Current Research in Geosciences

Original Research Paper

Using Robust Statistics, Exploring Document Analysis and Gis for Defining and Localizing Geochemical Anomalies-Case Study in Sin Quyen Copper Mine, Lao Cai Province, Viet Nam

Quoc Lap Kieu and Huu Tap Van

Faculty of Environment and Earth Science,
Thainguyen University of Sciences, Tan Thinh Ward, Thai Nguyen City, Vietnam

Article history
Received: 20-06-2017
Revised: 24-07-2017
Accepted: 31-07-2017

Corresponding author:
Huu Tap Van
Faculty of Environment and Earth Science, Thai Nguyen University of Sciences, Tan Thinh Ward, Thai Nguyen City, Vietnam
Email: vanhuutap@gmail.com

Abstract: Application of robust statistic and Exploratory Data Analysis (EDA) as well as the support of Geography Information System (GIS) for defining and assessing geochemical anomalies. The research was conducted in Sin Quyen copper mine, Lao Cai province, Viet Nam with 1,284 combination of aquatic sedimentary geochemical sample in scale of 1/200,000. The area of 5,717.8 km^2 was chosen to make samples for defining and localizing geochemical anomalies of copper element. This method that was used for preventing interference is simple and not requires a document to have to follow any distribution law. It was compared to classical statistic method to assess advantages and disadvantages. Results from the research were suitable in the fact that areas of geochemical anomalies which were defined by robust statistics and EDA (487.9 km^2) were higher than by classical method ((251 and 272 km^2) and equivalent to discovered mine in the region.

Keyword: EDA, GIS, Geochemical Anomaly, Normal Distribution Law, Robust Statistics

Introduction

Definition of geochemical anomaly is one of basic problems in solving the data of geochemical explorations. Since 1950, geochemists were agreed that the distribution of chemical elements in geochemical field followed Normal Distribution Law (NDL) or Lognormal Distribution Law (LDL) (Ahrens 1953; 1954). An analysis of correlate from the classical statistic method required a data to follow normal distribution law. Thus, when use of the classical statistic to determine chemical anomalies, the first step of this method is to research meticulous distribution forms of data and test a supposition of data for following normal distribution law or not. Many previous researchers agreed that geochemical data often followed normal distribution law but in recently some researchers proved that the amount of elements from geology was not limited by Normal Distribution Law (NDL) or Lognormal Distribution Law (LDL) (Jun, 2007; Huan *et al*., 2009). Therefore, a transfer of datas (logarite, cosine, square root etc…) and removal of particular lognormal picks (high and low values) to data to follow normal and lognormal distribution law were two common resolution options (Jun, 2007; Huan *et al*., 2009; Tao *et al*., 2011; Stanley, 2006).

Many statistic methods that were used to determine geochemical anomalies include probability chart, multivariate and univariate statistical analyzes (Sinclair, 1991; Stanley and Sinclair, 1989), statistical analysis methods used space in moving average or Kriging geostatistics (Agterberg, 2012; Grunsky and Agterberg, 1988). In 1994, the first time, Cheng *et al*. (1994a; 1994b; 1994c) introduced concentration-area multiracial method with using division multiple geometric principles for determining geochemical anomalies by classification of background and anomaly and has been used effectively in Canada.

The objective of this paper was to introduce robust statistic and exploratory data analysis-EDA as well as the support of Geography Information System (GIS) to determine and localized geochemical anomalies of copper element in Sin Quyen mine with 1,284 combinations of aquatic sedimentary geochemical samples in scale of 1/200,000.

DOI: 10.3844/ajgsp.2017.25.31

Determiniation and Lolization Methods of Geochemical Anomaly

Determination of Geochemical Anomaly by Classical Statistic Methods

With classical statistic method, the data are followed NDL is considered the main condition for conduction. Therefore, the distribution of data needs to be researched in details before data processing. The analysis of data distribution forms for approximately NDL is very important. Much distribution test methods were used in recently, including qualitative and quantitative methods.

