Journal of Geochemical Exploration 180 (2017) 15–23

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Comparing compositional multivariate outliers with autoencoder networks in anomaly detection at Hamich exploration area, east of Iran

Hamid Moeini, Farhad Mohammad Torab*

Department of Mining and Metallurgical Engineering, Yazd University, Yazd, Iran

ARTICLE INFO

Keywords:
Lithogeochemical exploration
Autoencoder
Compositional data
Hamich
CRBM

ABSTRACT

Newly presented machine learning methods based on Deep Belief Networks like autoencoders have opened a new window on anomaly identification in different fields of the science. They reconstruct the normal probability distribution pattern of the input data using stacks of Continuous Restricted Boltzmann Machines (CRBM) and thus determining the outliers. Therefore using this machine on geochemical samples taken in regional exploration scale, might be an acceptable way to delineate the multivariate anomalies and propose the next targets for detailed exploration. On the other hand, due to compositional nature of geochemical data, compositional data analysis (CoDa) has been developed to identify multivariate outliers or anomalies in recent years.

A comparison between both methods has been made applying them on lithogeochemical samples of Hamich area in Southern Khorasan, East of Iran. The area was explored in details some years ago and veinlets of galena-sphalerite-pyrite at depth, based on the outcrops of Cu-Pb, were verified by additional core drillings. We used its final report to validate the results of both methods. They showed that the two completely different methods could get the same acceptable targets. However the CoDa approach needs less parameters and shows which elements are responsible for the anomalies.

1. Introduction

Anomaly recognition is one of the main goals in regional geochemical exploration and the first step in making decisions for locating the next stage targets. Therefore to reduce the risk and uncertainties of exploration and costs of drilling, it demands applying precise analysis methods. The multivariate methods usually require multivariate geochemical data satisfying a known statistical distribution, such as a multivariate normal distribution (Xiong and Zuo, 2016).

Deep learning methods aim at learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower level features. Depth of architecture refers to the number of levels of composition of non-linear operations in the function learned. In 2006 Hinton et al. at University of Toronto introduced Deep Belief Networks (DBNs) (Hinton et al., 2006), with a learning algorithm that greedily trains one layer at a time, exploiting an unsupervised learning algorithm for each layer, a Restricted Boltzmann Machine (RBM) (Freund and Haussler, 1994). Shortly after, related algorithms based on auto-encoders were proposed (Bengio et al., 2007), apparently exploiting the same principle: guiding the training of intermediate levels of representation using unsupervised learning, which can be performed locally at each level (Bengio, 2009).

Since 2006, deep networks have been applied with success not only in classification tasks (Ahmed et al., 2008; Bengio et al., 2007; Boureau et al., 2008; Larochelle et al., 2007; Lee et al., 2009; Poultney et al., 2006; Vincent et al., 2008), but also in regression (Hinton and Salakhutdinov, 2008), dimensionality reduction (Hinton and Salakhutdinov, 2006; Salakhutdinov and Hinton, 2007) and geochemical anomaly recognition (Xiong and Zuo, 2016). Although auto-encoders, RBMs and DBNs can be trained with unlabeled data, in many of the above applications, they have been successfully used to initialize deep supervised feedforward neural networks applied to a specific task (Bengio, 2009).

On the other hand, in the convenient geochemical researches, datasets of samples taken of an environment (litho, sediment, soil, gas or water) and analyzed for constituent elements are often compositional data which have long been of concern in the geochemical field (Aitchison, 1982; Buccianti, 2011; Buccianti et al., 2006; Carranza, 2011; Chayes and Trochimczyk, 1978; Rollinson, 1992). It is an example of a closed number system because it contains compositional variables that are parts of a whole (Carranza, 2011). The statistical analysis of compositional multivariate data is a much discussed topic in the field of multivariate statistics (Filzmoser et al., 2005). In practice, log-ratio transformations are commonly employed in geochemical data

* Corresponding author.
E-mail address: fmtorab@yazd.ac.ir (F.M. Torab).

http://dx.doi.org/10.1016/j.gexplo.2017.05.008
Received 19 January 2017; Received in revised form 27 May 2017; Accepted 28 May 2017
Available online 30 May 2017
0375-6742/

processing to open closed systems for better understanding of realistic relationships among compositions (Filzmoser et al., 2012; Gallo and Buccianti, 2013).

In present case study to identify exploratory target areas in Hamich area located in east of Iran, we compared two completely different approaches in multivariate outlier detection. The first is the compositional method of Filzmoser (Filzmoser et al., 2012) and the second is using autoencoder networks which are based on Boltzmann machines structures. They are a kind of Deep Belief Networks that have been recently developed and are well known for anomaly detection applications. Although we used both ilr-transformed and raw data as input of the network, the results were almost the same. It should be noted that the compositional approach is much more interpretable since the anomalous variables causing the multivariate anomaly also could be detected.

2. Methodology

Among the regional geochemical data processing, there are a variety of statistical and data mining approaches as well as different mapping techniques which serve as presentations of the outputs (Fletcher, 2013). They include convenient methods such as statistical distribution thresholds of a gaussian distribution tails or extremes (Reimann et al., 2005). Extreme values are of interest in investigations where data are gathered under controlled conditions. In contrast, geochemists are typically interested in outliers as indicators of rare geochemical processes (Filzmoser et al., 2005).

2.1. Deep autoencoders

Deep Learning is used for unsupervised feature learning or more specifically, nonlinear dimensionality reduction. Deep Belief Network (DBN) is a probabilistic generative model composed of stacked Restricted Boltzmann Machines (RBMs). A DBN can recognize high level features of the inputs with the RBMs using a greedy layer-wise unsupervised training algorithm. The deep autoencoder network based on DBN is trained by minimizing the difference between the input and the output data. Hinton and Salakhutdinov suggested the use of RBMs for deep autoencoder networks with binary inputs and outputs (Hinton and Salakhutdinov, 2006). Furthermore, it is possible to use continuous RBMs (CRBMs), rather than RBMs as the unsupervised building block of the autoencoder network (Xiong and Zuo, 2016).

Stacks of CRBMs are in fact multi-layer feedforward neural networks without any inter-layer connections consisting many layers of interconnected neuron units (as shown in Fig. 1), starting with an input layer to match the feature space, followed by multiple hidden layers of nonlinearity, and ending with reverse copies of encoder layers to match the output space in a decoding structure. The inputs and outputs of the model's units follow the basic logic of the single real neuron (Candel et al., 2016).

