Journal of Earth & Statistics

Journal homepage: http://www.Jes.arakut.ac.ir

Doi: 10.22034/jes.2022.253179

Ministry of Science, Research & Technology
Arak University of Technology

Geochemical analysis of sandstones in Qom Formation to identify unknown data using comparison of data mining methods

Javad sorby

Department of Surveying Engineering, College of Earth Sciences Engineering, Arak University of Technology, Arak, Iran.

Article Info

Keywords:
Geochemical faces, Linear multiple correlation, Data mining, Curve fitting, Qom.

*Corresponding Author's Email Address:

javadsorby@gmail.com

Abstract

The purpose of this study is to identify the missing and unknown data of the sandstones of Qom Formation based on the available geochemical data and to complete the information of the rock units of this formation in the data mining methods that are used to predict and track the analyzes used in The sample obtained Some data are unknown. For this purpose, after conducting geological studies, 30 samples were taken for XRD-XRF analysis. The samples are then analyzed by two methods of multiple linear correlation and curve fitting models to identify the best model and description for such unspecified data that are in some elements of the sample. The results of this study show that multivariate linear regression (MLR) should be used to predict samples that have a large amount of uncertain data and univariate methods in such fits cannot provide accurate information.

Introduction

In geochemical explorations, measuring some oxides and elements is associated with problems such as increasing costs and increasing errors. There are many solutions and methods to solve such data, which can be referred to modeling and mathematical problems (Hassani Pak, 2011).), In multivariate linear regression analysis and curve fitting so that the element or oxide itself is a function of the amount of trace elements can be identified and predicted.

Such studies can be compared to studies (Sonja Cerar, 2018) that have compared the methods of predicting the composition of the isotope of oxygen in shallow groundwater (Saffari, 2013) using the landslide event line (Azimand, 1396) has estimated the concentration of ozone on the surface of the earth. (Ghadimi, 1399) Gold carat has been predicted in Zarshuran mine, Has studied karst formations. The purpose of this study is to model and analyze sedimentary facies and compare different fitting methods such as multiple linear regression (MLR) and curve fitting in MATLAB software, which can help to better identify and predict them. The data obtained from the chemical analysis of sedimentary withdrawals of Qom Formation in the northwest of Tafresh, which are shown in (Figure 1), have been investigated.

Journal of Earth & Statistics
Journal homepage: http://www.Jes.arakut.ac.ir
Doi: 10.22034/jes.2022.253179
Ministry of Science, Research & Technology
Arak University of Technology