Qualitative method is determinated through charts from EDA (Tukey, 1977; Clemens *et al*., 2008). The state of distribution data can be determined by eyes. The simplest way is to use probability distribution charts by judging the bell-shaped or U-shaped distribution; The position of the box and the median can be observed through the box plot scatter gram if the box is in the medial position and the median is located at the medial position of the box, it can be assumed that distribution data are symmetrical, whether the sideways distribution; using the stem-and-leaf scatter gram to observe the symmetrical distribution of the data; standard probability scatter grams are often used by observing if the data lies on or around the diagonal line in the first quarter may be considered as NDL. With the simple, visual and non-computationally scatter gram method, the amount of provided information is only an important addition to NDL test. Many scatter grams, including probability distribution, density trace, box plot and one-dimensional scatter gram and the cumulative distribution function, can be combined to makes it more efficient for studying the distribution pattern of the data Fig. 2.

Quantitative methods including robust statistics, non-parametric method (kolmogrov-Smirnov, Chi-square và Shapiro-Wilk methods) based on the principle of comparison between the two samples, one with known distribution (NDL) and other with unknown distribution. If P value (NDL level of significance) is more than 0.05, the sample thesis with standard is accepted. It is possible to test the obliqueness, curvature of the standard curve, the extreme distribution of S, K and Jarque-bera tests are two common methods.

If the obliqueness and curvature are approximately 0, studying sample is considered to follow NDL. However, the result is easy to be interference by anomaly point through reducing reliability. In this paper, both qualitative and quantitative analysis methods were combined to test the distribution of data for following (or similar) NDL to be persuasive.

Normally, middle elements are usually close to NDL while micro elements close to LDL. In reality, it is not entirely NDL or LDL, therefore, the first step is to process "raw" data for following NDL. Abnormal (A) is determined by the mean of the average value of the standard deviation (*sdev*) of the formula (1), (2), (3):

$$mean = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$sdev = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

$$A = mean + 2 * sdev \tag{3}$$

With, x_i is the amount of i sample, n is the total number of samples.

In the case of the amount of elements follows NDL, the data needs to be converted. Using data after transformation through formulas (1), (2), (3) to calculate anomaly, this value needs to be transformed back to (Rong, 2007).

Geochemical Anomalies are Determined by Anti-Interference Statistic and EDA

With Anti-interference Statistic (AS), distribution forms of the data need to be followed NDL to make prerequisites, but usually the complete (or almost) data does not follow NDL (even after transforming data). Therefore, the reliability and randomness of results were determined by this method need to be discussed. Due to the limitations of classical statistic, AS and EDA techniques were used to identify geochemical anomalies in this study. Characteristics of the method is that only a small amount of anomalies leading to deviation of the ideal form of distribution so that the results are affected less. In the case of large amount of anomaly exists, there will be no accidental consequences. Classical statistic follows the rule of weight increasing in the data, therefore, the robust of anomaly points is reduced with different levels, including zero.

Classical statistic defines two anti-interference factors are the median absolute deviation-mad and inter-quartile range-iqr. The geochematic anomaly can be determined by formulas [*median*+2**mad*] and [*median*+1.5**iqr*]. The quarter range is determined by the absolute difference of the first and third quarter. Mad and *iqr* are almost identical to normal deviations in classical statistic, so [*median*+2*mad*] and [*median*+1.5 * *iqr*] are similar to [*mean*+2*sdev*], *mad* are defined by formula (4). From the mad and *iqr* formulas, it is easy to see that these two quantities are less disturbed by the anomalies than normal deviations:

$$mad = 1.483 * median\left[\left| x_i - median(x_i)\right|\right] \tag{4}$$

DOI: 10.3844/ajgsp.2017.25.31

1.483 is the scaling factor (multiplied by this value making mad and normal deviation) (Keng, 1991).