Bias units are included in each non-output layer of the network. The weights linking neurons and biases with other neurons fully determine the output of the entire network. Learning occurs when these weights are adapted to minimize the error on the labeled training data. More specifically, for each training example j, the objective is to minimize a loss function, $L(W,B|j)$.

Here, W is the collection $\{W_i\}_{1:N-1}$, where W_i denotes the weight matrix connecting layers i and $i+1$ for a network of N layers. Similarly B is the collection $\{b_i\}_{1:N-1}$, where b_i denotes the column vector of biases for layer $i+1$. This basic framework of multi-layer neural networks can be used to accomplish Deep Learning tasks. Deep Learning architectures are models of hierarchical feature extraction, typically involving multiple levels of nonlinearity (Candel et al., 2016).

If the input data is treated as labeled with the same input values, then the network is forced to learn the identity via a nonlinear, reduced representation of the original data. The most convenient activation function f is a sigmoid function or *tanh* defined as Eq. (1) (Candel et al., 2016).

$$f(\alpha) = \frac{e^{\alpha}-e^{-\alpha}}{e^{\alpha}+e^{-\alpha}} \quad with \quad f(\cdot) \in [-1, 1] \quad and \quad \alpha = \sum_i w_i x_i + b \tag{1}$$

x_i and w_i represent the firing neuron's input values and their weights, respectively; α denotes the weighted combination. In fact the *tanh* function is a rescaled and shifted logistic function; its symmetry around 0 allows the training algorithm to converge faster (Candel et al., 2016).

H2O's Deep Learning framework interface in R supports regularization techniques to prevent overfitting. ℓ_1(L1: Lasso) is a regularization method that constrains the absolute value of the weights and has the net effect of dropping some weights (setting them to zero) from a model to reduce complexity and avoid overfitting. H2O's Deep Learning preprocesses the data to standardize it for compatibility with the activation functions. Since the activation function generally does not map into the full spectrum of real numbers, $\mathbb{R}$, the algorithm first standardize the data to be drawn from $\mathcal{N}(0, 1)$. Although for autoencoding, the data is normalized (instead of standardized) to the compact interval of $\mathcal{U}(-0.5, 0.5)$ to allow bounded activation functions like *tanh* to better reconstruct the data. The stopping rules are convergence-based or time-based that can be set in training the model. There is no general rule for setting the number of hidden layers, their sizes or the number of epochs. Experimenting by building Deep Learning models using different network topologies and different datasets will lead to insights about these parameters (Candel et al., 2016).

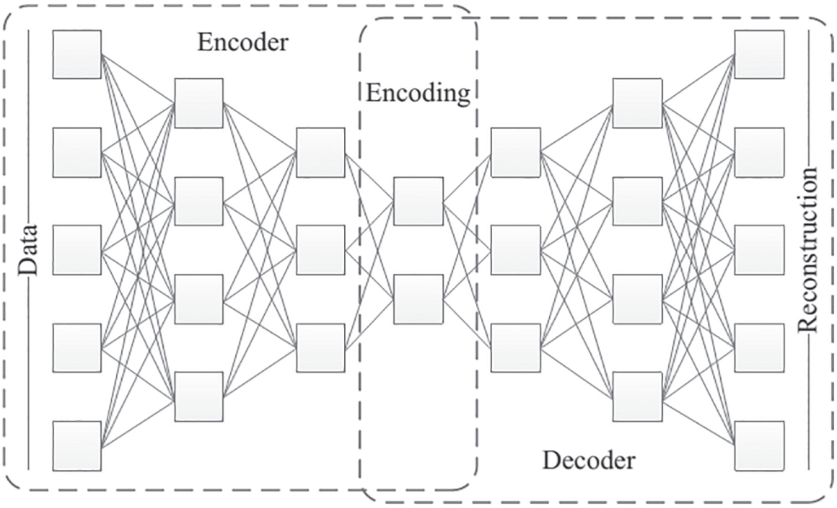

Fig. 1. A simplified architecture of an autoencoder network (Xiong and Zuo, 2016).

In other words an autoencoder network is designed to minimize the difference between the input and output. The main difference between an autoencoder network and a traditional network is the size of the output layer. The size of an autoencoder's output layer is always the same as the input layer. An autoencoder network is composed of an encoder and a decoder. The encoder network transforms the input data into new code, and the decoder network recovers the data from the code. Then if an anomalous test point does not match the learned pattern, the autoencoder will likely have a high error rate in reconstructing this data, indicating anomalous data. This framework is used to develop an anomaly detection demonstration using a deep autoencoder (Xiong and Zuo, 2016).

2.2. Compositional data multivariate outlier detection

The geochemical data have an intrinsic compositional property. They are multivariate observations that describe quantitatively the parts of some whole. Thus, their components carry exclusively relative information about the parts (Aitchison, 1982). Typically these observations are expressed as data with a constant sum constraint such as proportions, percentages, or mg/kg, i.e:

$$X = (x_1,...,x_D)', \quad x_i > 0, \quad \sum_{i=1}^{D} x_i = constant \tag{2}$$

The D-dimensional space defined in Eq. (2) is called a *Simplex*. Due to the fact that geochemical data follow the simplex geometry, classical statistical methods that rely mostly on the euclidean geometry produce spurious results when they are applied to raw compositional data (Filzmoser and Hron, 2008; Filzmoser et al., 2009a). Whether or not the data follow a normal distribution is of no importance at all (Filzmoser and Gschwandtner, 2012). In order to work with these data there has been proposed three log-ratio transformations that open the data and define some new coordinates in euclidean vector space. They are named as: *clr: centered logratio*, *ilr: isometric logratio* and *alr: additive logratio*.

Due to the definition of compositional data, all the relevant information about x_1 is contained in the ratios to each of the remaining parts $x_2, ...,x_D$. Accordingly, this relative information for all remaining parts needs to be considered also for univariate data analysis (Filzmoser et al., 2009b). The ilr transformation of the elements of the simplex allows more effective handling of compositional data and preserves all metric properties so that it can be analyzed as a standard multivariate dataset in real space. Therefore the ilr variable:

$$z_1 = \sqrt{\frac{D-1}{D}}\, ln \frac{x_1}{\sqrt[D-1]{\prod_{j=2}^{D} x_j}} \tag{3}$$

contains all the relative information between x_1 and $x_2, ...,x_D$, because none of $z_2, ...,z_{D-1}$ includes x_1. In this way, each compositional part can be expressed by a single ilr variable as defined in Eq. (3) which here is used for univariate analysis of global outliers (Filzmoser et al., 2009b).