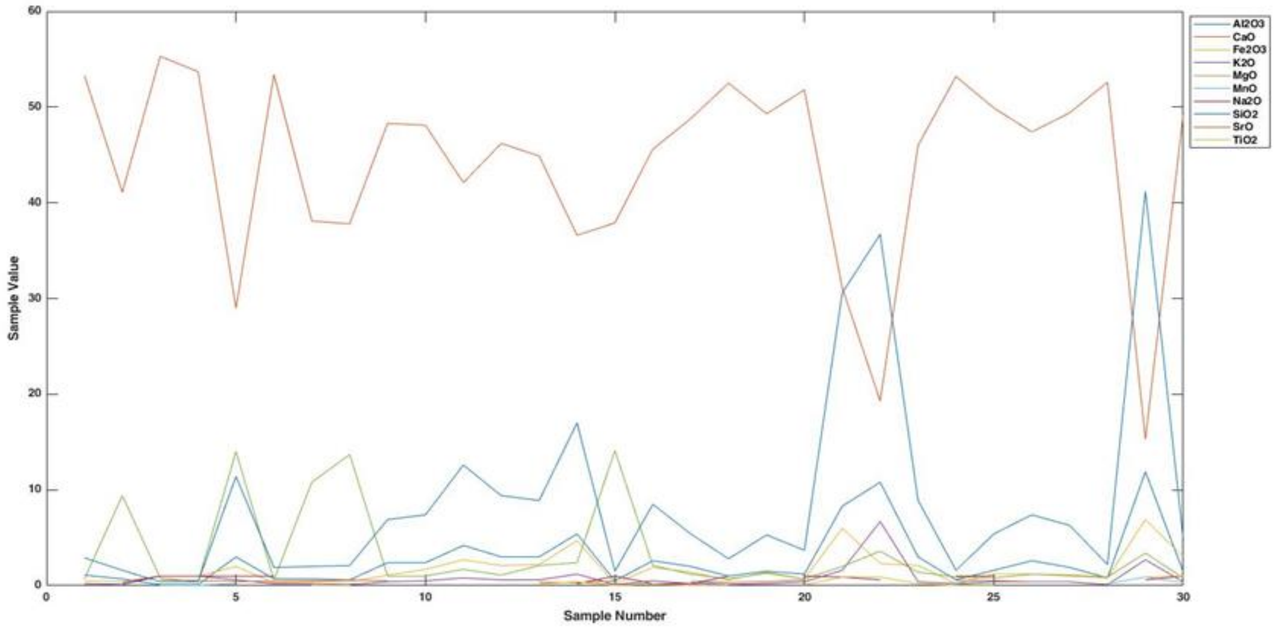

Fig 1. Graph of values obtained from chemical analysis of samples taken (processing in MATLAB software)

Investigation roach

One of the methods used to analyze geochemical data is modeling by linear multiple correlation (MLR). A regression model describes the relationship between a response and a predictor. Linearity in a model refers to the linearity of predictor coefficients. With this method, several different variables can be analyzed and studied simultaneously. To obtain better results through MLR, the samples must be sufficient in number and accuracy because this method is highly sensitive to incorrect information and the entry of such data may lead to high errors in the results obtained. In addition, to use this method, the variables must have a normal distribution and their change must follow a linear relationship. Multiple regression, in fact, expresses the relationship between a series of predictor variables and the desired response variable (Balan et al, 1995). Used.

A) Methods based on eigenvalues:

Using special values and special vectors, orientation is detected with maximum variability. Then, by defining new variables, which are a linear combination of the initial variables, the number of variables is reduced and the role of each variable in variability is determined. The first axis rotates in such a way that the greatest expansion and elongation of data appears along it. In these new coordinates, the origin is transferred to the point ($\bar{X}1$, $\bar{x}2$). The direction of this transfer does not affect the correlation of the principal component variables. (Hassani Pak, 1390)

Initially, this component is the most variable and most of the variability can be explained along this line, which from an exploratory point of view, the length of this component corresponds to the length of the most variability. Are the main variables, so the percentage of variability that can be justified by the first component. (Table 2 and Figure 2).

B) Curve fitting

Is a program that provides functions for fitting curves and data levels. This type of fitting allows you to analyze exploratory data, preprocessing and post-processing data, compare candidate models, and eliminate inconsistencies in data mining (Curve Fitting Toolbox, 2020a). In this type of Coefficients processing, the number of coefficients is modeled, the minimum number of coefficients is more appropriate to help make decisions, and the correlation coefficient will also show more. The sum of the squares of error (SSE) is that the closer the number is to zero, the more confident the prediction is. The square of the correlation (R-square) between the response and prediction values. A number has a value close to one more. Adjusted degrees of freedom (Adjusted R-square) is a square of correlation and as such has a value close to a better fit. It is more useful.

Results and Discussion

In the study of geochemistry of sandstones, oxide of the main elements is very useful in determining the characteristics of such rocks and even determining their mineralogy. In the present study, the initial multivariate study of the relationships of each of the main oxides in the content of these sandstones has been investigated, the results of which are presented in the comparative matrix (Table 1). According to the geochemical analysis performed for the samples, some data of SrO, TiO2 and MnO variables are unknown and have been selected for prediction. According to the correlation table (Table 1) and the main component diagrams (Figure 1), they show the highest correlation with Al2O3, Na2O and Fe2O3 oxides and have more complete data to fit and find unknown data. Due to the constant nature of aluminum oxide and iron, many variables have been compared with these oxides. Calcium and Potassium oxidation of Al2O3 in the case of the study of the abundance of the mines in these cases of these elements in the aluminusiacaters, It is the main in detrital sediments and the positive adaptation of Al2O3 with Na2O and TiO2, SiO2, MnO, K2O, Fe2O3 in the studied samples shows the presence of this set of elements within the detrital stages.