EDA is an extraordinary and non-parametric statistic method for processing data, using statistics against anomaly and introducing simple and effective types of charts. From there, the characteristics, distribution and structure of the data can be quickly detected. The EDA does not require any data to follow any form of distribution, but rather relies on an inherent pattern to distinguish anomalies, thereby determining the total anomaly and overall ground coverage. Box plot is used in EDA to identify anomalies and studied the distribution of data such as location, dispersion, deviation, tail length and out-of-box anomalies Fig. 1. Geochemical anomalies can be identified through formulas (5), (6):

$$lif = Q_1 - (1.5 * iqr) \quad (5)$$

$$uif = Q_3 + (1.5 * iqr) \quad (6)$$

At the two tails of the box (smaller than *lif* and larger than *uif*), there is the possibility of a "free anomaly" case, which indicates that the data is relatively uniform. If wanting an anomaly, the percentage method or the cumulative probability charts are used to determine the boundary between the background and the anomaly. The percentage method, which can be used as 95 or 98% position as the dividing line on the probability accumulation chart, will easily see anomalies.

Application of GIS for Screening and Evaluating Geochomatic Anomaly

According to the method of geochemical sedimentation of hydro-scale sediment with the scale of 1/200,000, 1,484 samples were taken at the mineral mine of Sin Quyen in Lao Cai province. The average density is about 4km^2/sample. Each sample consists of Ag, As, Au, Be, Cd, Cu, Hg, Li, Mn, Mo, Nb, Pb, etc ... In the study area, Copper (Cu) is one of three elements for forming ore (Cu, Ag and Au). Classical statistic and EDA methods were used to identify geochemical anomalies, only for the Cu element was studied in this paper. Based on the coordinates and content, interpolate with the size of each square is 0.1 km^2 to create contour lines and elevation model. The anomaly was identified by above mentioned methods; delineation of anomaly is conducted to determine the area. Incorporation of the contour maps and map layers of discovered mines in the area and concurrently combination to the geology of the study area to screen, evaluate and compare the results of the anomaly calculations by the above methods.

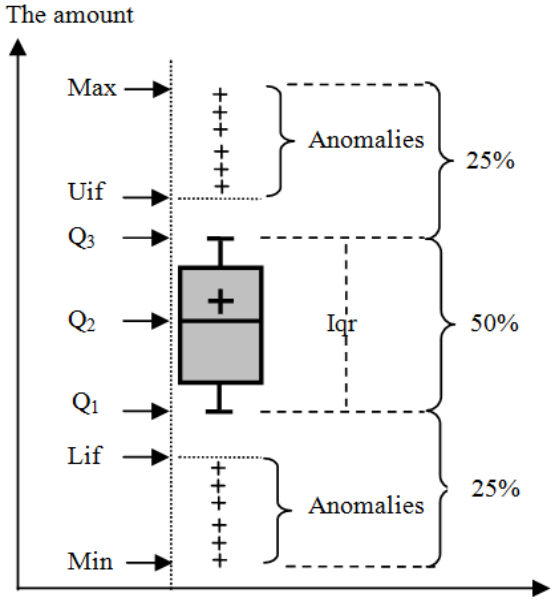

Fig. 1. Boxplot chart Max- the largest value; Uif-anomalies belowing the level; Q3- the third quarter; Q2-quartier (median); Q1- the first quarter; Iqr-quarter range; Lif-anomalies on the line; Min- the smallest values

Results and Discussion

The qualitative analysis method was used to study the distribution of the amount of Cu element by combining the use of multiple charts, including charts of probability distribution, density, box plot, one-dimensional dispersion and cumulative distribution function. From Fig. 2 and Table 1 the summary of the calculation results shows that the distribution of the amount of Cu did not follow NDL Fig. 2a and 2b. The difference between the average value and the median (32.0 and 27.1, respectively), when eliminating 12 typical anomalies and the standard deviation from the absolute median deviation, was significantly. In order to increase the reliability of the distribution formulation of Cu element, the Kolmogrov-Smirnov and Shapiro-Wilk calibration methods combined the calculation of the obliquity and curvature of the standard curve was used. Table 1 shows that the standard distribution with significance level (P) was zero (It is mean that NDL was rejected because P value was less than 0.05), the curvature coefficient of 258.3 and the obliquity of 15.2 were very large. This result demonstrated that the amount of Cu did not follow NDL, it tended to deviate to the right. In order to proceed with the next step for evaluating the amount of Cu with following to NDL, data conversion was carried out by logarite method. The qualitative analysis method Fig. 2b and 2c shows that the data was concentrated better at the average value and