For all logratio transformations, the problem of missing or zero values should be solved prior to any analysis. Fortunately there have been proposed many solutions and tools to impute them in the best way. One of them is *zCompositions* package written and developed by Javier Palarea-Albaladejo and Josep Antoni Martín-Fernández that provides principled methods to deal with zeros and nondetects in compositional datasets (Palarea-Albaladejo and Martín-Fernández, 2015).

In contrast to univariate outliers, multivariate outliers are not necessarily extreme along single coordinates. Rather, they could deviate from the multivariate data structure formed by the majority of observations (Filzmoser et al., 2009b). The estimated covariance structure is used to assign a distance to each observation indicating how far the observation is from the center of the data cloud with respect to the covariance structure (Filzmoser et al., 2012). This distance measure is the well-known Mahalanobis distance, defined for a sample $x_1, ...,x_n$ of n observations in the d-dimensional real space R^D as

$$MD(x_i) = [(x_i - T)'C^{-1}(x_i - T)]^{1/2} \quad \text{for i} = 1, ...,n \tag{4}$$

T and C in Eq. (4) are location and spread estimators, respectively. In the case of multivariate normally distributed data, the arithmetic mean and the sample covariance matrix are the best choices, leading to the best statistical efficiency (Filzmoser et al., 2012). In this case, the squared Mahalanobis distances approximate a chi-square distribution χ^2_d with d degrees of freedom. A certain cut-off value like the 97.5% quantile of χ^2_d can be taken as an indication of extremeness: data points with higher (squared) Mahalanobis distance than the cut-off value are considered as potential outliers (Rosseeuw and Van Zomeren, 1990).

Both the arithmetic mean and the sample covariance matrix are highly sensitive to outlying observations. A number of robust estimators of covariance have been proposed in the literature, like the MCD[1] estimator (Maronna et al., 2006). It looks for a subset h out of n observations with the smallest determinant of their sample covariance matrix. The subset size h can vary between half the sample size and n, and it will determine not only the robustness of the estimates, but also their efficiency. It is common to use the same cut-off value from the χ^2_d distribution (Rosseeuw and Van Zomeren, 1990). Filzmoser introduced a more advanced approach to the cut-off value that could lead to more accurate values. This method called “adaptive outlier detection”, accounts for the actual numbers of observations and variables in the dataset, and it tries to distinguish among extremes of the data distribution and outliers coming from a different distribution (Filzmoser et al., 2005).

Several methods have been proposed for the identification of multivariate outliers, making use of robust statistics (Maronna et al., 2006). Such tools have also been developed in the context of compositional data (Filzmoser and Hron, 2008). The package *mvoutlier* in **R** software presented by Filzmoser and Gschwandtner is a free and precious tool to deal with outliers especially in compositional data. It is designed on a comprehensive mathematical basis that detects, maps and plots the global outliers in a multivariate dataset. This is done by computing for each observation the robust squared Mahalanobis distances to the medians along the single ilr variables. The distances are computed and split by four values: the quantiles 0.25, 0.5, 0.75 and the aforementioned adaptive outlier cutoff, i.e., an outlier with a high value means that most univariate parts have higher values than the average. The maps used in this study are marked by only the values above median that is coloured light green. This characterization helps to interpret multivariate outliers (Filzmoser et al., 2012).

3. Study area

The sampled area (Fig. 2) is situated within the eastern part of the so-called Lut block of eastern Iran. Eastern Iran and particularly the Lut block, has a great potential for different types of mineralization as a result of its past subduction zone tectonic setting, which lead to extensive magmatic activities forming igneous rocks of different geochemical compositions. The Lut block is characterized by extensive exposure tertiary volcanic and subvolcanic rocks formed due to the subduction prior to the collision of the Arabian and Asian plates (Camp and Griffis, 1982; Tirrul et al., 1983). Most of the study area is covered by upper Eocene-Oligocene altered volcanic rocks including andesite, dacite, tuff and ignimbrite. These rocks are intruded by felsic to intermediate intrusive porphyritic rocks consisting of monzonite, diorite and microgranodiorite porphyry stocks. Sedimentary rocks in this area consist of conglomerates, minor middle Eocene to upper Eocene tuffaceous marls in the southeastern to eastern area and Quaternary sediments. The prospect area is similar to low-sulfidation epithermal

[1] Minimum Covariance Determinant.

