represents the indices of the samples. The red lines show the 0.975 quantile of the reconstructed errors as a cut-off threshold. The samples (or indices) that cut the threshold (i.e. are above the threshold) are anomalies.

The high-errors identified in Figure 5 were mapped. They showed the anomalies (Figure 6) that lie mostly on the Red Bed Neogen Conglomerate formation. In Figure 6, The red circles illustrate the mineral potential area for the Cu, Pb, Zn, Sn, and Sb elements. The field observations and the existing mining indices in the study area confirmly proved the results obtained by CRBM. There are good aggreements between the field study and CRBM results.

Table 2. Compositional descriptive statistics of the data.

Variable	Center	Q1	Median	Q3	Clr Variances
Al	0.0701	0.0543	0.0697	0.0389	0.0402
As	0.0001	0.0001	0.0001	0.0001	0.0511
Ba	0.0013	0.0009	0.0012	0.0016	0.1376
Be	0	0	0	0	0.0969
Bi	0	0	0	0	0.698
Ca	0.6111	0.526	0.5985		0.1874
Cd	0	0	0	0.6625	0.0644
Ce	0.0001	0.0001	0.0001	0	0.0479
Co	0.0001	0.0001	0.0001	0.0001	0.033
Cr	0.0002	0.0002	0.0002	0.0001	0.048
Cs	0	0	0	0.0003	0.1549
Cu	0.0002	0.0001	0.0002	0	0.3444
Fe	0.1971	0.1435	0.1896	0.0002	0.0498
Ga	0	0	0	0.2528	0.0292
K	0.0111	0.0083	0.0107	0	0.0529
La	0.0001	0	0.0001	0.0135	0.0587
Li	0.0001	0.0001	0.0001	0.0001	0.0921
Mg	0.0741	0.0575	0.0732	0.0001	0.1466
Mn	0.0046	0.0035	0.0044	0.0865	0.0436
Mo	0	0	0	0.0058	0.0239
Na	0.0092	0.0062	0.0082	0	0.2205
Ni	0.0002	0.0002	0.0002	0.0116	0.0336
P	0.0029	0.0023	0.0029	0.0003	0.026
Pb	0.0001	0.0001	0.0001	0.0034	0.067
Rb	0.0001	0.0001	0.0001	0.0002	0.0863
S	0.0114	0.0047	0.0102	0.0001	1.1596
Sb	0	0	0	0.0235	0.0342
Sn	0	0	0	0	0.3347
Sr	0.0022	0.0017	0.002	0.0025	0.1719
Te	0	0	0	0	0.1202
Ti	0.0026	0.0018	0.0025	0.0036	0.3156
U	0	0	0	0	0.055
V	0.0004	0.0003	0.0003	0.0004	0.1
Y	0.0001	0	0.0001	0.0001	0.0151
Zn	0.0004	0.0003	0.0004	0.0005	0.0711
Zr	0	0	0	0	0.0879
			Total Variance		4.671

Table 3. Radius test of normality with three statistics.

Anderson-Darling		Cramer-von Mises		Watson	
A^{2*}	p	W^{2*}	p	U^{2*}	p
Radius Test					
∞	<0.01	15.7379	<0.01	10.1711	<0.01