median, but tended to shift to the right. The average value (P) of zero, a deviation of 3.36 and a curvature of 22.5 resulted in Kolmogorov-Smirnov and Shapiro-Wilk quantitative methods, was very high. The above indicators demonstrated that data did not follow NDL after transformation. It means that data did not follow NDL before converting. Therefore, the use of classical statistics to identify anomalies was not consistent with the statistical hypothesis for geochemical exploration data before and after transformation. To compare the classical statistic methods to EDA, the method of logarite transformation and data the elimination of typical anomalys was conducted. Results from Table 1 and Fig. 4 showed that two methods of effect elimination of anomaly had relatively large differences.

Transforming method of the data using classical statistic discovered more anomalous than that of apparent abnormalities elimination method Fig. 4a and 4b but the anomaly area in the south was relatively small (almost none) and not exactly match the reality of the ore mines discovered at this location. The reliability of classical statistic depends on how many abnormalities and distributions of the data are eliminated, so results were more random. Robust statistics [*Median*+2*mad*] and [*median*+2*iqr*] detected more an area of the anomaly than by the classical method, the area of 487.9 km^2 compared to the area of anomalies determined by cyberspace: the method of typical anomaly elimination with the area of 251 km^2 and transforms of the area of 272 km^2.

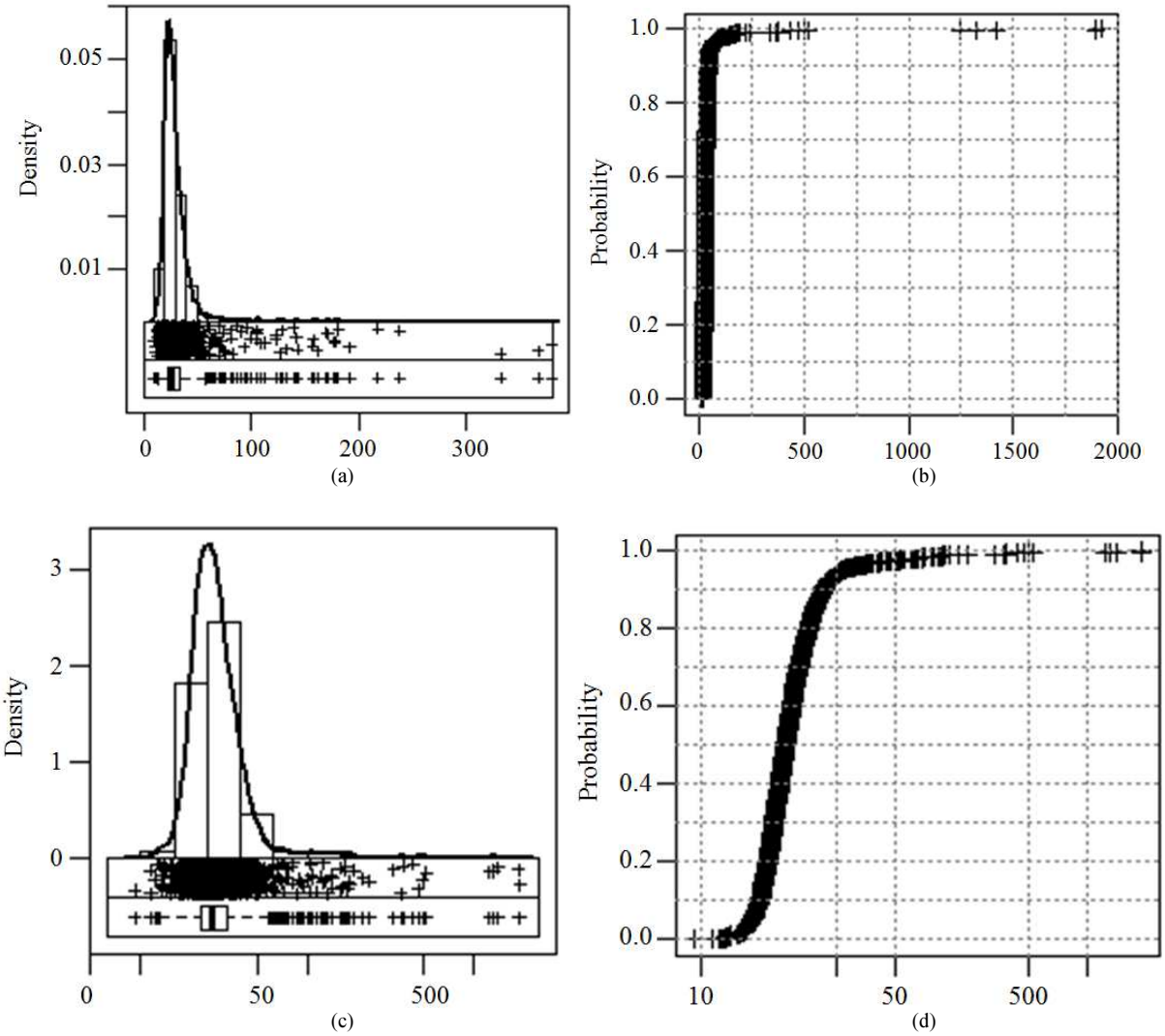

Fig. 2. Combination of probability, density, box plot, one-dimensional distribution chart and cumulative distribution function (figure a, b-use data before transformation, c, d-data after transformation)

DOI: 10.3844/ajgsp.2017.25.31

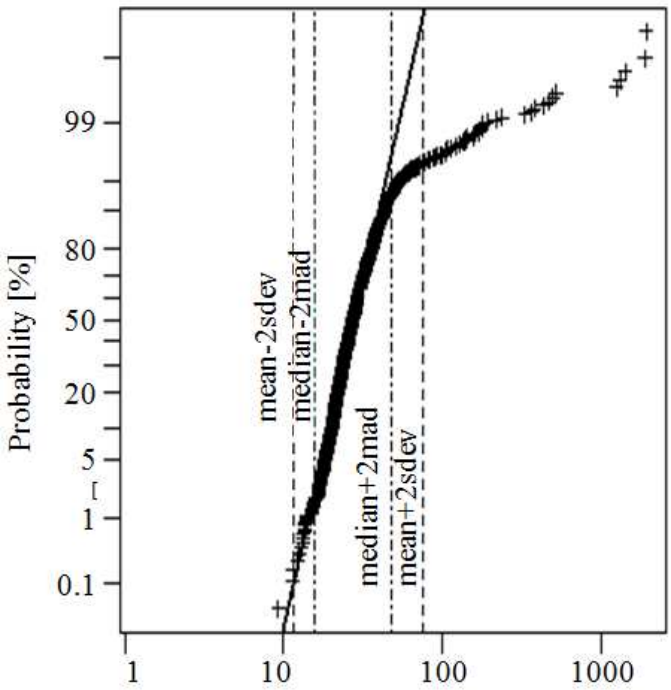

Fig. 3. Comparison of results of anomaly detection by methods of classical and robust statistics using the probability accumulation charts