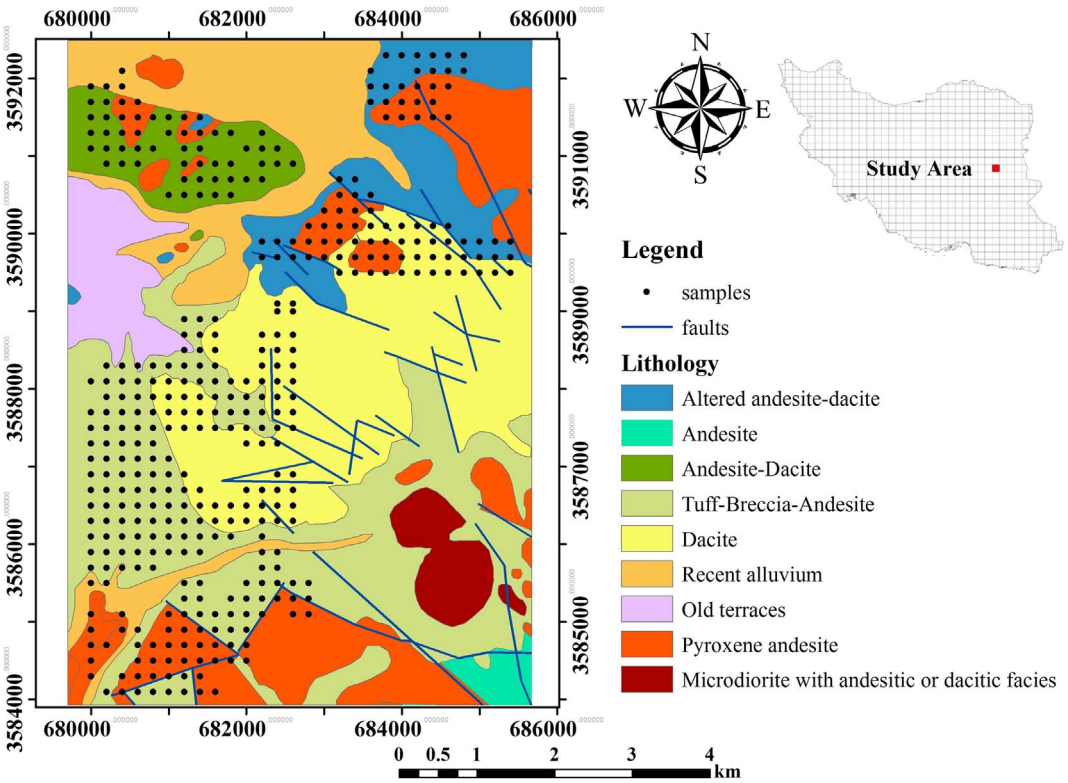

Fig. 2. Lithological map of the study area and litho-geochemical sampling locations.

systems. The rocks are dominantly altered andesite and dacite. Argillization, sericitization and silicification are common hydrothermal alterations in this area. Mineralization is not seen at surface (Karimpour and Mazaheri, 2009).

The area comprises the moderately folded Tertiary volcanic zone of Kuh-e-Shah in the north, a zone of strongly tectonized and partly Eocene andesites and dacite-andesites. It is a zone of gently warped and tilted Upper Tertiary andesitic formations belonging to the Lut block. The tuff in the center of the area grades laterally and upwards into thick volcanic breccia and locally, conglomerate with pebbles of alveolina and nummulitic limestones of Paleocene-early Eocene age. These clastic rocks are overlain by widespread dacites and dacite tuffs which because of the gradational contact relations with the underlying beds are thought to be also of Paleocene age. The rocks tentatively attributed to Neogene are mainly various types of andesite. In the present area the andesites seem to be genetically related to a characteristic formation of microdiorites which protrude through the Paleogene volcanics (Vassighi et al., 1975). The volcanic breccias underlying the dacites extend eastward where they are again associated with dacitic rocks, but also with abundant, strongly altered andesitic material. The Neogene andesites have been divided into several petrographic varieties, which seem to belong to different extrusive centers and also differ slightly in age. Most widespread are pyroxene andesites. Some ancient workings for copper are found in the south of this area, apparently related to the aplitic intrusions in the area. Traces of malachite and chalcopyrite occur in several small ancient workings south and southeast of Hamich in dacitic and andesitic volcanic rocks of the Paleogene. Minor lead-zinc mineralisations with galena, cerussite and smithsonite are found in dacite and pyroxene andesite units south of this area (Vassighi et al., 1975).

4. Data process and results

The dataset used in this research is taken from a part of an exploration project carried out in southwestern of Birjand, South Khorasan by industry, mine and trade (IMT) organization. The data consists of 396 lithogeochemical samples and 20 duplicate samples analyzed for 44 trace elements with ICP-MS method in Australia's Amdel lab. Maximum accepted analytical error $\left(e,=,\frac{2}{n},\sum,\frac{|x_i-y_i|}{x_i+y_i}\right)$ for the analysis was 20%. According to Table 1, Au, Cr and Ni with about 54%, 33% and 36% of error respectively, had low to moderate accuracy probably due to their behaviour in the dispersion environment. Other elements were in acceptable ranges (Consulting Engineers Co., 2009). The study area covers a 40 km^2 rectangle with Hamich, a village and the only populated area, located at the west. Geographically it is an arid region with an almost hill and creek topography. The samples locations are shown in Fig. 2.

Out of all the variables, Ag, B, Bi, Cd, Hg, Sn, Te were left out because they had more than about 60% of missing values. 37 remained variables were used in imputation process. 53 samples of Au had zero values and 1 of As, 6 of Co, 185 of Cr, 20 of Mo, 28 of Sb, 33 of Tl, 15 of W were below detection limit (BDL). These missing values were replaced using recent technique of ilr-Em imputation in *zCompositions* package in R (Palarea-Albaladejo and Martín-Fernández, 2015).

Then a subcomposition of variables was selected based on their geological relations and paragenetical properties. It included 12 elements of Au, As, Cu, Mo, Fe, Mg, Pb, Zn, Ni, Co, Cr, W. First using *mvoutlier* package in R (Filzmoser and Gschwandtner, 2012), the matrix of the selected variables was ilr-transformed and univariate outliers were determined and plotted (Fig. 3). The global multivariate outliers were identified and mapped (Fig. 4) too. It shows the outliers that are out of the 0.975 quantile of the multivariate chi-square distributed Mahalanobis distance (MD) based on minimum covariance determinant (MCD) estimator. Comparing Figs. 3 and 4 gives us some clue that indices 1 to 12 and 73 in the southern part show anomalies of Cu, Fe, Pb, Zn, Au and indices 28, 32, 34, 39, 42, 43 show anomalies of Ni, Co, Cr, Pb, Zn. Index 56 shows a strong anomalies of As and Ni.

At the next step, using *H2O R interface*,[2] different structures for autoencoder network with different parameters were designed and tried. In order to consider the closed nature of geochemical data, two

Table 1
Analyzed elements and measurement errors (Consulting Engineers Co., 2009).