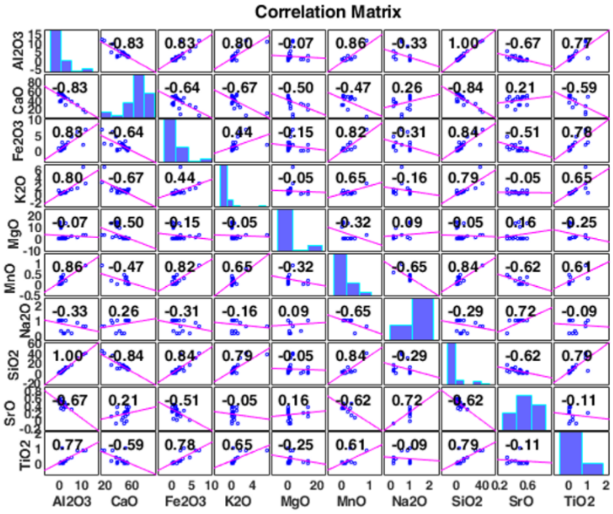

Fig. 2. Correlation matrix of variables Al2O3, CaO, Fe2O3, K2O, MgO, MnO, Na2O, SiO2, SrO, TiO2

As can be seen from Figure 2, the red line in this diagram is a correlation process that varies in a positive or negative direction, ie the more this line tends to a positive number and more, the higher the bit correlation of the two variables will be. +1 The closer the variables are to each other, the greater the effect. (Matlab software)

Al2O3 is often used as a suitable factor for comparing different types of rocks (Cardenas et al., 1996) because it is relatively unchanged during the processes of weathering, diagenesis and metamorphism. Na2O, K2O and CaO are known as the most variable phases in sandstones and mudstones. (Gateneh, 2000)

Al2O3 is positively correlated with Fe2O3, K2O, MnO, SiO2 and TiO2 and inversely related to Na2O, SrO, MgO, CaO. Aluminum is especially present in the structure of clays. A positive correlation between K2O and Al2O3 indicates the concentration of potassium ores in the samples under study, which has a significant effect on the dispersion of aluminum. (Das et al. 2006; Pettijohn et al., 1987)

Journal of Earth & Statistics
Journal homepage: http://www.Jes.arakut.ac.ir
Doi: 10.22034/jes.2022.253179

Ministry of Science, Research & Technology
Arak University of Technology

Table 1. Eigenvalues are related to the variables Al2O3, CaO, Fe2O3, K2O, MgO, MnO, Na2O, SiO2, SrO, TiO2

مقادیر ویژه	شاخص وضعیت	Al2O3	CaO	Fe2O3	K2O	MgO	MnO	Na2O	SrO	TiO2
2.59265	1	4.92362e-05	6.62619e-05	6.59992e-05	0.000148699	0.00026829	0.000498942	0.000235399	8.09548e-05	9.72544e-05
1.01444	2.55574	2.65854e-07	2.8714e-05	0.000432594	1.01034e-05	0.0120775	0.0012194	0.000604666	0.00020463	0.00349097
0.79241	3.27184	2.99688e-05	0.00051881	0.000269421	4.63956e-06	0.0153238	0.0128848	1.40856e-06	4.88131e-05	0.00478635
0.62097	4.1751	0.00211471	0.00100807	0.000486428	0.00211671	0.00205153	0.000658479	0.00450952	0.00124461	0.00319531
0.42609	6.08466	5.46505e-06	6.54647e-05	2.88578e-05	0.0148956	0.0337406	0.0601422	0.00421604	0.000440093	0.0010019
0.17353	14.9399	0.004917	0.0199108	0.0333675	0.0555906	0.0148881	0.20478	0.00504566	0.000249273	0.0155113
0.14479	17.9052	0.007594	0.00014634	0.000888488	0.0144984	0.00480221	1.57516e-07	0.288534	0.0773924	0.000322505
0.04335	59.8029	0.057749	0.686319	0.43776	0.146732	0.123094	0.46624	0.695844	0.222117	0.410552
0.0324	79.8426	0.927539	0.291936	0.526701	0.766003	0.793754	0.253576	0.0010086	0.698222	0.561043