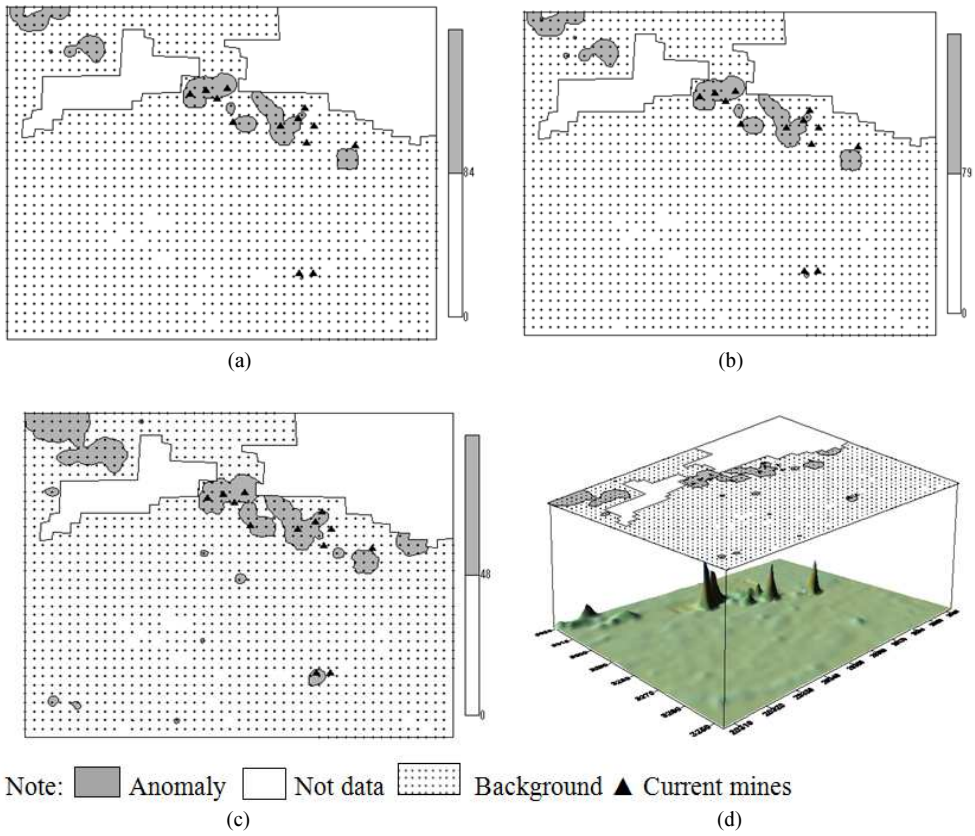

Fig. 4. Charts of anomalies: Classical statistic-Elimination of typical anomalies (a) Data transformation (b), robust statistics and EDA: (c), abnormal model (d)

Table 1. Summary table of standard deviation and results of anomaly determination by different methods

Index/anomaly	Cu	log (Cu)
Obliquity	15.2	3.6
Curvature	258.3	22.5
P k-s	0.0	0.0
P s-w	0.0	0.0
Mean	32.0	1.5
Sdev	26.0	0.2
Median	27.1	1.4
Mad	×	0.2
Iqr	×	0.2
Mean+2sdev	84	79.0
Median+2mad	×	48.0
Uif	×	48.0

Charts of q-q Fig. 3 proved that geochemical anomalies were identified by robust statistics method with higher reliability than classical statistic through the correlation between observed value and expected value of anomaly when starting to deviate from the diagonal line of the first quadrant.

Based on the current status of the ore mines found in the area (5 mines in north, 4 mines in northeast, 1 mine in east and 2 mines in south) Fig. 4, results were consistent with the situation. At the same time, some fairly large anomalies are found in the northwest, west, southwest and west east, but these mines are undetected by geological formations.

Conclusion

Based on the assumption that the distribution of geochemical data needs to be followed NDL but in fact that it did not strictly follow NDL as well as LDL, data transformation (logarite, cosine or square root) did not make the data to follow NDL after transforming, thereby, classical statistic was used to treat these data led to inaccurate results (in the south of the study area). Due to characteristics of geochemical data, average values and standard deviation could not be used to accurately assess the center and deviation from the center of data. Since statistical hypothesis was not guaranteed, the use of classical statistic to determine the anomaly should be reconsidered. In addition, the reliability of results was still to be taken into consideration from classical statistic method (the data had to be transformed before and after computation). The median values in robust statistics were estimated the data center position better than average values and the absolute median deviation. Quarter ranges were able to estimate better than the deviation from the central position. These quantities were capable of reducing the robust caused by the anomaly. This method was simple, not require the data to follow specific distribution laws. The research results through robust statistic method and EDA technique show that the ability to treat geochemical exploration data was more effective than the classical statistic method and suitable of the reality, the area of anomaly was larger when using classical statistic, matching discovered ore mines in the area.