Element	Au	Al	Ca	Co	Cr	Cu	Fe	K	Mg	Mn	Na	Ni	P	S
Detection limit (ppm)	1(ppb)	10(%)	10(%)	0.2	2	0.2	100(%)	10	10(%)	2	10(%)	2	5	50
Error percentage (%)	54.79	3.17	2.23	13.7	33.02	8.11	3.7	3.81	2.78	3.96	5.08	36.36	4.5	9.47
Below 10 DL (%)	58.2	3.2	2.23	13.7	32.2	8.11	3.7	3.81	2.78	3.96	4.53	38.1	4.5	9.47
Up 10 DL (%)	24.0	–	–	–	48.6	–	–	–	–	–	4.4	3.9	–	–
Number of below 10 DL	18	20	20	20	19	20	20	20	20	20	12	19	20	20
Number of up 10 DL	2	–	–	–	1	–	–	–	–	–	8	1	–	–
Element	Ti	V	Zn	Ag	As	Ba	Be	Bi	Cd	Ce	Cs	La	Li	Mo
Detection limit (ppm)	10(%)	0.1	0.2	0.01	0.5	0.2	0.2	0.1	0.1	0.5	0.1	10	0.5	0.1
Error percentage (%)	4.36	3.94	5.52	4.76	8.58	6.39	9.65	0.77	0	5.65	4.69	5.24	7.19	23.00
Below 10 DL (%)	4.36	3.94	5.52	4.76	9.12	6.58	9.65	0.77	0	5.65	4.69	5.24	7.19	24.2
Up 10 DL (%)	–	–	–	–	5.51	4.6	–	–	–	–	–	–	–	0.0
Number of below 10 DL	20	20	20	20	17	18	20	20	20	20	20	20	20	19
Number of up 10 DL	–	–	–	–	3	2	–	–	–	–	–	–	–	1
Element	Nb	Pb	Rb	Sb	Sc	Sn	Sr	Te	Th	Tl	U	W	Y	Zr
Detection limit (ppm)	0.5	0.2	0.1	0.1	1	0.2	0.1	0.2	0.02	0.1	0.02	0.1	0.05	5
Error percentage (%)	6.06	24.05	5.8	8.88	4.98	2.86	4.55	6.76	5.89	3.43	4.09	7.43	5.99	9.82
Below 10 DL (%)	6.06	25.2	5.8	18.1	4.98	2.86	4.55	6.76	5.89	3.43	4.09	7.43	5.99	9.82
Up 10 DL (%)	–	3.11	–	4.93	–	–	–	–	–	–	–	–	–	–
Number of below 10 DL	20	19	20	6	20	20	20	20	20	20	20	20	20	20
Number of up 10 DL	–	1	–	14	–	–	–	–	–	–	–	–	–	–

groups of data were prepared for the network as input:(1) raw data without any transformations, and (2) ilr-transformed data using Eq. (3). For each group, the performances of different autoencoder models were compared after training on all the 396 samples and the best model was selected for anomaly detection (Table 2).

In geochemical exploration, the multivariate geochemical background and anomaly values are separated by the samples with large and small probability respectively. 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 (Chen et al., 2014).

The parameters of the model are: activation function = "Sigmoid", number of hidden units in 3 layers = (100,80,60,80,100), L1 = 1e−3, number of iterations over the training dataset = 500. The rest of the parameters were left to defaults. A 90th percentile was selected as the threshold to indicate the outliers of errors. The location of multivariate outliers as geochemical anomalies are shown in Fig. 5.

5. Discussion

Geochemical data analysis in a regional scale leads to definition of anomaly locations or detailed exploration targets. A detailed exploration project was planned and conducted by IMTO of South Khorasan. After remote sensing the potential alteration zones in a large scale enfolding this area, lithogeochemical sampling and detailed geological mapping were done. However there existed some mineralized outcrops of veins and veinlets (mostly pyrite-quartz-malachite with silicic alteration and sphalerite-galena) and ancient diggings in eastern and southern parts of the area. After geochemical analysis, the altered dacite-andesite units were recognized as a potential zone and 9 points in the west-southern part of the area were marked for core drillings.

In brief, the most significant boreholes were BH1, BH3 and BH6 with 20, 125.5 and 158 m depths respectively. The other boreholes were near these and showed almost the same results.

BH1 mineralized zone was from surface to about 5 m in a fault zone with quartz veins and Fe-oxides in altered dacite-andesite. BH3 was drilled about 20 m far from hanging wall of a galena-sphalerite-pyrite vein outcrop. The vein direction was the same as strike-slip faults. The mineralized zones were in depths of 45–55 m and 75 m with remarkable values of Au in quartz veinlets and Pb-Zn in sulfides. Their host rock was latite-monzonite. BH6 cores were mostly tectonized pyrite with quartz veinlets. Au in all samples of this core was below detection limit of the lab. The host rocks through the core were altered sub-volcanics, andesite-dacites. The high value of Zn in analyzed samples was considerable. However the report concluded that after combining all the results of the cores, geophysics, geochemistry, geology and mineralography, there seemed to be no evidence of an accountable mineral reserve in the area and there exists just some minor veinlets of galena and sphalerite that have been formed along the faults.

Locations of the major proposed boreholes for identified anomalies and the mineralized Pb-Cu indexes of ancient minings in this area are shown in Fig. 6. The blue circles and squares are anomalies detected by autoencoder network on raw data and its ilr-transformed and hollow circles are anomalies detected by compositional data multivariate outlier detection. In western and southwestern parts, the detected anomalies are the same. Comparing the anomalies with the outcrops near boreholes and mineral indexes suggests the acceptable accuracy of their detection.

6. Conclusion

Anomaly detection is one of the first and most important stages in mineral exploration. Usually it is done in regional scale so as the prospected local points for drillings or further detailed exploration are determined. In this paper, the autoencoder network together with compositional data multivariate outlier detection were applied on an explored area to compare them as recognition tools. Although they are completely different in methodology, their aim is the same.

In this study, the parameters of the network defined within H2O-R interface, such as the number of iterations, the size of each hidden layer were changed and activation function, adaptive learning, standardization rule, regularization techniques (to prevent overfitting) remained as default. The reconstruction error of the optimal network with the least MSE, were used as a useful indicator of multivariate geochemical anomalies. Their corresponding samples were mapped.

The results from the autoencoder network performed on raw and ilr-transformed data were compared with the results from multivariate analysis of compositional data. In the latter method after defining a distance criteria in simplex, a robust identification is performed using an adaptive threshold and multivariate global outliers are detected and mapped.

[2] http://h2o.ai/.