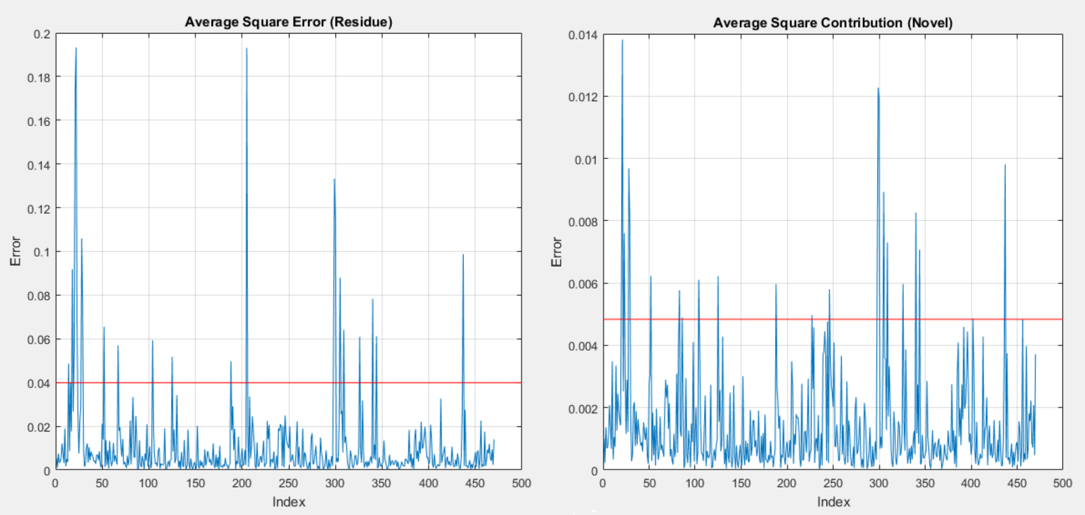

Figure 5. Two anomaly thresholds (ASC & ASE) defined for CRBM reconstructed error.

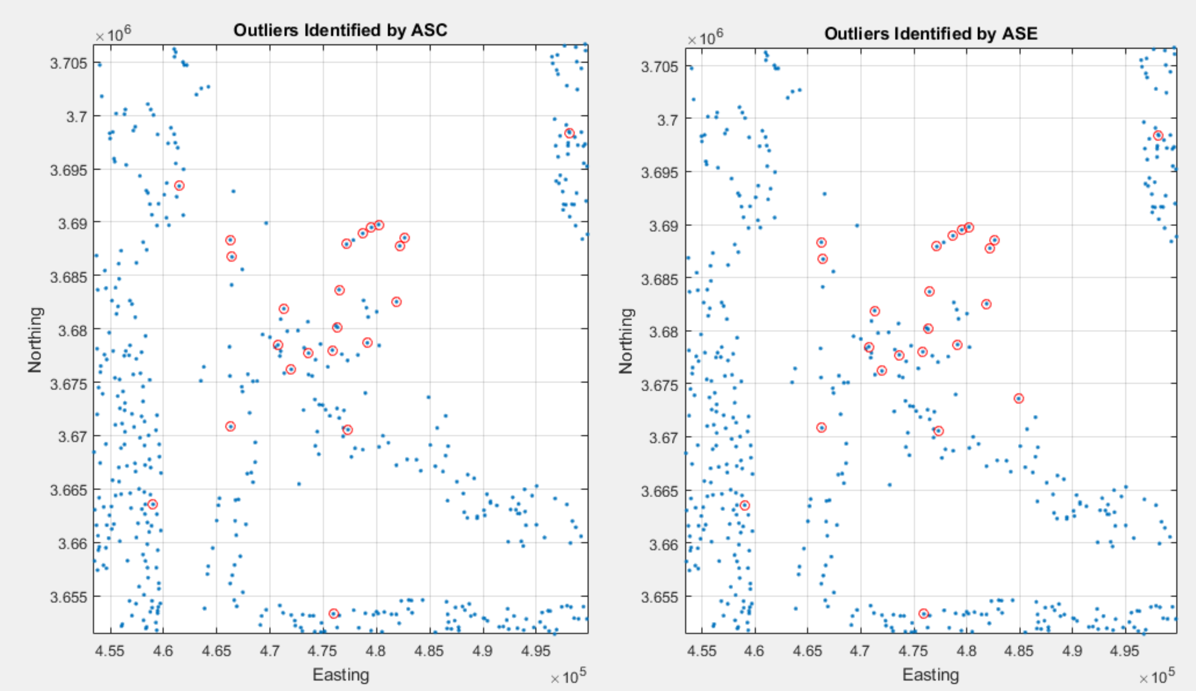

Figure 6. Anomalies detected by two thresholds on CRBM output.

5. Conclusions

In geochemical exploration, anomaly identification is the most important target of data analysis. The more precise and correct the targets are detected, the lower would be the costs of exploration operation in the next phases. Deep belief networks, and more specifically, restricted Boltzmann machines are applied to identify multivariate anomalies. Its specific recursive structure enables the model to recognize the probability distribution of the data. In order to find a meaningful relationship between the input and output data, we selected a paragenetic subset related through a factor. This helps to reduce the complex geological structure in addition to increase the overall process convergence and required memory of the network. The major drawback of a neural network could be setting its parameters. In this case, also the parameters should be configured to get the best performance. The most influencing parameters are the number of hidden neurons, learning rate, noise control parameter for visible units, noise control parameter for hidden units, and standard deviation of normal noise. The most recommended activation function in the literature is sigmoid, which is a rescaled and shifted logistic function and allows the training algorithm to converge faster. The stabilized best performed network resulted in the reconstructed probability distribution of the input data. The difference between them, which is called ‘reconstructed error’, was then mapped and showed anomalous samples of the studied area. These samples are mainly on the red neogene conglomerate, and one is on the Hojedk formation. The field investigations confirmed the results concluded by CRBM.