Figure 3 shows the size chart of each of the 8th and 9th row variables of the principal components of Table 2

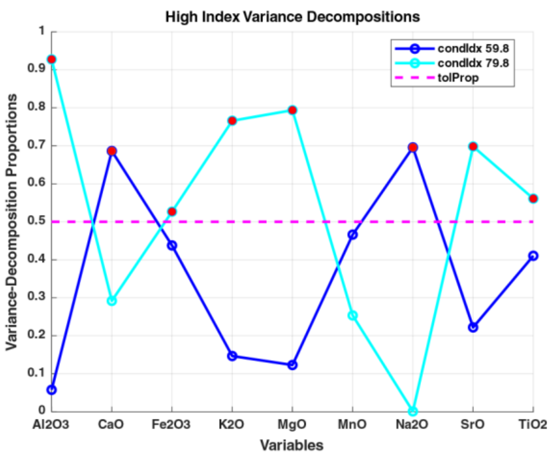

Fig. 3. The size chart of each of the 8th and 9th row variables of the principal components of Table 2

In Figure 3, the bold blue of the first component, the light blue of the second component, and the red dashed line of the confidence area indicate the variance of the data. In the first component, the two variables CaO and NaO2 have increased equally, and in the second component, the variables Al2O3, MgO, K2O, MgO, SrO and TiO2 are correlated with each other. The negative correlation between Al2O3 and Na2O indicates that sodium is not present in the structure of aluminosilicates and the abundance of this element is probably controlled by the content of clay minerals such as kaolinite and elite and can be determined by the ratio K2 to Al2. Jin et al, 2006; Adel et al., 2008))

Linear regression model

Simple linear regression measures the effect of an independent variable on a dependent variable and measures the correlation between them, as in Eq. 2, where χ is an independent variable and y is a dependent variable.

$$Y=\alpha+\beta\chi+\delta \tag{1}$$

In the previous equation α is a random variable with a normal distribution that represents the width of the origin of the regression line; β is the slope of the regression line and σ is a random variable called the residual component or random effect. (Rencher, 2012)

Multiple linear regression model for variable (MLR_ MnO)

Considering that in the present study, some of the studied samples had no MnO values and on the other hand, there was a significant relationship between MnO and Fe2O3, which are both unstable oxides during weathering and diagenetic processes, and the linear multiple regression model. They are analyzed based on Equation 2 and the results are shown in Table 2. The Value column is the value that each variable can have to predict. The more positive and large this value is, the greater the correlation between the two variables. And the greater the distance between the two tests T and P, the greater the correlation.

$$y_t = c + X_1\beta_1 + \varepsilon_t \tag{2}$$

Table 2. The result of evaluating the linear multiple regression of the MnO variable relative to other variables

Parameter	**Value**	**Standard Error**	**t Statistic**	**P-Value**
Intercept	0.071248	0.049939	1.4267	0.17917
Beta{Fe2O3}	0.11326	0.022676	4.9946	0.00031211

Multiple linear regression model for variable (MLR _ SrO)

In many common regression problems, the input is multivariate. Because Al2O3 is relatively unchanged during weathering, diagenesis, and metamorphism, it is commonly used as a factor for comparison between different rocks (Cardenas et al., 1996). And MgO, CaO are immobile (Bauluz et al. 2000)