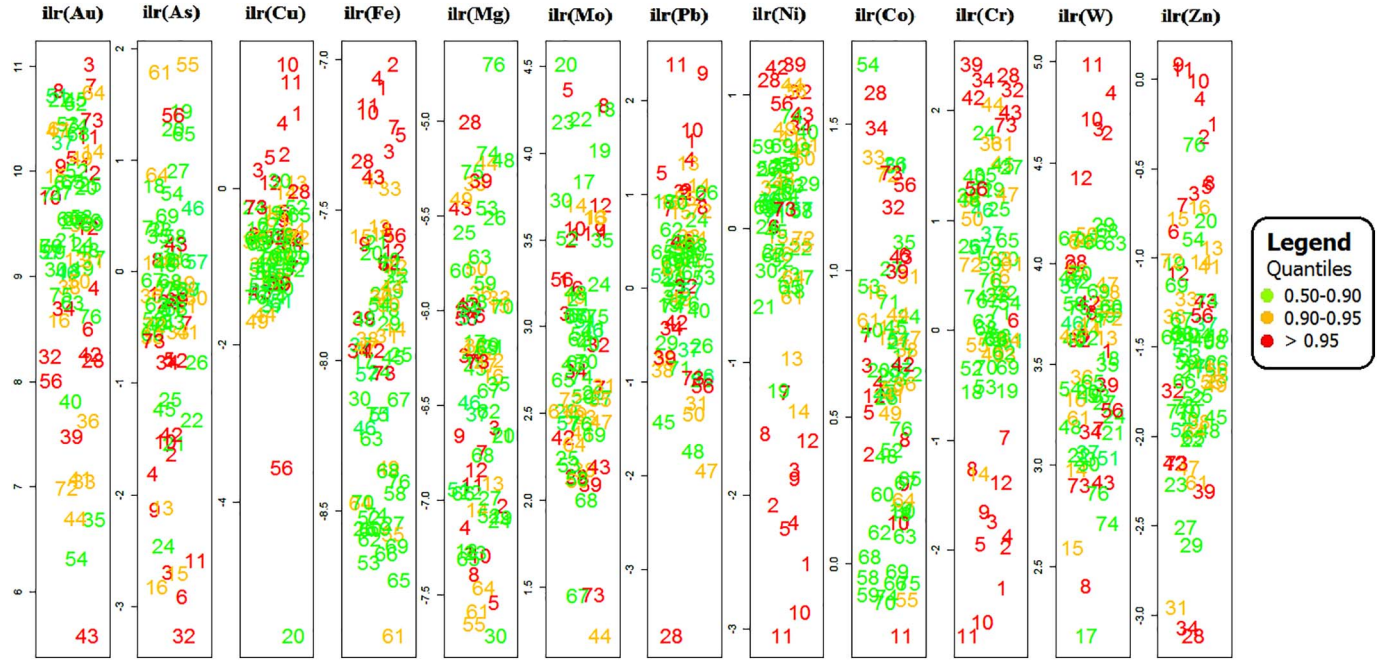

Fig. 3. Univariate ilr-transformed data showing outliers.

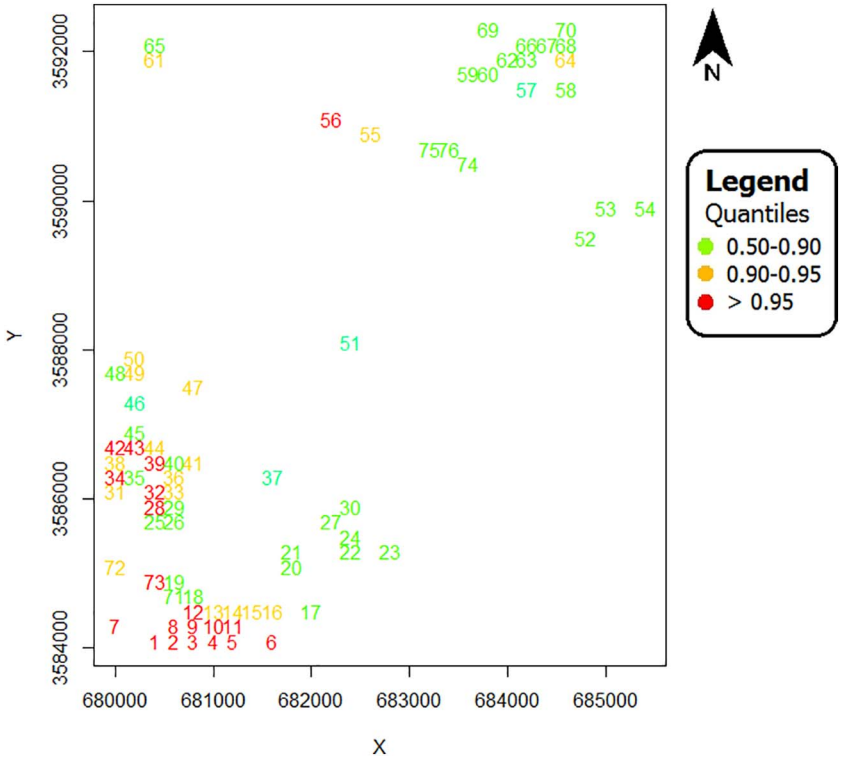

Fig. 4. Global outliers map of the study area indicating the locations of the outlying samples.

Table 2
Performances of the network.

	MSE	RMSE
Raw data	0.008855246	0.09410232
ilr-Transformed data	0.02741058	0.1655614

The spatial distribution of the geochemical anomalies obtained by both methods was similar in the study area. Although compositional data analysis method has slight preferences over autoencoder as it can show univariately which elements are responsible for the anomaly too. Performance of the autoencoder model on ilr-transformed data was lower and showed more noise after decoding than on raw data. This led to wider anomaly regions and lower accuracy. However from practical point of view, compositional data outlier detection needs less parameters than autoencoder and is closer to the nature of the geochemical simplex space.