Acknowledgments

We would like to thank Dr. Yongliang Chen (Chinese academy of Science) for his invaluable recommendations on scripting and precious helps on the subject. Also special thanks go to Industry, Mine & Trade Organization of South Khorasan for providing the required data.

References

[1]. Ghadimi, F., Ghomi, M. and Aref Sedigh, M. (2016). Identification of Ti-anomaly in stream sediment geochemistry using of stepwise factor analysis and multifractal model in Delijan district, Iran. International Journal of Mining & Geo-Engineering. 50: 77-95.

[2]. Carranza, E.J.M. (2008). Geochemical anomaly and mineral prospectivity mapping in GIS (Elsevier). Amsterdam.

[3]. Tang, T.B. and Murray, A.F. (2007). Adaptive sensor modelling and classification using a continuous restricted Boltzmann machine (CRBM). Neurocomputing. 70: 1198-1206.

[4]. Reimann, C., Filzmoser, P. and Garrett, R.G. (2002). Factor analysis applied to regional geochemical data: problems and possibilities. Applied Geochemistry. 17: 185-206.

[5]. Tripathi, V.S. (1979). Factor analysis in geochemical exploration. Journal of Geochemical Exploration. 11: 263-275.

[6]. Yousefi, M., Kamkar-Rouhani, A. and Carranza, E.J.M. (2012). Geochemical mineralization probability index (GMPI): a new approach to generate enhanced stream sediment geochemical evidential map for increasing probability of success in mineral potential mapping. Journal of Geochemical Exploration. 115: 24-35.

[7]. Murray, A.F. (2001). Novelty detection using products of simple experts-a potential architecture for embedded systems. Neural Networks. 14: 1257-1264.

[8]. Filzmoser, P., Hron, K., Reimann, C. and Garrett, R. (2009). Robust factor analysis for compositional data. Computers & Geosciences. 35: 1854-1861.

[9]. Brown, W.M., Gedeon, T., Groves, D. and Barnes, R. (2000). Artificial neural networks: a new method for mineral prospectivity mapping. Australian journal of earth sciences. 47: 757-770.

[10]. Larochelle, H. and Bengio, Y. (2008). Classification using discriminative restricted Boltzmann machines. In Proceedings of the 25th international conference on Machine learning (ACM). pp. 536-543.

[11]. Hinton, G.E. (2007). Learning multiple layers of representation. Trends in cognitive sciences. 11: 428-434.

[12]. Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory (DTIC Document).

[13]. Hinton, G.E. and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science. 313 (5786): 504-507.

[14]. Hinton, G.E. and Salakhutdinov, R.R. (2009). Replicated softmax: an undirected topic model. In Advances in neural information processing systems. pp. 1607-1614.

[15]. Coates, A., Ng, A. and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215-223.

[16]. Chen, Y., Lu, L. and Li, X. (2014). Application of continuous restricted Boltzmann machine to identify multivariate geochemical anomaly. Journal of Geochemical Exploration. 140: 56-63.

[17]. Xiong, Y. and Zuo, R. (2016). Recognition of geochemical anomalies using a deep autoencoder network. Computers & Geosciences. 86: 75-82.

[18]. Ghorbani, M. (2013). The economic geology of Iran: mineral deposits and natural resources (Springer Science & Business Media).

[19]. Aghanabati, A. (1998). Jurassic stratigraphy of Iran (Geological Survey of Iran).

[20]. Aghanabati, A. (2004). Geology of Iran (Geological survey of Iran).

[21]. Ghorbani, M. (2002). Economic geology of Iran (Geological Survey of Iran).

[22]. Shemirani, A. (1988). Cretaceous in Iran.

[23]. Hajian, J. (1996). Paleocene-Eocene deposit in Iran.

[24]. Haddadan, M., Mussavi Harami, S.R. and Ghaemi, F. (1385). Geological Map of Korit.

[25]. Ghorbani, M. and Momenzadeh, M. (1994). Mineralization phases of Iran. (Geological Survey of Iran).

[26]. Ghorbani, M., Tajbakhsh, P. and Khoi, N. (2000). Lead-zinc deposits in Iran (Geological Survey of Iran).