Acknowledgement

This article is the result of the project "Developing strategy for protection of natural resources and environment in Lao Cai province. We would like to express our sincere thanks to Sin Quyen mine Management Board for providing this data and supporting research materials.

Funding Information

The research was funded by Thai Nguyen University Project in 2017.

Author's Contributions

Quoc Lap Kieu: Study survey overview, model building, data collection.

Huu Tap Van: Outlining the proporal and revising the paper.

References

Ahrens, L.H., 1953. A fundamental law of geochemistry. Nature, 172: 1149-50. DOI: 10.1038/1721148a0

Ahrens, L.H., 1954. The lognormal distribution of the elements (a fundamental law of geochemistry and its subsidiary). Geochim. Cosmochim. Acta, 5: 49-73. DOI: 10.1016/0016-7037(54)90040-X

Agterberg, F.P., 2012. Multifractals and geostatistics. J. Geochem. Exploration 122: 113-122. DOI: 10.1016/j.gexplo.2012.04.001

Cheng, Q., F.P. Agterherg and S.B. Ballantyne, 1994a. The separation of geochemical anomalies from hack ground by fractal methods. J. Geochem. Explor., 51: 109-130.

Cheng, Q., G.F. Bonham-Carter and F.P. Agterberg, 1994b. Fractal modelling in the geosciences and implementation with GIS. Proceedings 6th Canadian Conference on Geographic Information Systems, (GIS' 94), Ottawa I, pp: 565-577.

Cheng, Q., 1994c. Multifractal modelling and spatial analysis with GIS: Gold potential eatimation in the mitchell-sulphurets area, Northwestern British Columbia. PhD Thesis, University of Ottawa, Ottawa.

Clemens, R., F. Peter and G. Robert, 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. 1st Edn., Wiley, Chichester, ISBN-10: 0470987596, pp: 362.

Grunsky, E.C. and F.P. Agterberg, 1988. Spatial and multivariate analysis of geochemical data from metavolcanic rocks in the Ben Nevis area, Ontario. Math. Geol., 20: 415-446.

Huan, Y.D., G. Min and L. Rui, 2009. A new method to determine geochemical anomaly threshold-contend-sequence method. Comput. Techniques Geophys. Geochem Explor., 31: 154-157.

Jun, S.Z., 2007. Multifractal method of geochemical threshold in mineral exploration. Comput. Techniques Geophys. Geochem. Explor., 29: 54-57.

Keng, H.Y., 1991. Multivariate Statistic Analysis in Geochemistry. 1st Edn., China University of Geosciences Press, Wuhan.

Rong, L.X., 2007. Geochemical Exploration. 1st Edn., Metallurgical Industry Press, Bejing.

Stanley, C.R., 2006. Numerical transformation of geochemical data: Maximizing geochemical contrast to facilitate information extraction and improve data presentation. Geochem. Explor. Environ, Anal., 6: 69-78.

Sinclair, A.J., 1991. A fundamental approach to threshold estimation in exploration geochemistry: Probability plots revisited. J. Geochem. Explor., 41: 1-22. DOI: 10.1016/0375-6742(91)90071-2

Stanley, C.R. and A.J. Sinclair, 1989. Comparison of probability plots and gap statistics in the selection of threshold for exploration geochemistry data. J. Geochem. Exploration 32: 355-357.

Tao. Y., C.S. Yu and L.R.Y. Zi, 2011. Methods for determination of the lower geochemical anomaly limit and the rationality discussion. Contributions Geolo. Mineral Resources Res., 26: 96-101.

Tukey, J.W., 1977. Exploratory Data Analysis. 18th Edn., Addison-Wesley Publishing Company, Reading, ISBN-10: 0201076160, pp: 688.