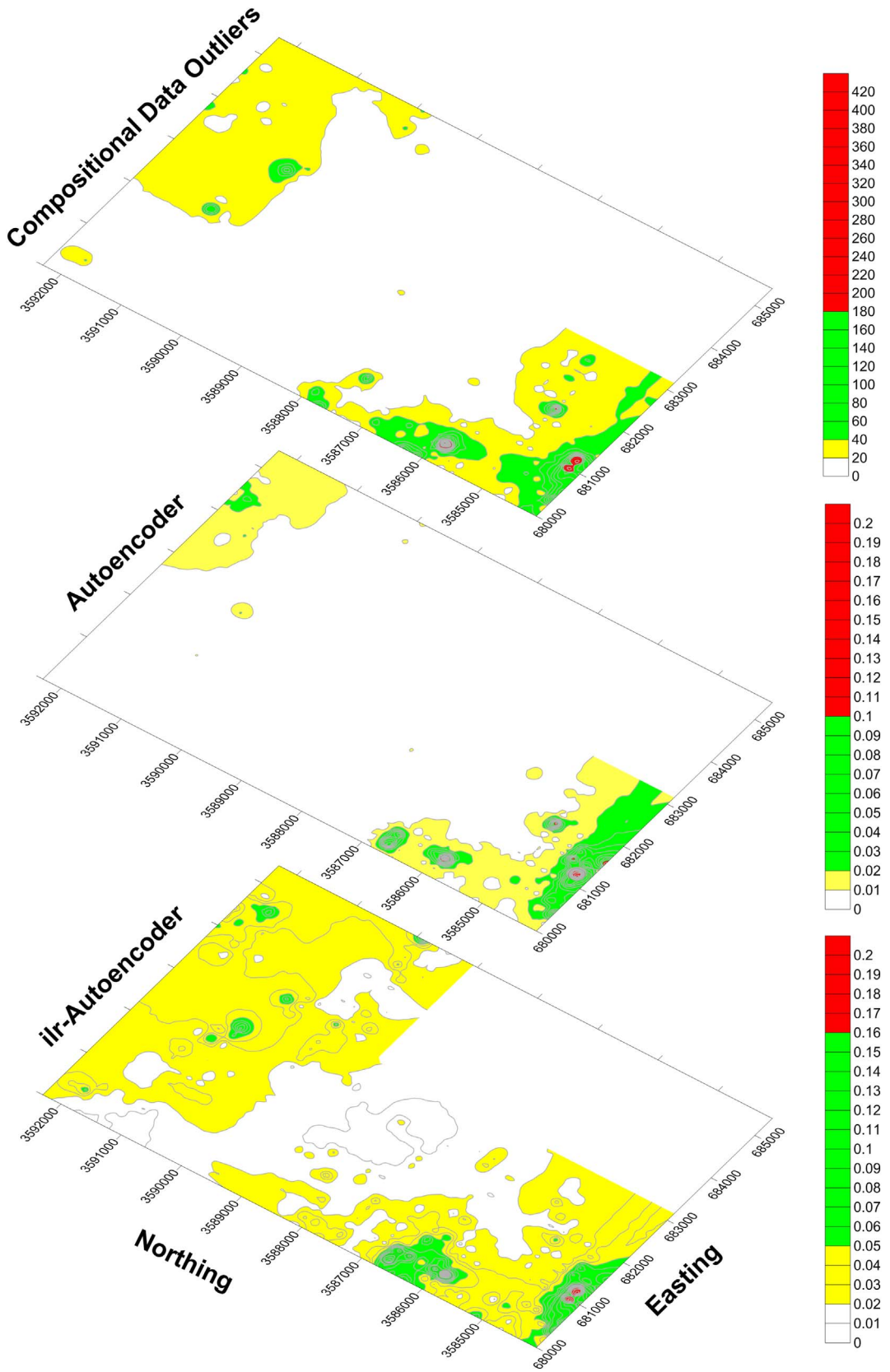

Fig. 5. Maps representing the location of anomalies or high errors in Hamich area.

There was a strong correlation between the known Cu-Pb indexes, vein outcrops and the anomalous samples, indicating that the methods used in this study are powerful tools for recognition of multivariate geochemical anomalies. It is also considerable that all three anomaly maps in this study (Fig. 5) showed an anomaly region in the northeast of Hamich area that was not referred to in the exploration report and it needs further studies in the future. This highlights the preference of novel over conventional methods that will require reconsidering the old

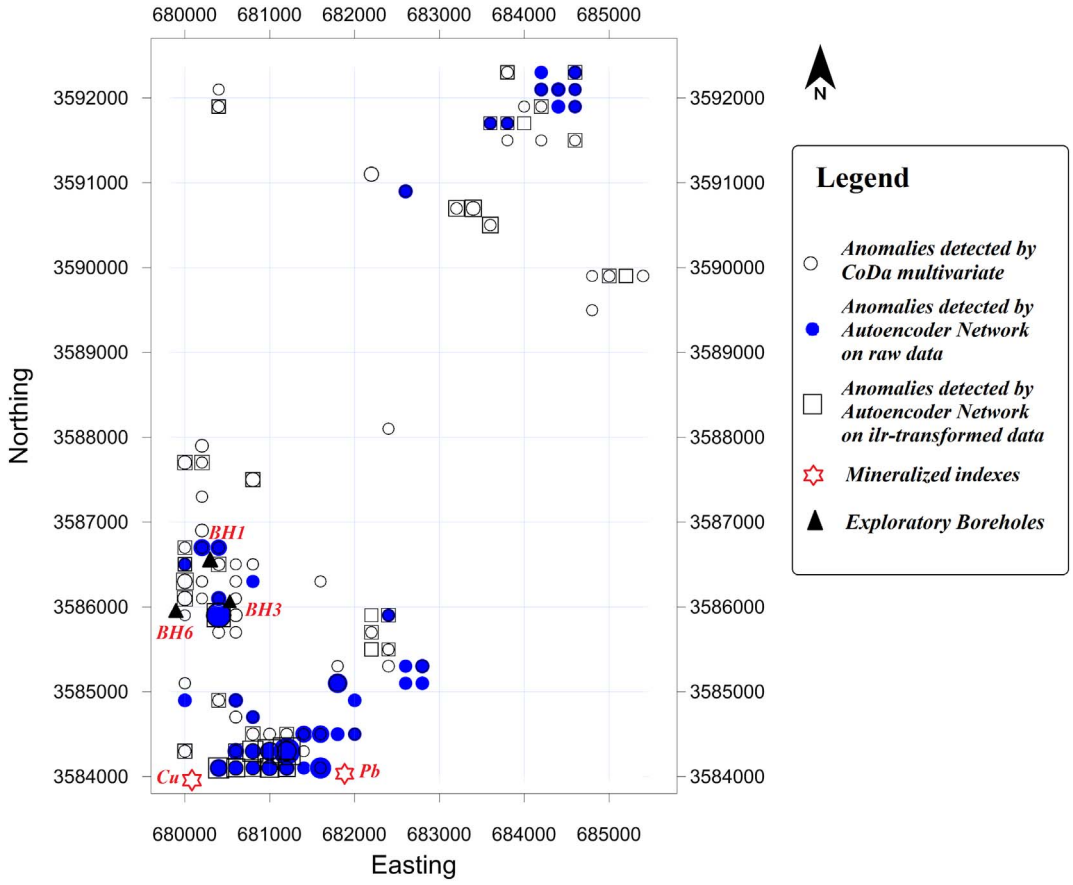

Fig. 6. Overlaid maps of anomalous samples detected by both methods and locations of drilled boreholes containing mineral veins and mineralized indexes. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

exploration projects.