[27]. Pison, G., Rousseeuw, P.J., Filzmoser, P. and Croux, C. (2003). Robust factor analysis. Journal of Multivariate Analysis. 84: 145-172.

[28]. Rousseeuw, P.J. and Driessen, K.V. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics. 41: 212-223.

[29]. Reimann, C., Filzmoser, P., Garrett, R. and Dutter, R. (2011). Statistical data analysis explained: applied environmental statistics with R (John Wiley & Sons).

[30]. Templ, M., Hron, K. and Filzmoser, P. (2011). robCompositions: An R-package for robust statistical analysis of compositional data (na).

[31]. Comas-Cufí, M. and Thió-Henestrosa, S. (2011). CoDaPack 2.0: a stand-alone, multi-platform compositional software. In CoDaWork'11: 4th International Workshop on Compositional Data Analysis, J. J. Egozcue, R. Tolosana-Delgado, and M. I. Ortego, eds. (Sant Feliu de Guíxols).

[32]. Hinton, G.E. (1999). Products of experts. pp. 1-6.

[33]. Chen, H. and Murray, A.F. (2003). Continuous restricted Boltzmann machine with an implementable training algorithm. IEE Proceedings-Vision, Image and Signal Processing. 150: 153-158.

[34]. Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. Neural computation. 14: 1771-1800.

[35]. Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2006). Frequency distributions and natural laws in geochemistry. Geological Society, London, Special Publications. 264: 175-189.

به‌کارگیری ماشین بولتزمن محدود پیوسته برای شناسایی آنومالی‌های چند متغیره از داده‌های ژئوشیمی رسوبات آبراهه‌ای در منطقه کریت، شرق ایران

احمد آریافر[1]* و حمید معینی[2]

۱-گروه معدن، دانشکده مهندسی، دانشگاه بیرجند، ایران
۲-گروه معدن، دانشکده مهندسی معدن و متالورژی، دانشگاه یزد، ایران

ارسال ۲۰۱۷/۵/۲، پذیرش ۲۰۱۷/۷/۵

* نویسنده مسئول مکاتبات: aaryafar@birjand.ac.ir

چکیده:

جدایش آنومالی داده‌های ژئوشیمیایی رسوبات آبراهه‌ای نقش اساسی در اکتشافات ناحیه‌ای دارد. روش‌های مختلف فراوانی برای تفکیک نواحی آنومالی از منطقه مورد مطالعه پیشنهاد شده‌اند. در این تحقیق، از ماشین بولتزمن محدود پیوسته (CRBM) که یک شبکه عصبی مصنوعی تصادفی است، برای تشخیص نواحی معدنی بالقوه برگه زمین‌شناسی ۱:۱۰۰،۰۰۰ کریت در ۱۵ کیلومتری جنوب طبس واقع در استان خراسان جنوبی (شرق ایران) استفاده شده است. به این منظور، ۴۷۰ نمونه ژئوشیمیایی رسوب آبراهه‌ای از منطقه مورد مطالعه برداشت و برای ۳۶ عنصر آنالیز شده است. در نخستین گام، برای نیل به هدف پژوهش، پس از بررسی آمار توصیفی داده‌های ترکیبی (مرکز، واریانس clr و واریانس کلی سیمپلکس)، تحلیل فاکتوری مقاوم برای کاهش بعد داده‌ها و محدود کردن تحلیل چند متغیره به فاکتور اصلی کانی‌سازی انجام شد. سپس فاکتور سوم (از ۶ فاکتور) شامل Cu، Pb، Zn، Sn و Sb که مرتبط با ویژگی‌های متالوژنی منطقه بود، به عنوان ورودی ماشین بولتزمن CRBM در نظر گرفته شد. آنگاه پس از امتحان پارامترهای مختلف، ساختار نهایی یک CRBM با بهترین کارایی به دست آمد. مقادیر خطا یا آنومالی‌های دو حد آستانه (ASC و ASE) پس از آموزش شبکه روی کل داده‌ها و بازسازی آن‌ها، بر اساس یک معیار تفکیک آماری استخراج شد. در ادامه آنومالی‌های به دست آمده در دو نقشه ترسیم شد که نقاط امید بخش در آن نشان داده شده است. مطالعات صحرایی و اندیس‌های معدنی موجود، نتایج به دست آمده با CRBM را تائید کرد.

کلمات کلیدی: رسوب آبراهه‌ای، CRBM، تحلیل فاکتوری مقاوم، کریت.