Acknowledgments

The authors would like to say the special thanks to South Khorasan Industry, Mine and Trade Organization (IMTO) for providing the data.

References

Ahmed, A.A., Yu, K.K., Xu, W.W., Gong, Y.Y., Xing, E.E., 2008. Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks. In: Computer Vision-ECCV 2008. Springer, pp. 69–82.

Aitchison, J.J., 1982. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological) 139–177.

Bengio, Y.Y., 2009. Learning deep architectures for AI. Foundations and trends® in Machine Learning 2 (1), 1–127.

Bengio, Y.Y., Lamblin, P.P., Popovici, D.D., Larochelle, H.H., et al., 2007. Greedy layer-wise training of deep networks. Advances in neural information processing systems 19, 153.

Boureau, Y.Y., Cun, Y.L.Y.L., et al., 2008. Sparse feature learning for deep belief networks. In: Advances in Neural Information Processing Systems, pp. 1185–1192.

Buccianti, A.A., 2011. Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach. Compositional Data Analysis: Theory and Applications 255–266.

Buccianti, A.A., Mateu-Figueras, G.G., Pawlowsky-Glahn, V.V., 2006. Compositional Data Analysis in the Geosciences: From Theory to Practice. vol. 264 Geological Society of London.

Camp, V.V., Griffis, R.R., 1982. Character, genesis and tectonic setting of igneous rocks in the Sistan Suture Zone, Eastern Iran. Lithos 15 (3), 221–239.

Candel, A.A., Parmar, V.V., LeDell, E.E., Arora, A.A., 2016. Deep Learning with H2O. H2O.ai Inc.

Carranza, E.J.M.E.J.M., 2011. Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. Journal of Geochemical Exploration 110 (2), 167–185.

Chayes, F.F., Trochimczyk, J.J., 1978. An effect of closure on the structure of principal components. Journal of the International Association for Mathematical Geology 10 (4), 323–333.

Chen, Y.Y., Lu, L.L., Li, X.X., 2014. Application of continuous restricted Boltzmann machine to identify multivariate geochemical anomaly. Journal of Geochemical Exploration 140, 56–63.

Consulting Engineers Co., K., 2009. Detailed Exploration of Mineral Deposits of Birjand, Hamich Area, South Khorasan. report.

Filzmoser, P.P., Garrett, R.G.R.G., Reimann, C.C., 2005. Multivariate outlier detection in exploration geochemistry. Computers & geosciences 31 (5), 579–587.

Filzmoser, P.P., Gschwandtner, M.M., 2012. mvoutlier: multivariate outlier detection based on robust methods. R package version 1 (7).

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

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

Filzmoser, P.P., Hron, K.K., Reimann, C.C., 2009b. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environment 407 (23), 6100–6108.

Filzmoser, P.P., Hron, K.K., Reimann, C.C., 2012. Interpretation of multivariate outliers for compositional data. Computers & Geosciences 39, 77–85.

Fletcher, W.K.W.K., 2013. Analytical Methods in Geochemical Prospecting. Elsevier.

Freund, Y.Y., Haussler, D.D., 1994. Unsupervised Learning of Distributions of Binary Vectors Using Two Layer Networks.

Gallo, M.M., Buccianti, A.A., 2013. Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno river (Tuscany, central Italy). Environmetrics 24 (4), 269–277.

Hinton, G.E.G.E., Osindero, S.S., Teh, Y.-W.Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural computation 18 (7), 1527–1554.

Hinton, G.E.G.E., Salakhutdinov, R.R.R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.

Hinton, G.E.G.E., Salakhutdinov, R.R.R.R., 2008. Using deep belief nets to learn covariance kernels for Gaussian processes. In: Advances in Neural Information Processing Systems, pp. 1249–1256.

Karimpour, M.H.M.H., Mazaheri, A.A., 2009. Hydrothermal alteration mapping in SW Birjand, Iran, using the advanced spaceborne thermal emission and reflection radiometer (ASTER) image processing. Journal of applied sciences 9.

Larochelle, H.H., Erhan, D.D., Courville, A.A., Bergstra, J.J., Bengio, Y.Y., 2007. An empirical evaluation of deep architectures on problems with many factors of variation. In: Proceedings of the 24th International Conference on Machine Learning, pp. 473–480 ACM.

Lee, H.H., Grosse, R.R., Ranganath, R.R., Ng, A.Y.A.Y., 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp.

609–616 ACM.
Maronna, R.A.R.A., Martin, R.D.R.D., Yohai, V.J.V.J., 2006. Robust Statistics: Theory and Methods. J. Wiley.
Palarea-Albaladejo, J.J., Martín-Fernández, J.A.J.A., 2015. zCompositions-R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems 143, 85–96.
Poultney, C.C., Chopra, S.S., Cun, Y.L.Y.L., et al., 2006. Efficient learning of sparse representations with an energy-based model. In: Advances in Neural Information Processing Systems, pp. 1137–1144.
Reimann, C.C., Filzmoser, P.P., Garrett, R.G.R.G., 2005. Background and threshold: critical comparison of methods of determination. Science of the Total Environment 346 (1), 1–16.
Rollinson, H.H., 1992. Another look at the constant sum problem in geochemistry. Mineralogical Magazine London 56, 469-469.
Rosseeuw, P.P., Van Zomeren, B.B., 1990. Unmasking multivariate outliers and leverate points. Journal of the American Statistical Association 85, 633–639.
Salakhutdinov, R.R., Hinton, G.E.G.E., 2007. Learning a nonlinear embedding by preserving class neighbourhood structure. In: International Conference on Artificial Intelligence and Statistics, pp. 412–419.
Tirrul, R.R., Bell, I.I., Griffis, R.R., Camp, V.V., 1983. The Sistan suture zone of eastern Iran. Geological Society of America Bulletin 94 (1), 134–150.
Vassighi, H.H., Soheili, M.M., Eftekharnejad, J.J., Stoecklin, J.J., 1975. Geological 1:100000 sheet of Sar-e-Chah-e-Shur no.7754. Geological Map Report, Geological Survey of Iran.
Vincent, P.P., Larochelle, H.H., Bengio, Y.Y., Manzagol, P.-A.P.-A., 2008. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103 ACM.
Xiong, Y.Y., Zuo, R.R., 2016. Recognition of geochemical anomalies using a deep autoencoder network. Computers & Geosciences 86, 75–82.