After modeling, the relationship 4 between the SrO variable with Al2O3 and Na2O is obtained and the regression results are shown in Table 3, where the Value column corresponds to the value that each variable can have to predict. The more positive and larger this value is, the greater the correlation. There are two variables. And the greater the distance between the two tests T and P, the greater the correlation

$$y_t = c + X_1\beta_1 + X_2\beta_2 + \varepsilon_t \tag{3}$$

Table 3. The result of evaluating the linear multiple regression of the SrO variable relative to other variables

Parameter	Value	Standard Error	t Statistic	P-Value
Intercept	0.35015	0.11004	3.1819	0.009788
Beta{Al2O3}	-0.020169	0.020396	-0.98883	0.34607
Beta{Na2O}	0.17238	0.098228	1.7549	0.1098

Multiple linear regression model for variable (MLR_TiO2)

As mentioned in the previous section, some data of the TiO2 variable are unknown, so the Fe2O3 variable has been used in forecasting and evaluation. After fitting, Equation 4 is established for them and the evaluation results according to Table 4 show that the Value column is related to the value that each variable can have to predict. The more positive and larger this value is, the greater the correlation between the two variables. And the greater the distance between the two tests T and P, the greater the correlation.

$$y_t = c + X_1\beta_1 + \varepsilon_t \tag{4}$$

Table 4. The result of evaluating the linear multiple regression of the TiO2 variable compared to other variables

Parameter	Value	Standard Error	t Statistic	P-Value
Intercept	0.0016242	0.081622	0.019899	0.98437
Beta{Fe2O3}	0.14218	0.028389	5.0082	0.00012872

Modeling of binomial linear dual regression for the variable SrO. In this model, the fit of changes of one variable to another is measured to obtain a more accurate prediction. So fitting the curves of the two variables to each other, the formula f (x) = p1 * x ^ 2 + p2 * x + p3 with the coefficients of Table 6 and 95% correlation, the root mean square standard error (RMSE: 0.04976), the degree of freedom of the correlation square (Adjusted R-square: 0.9745), sum of error squares (SSE: 0.02228) and correlation square (R-square: 0.9791). The relationship between the two variables is shown in Figure 4, which is fitted by binomial modeling. In this study, the blue line is the best type of regression without two variables, which are drawn by removing the heterogeneous data of red dots.

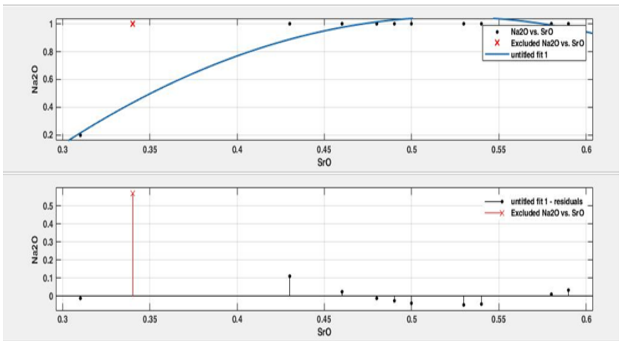

Fig. 4. The relationship of the variable Na2O with SRO

Modeling of binomial linear dual regression for the TiO2 variable. In this model, the fit of changes of one variable to another is measured to obtain a more accurate prediction. So fitting the curves of the two variables to each other, the formula f (x) = p1 * x ^ 2 + p2 * x + p3 with the coefficients of Table 5 and 95% correlation, the root mean square standard error (RMSE: 0.5336), the degree of freedom of the correlation square (Adjusted R-square: 0.9332), the sum of the error squares (SSE: 3.702) and the correlation square (R-square: 0.9421). The relationship between the two variables is shown in Figure 5, which is fitted by binomial modeling. In this study, the blue line is the best type of regression without two variables, which are drawn by removing heterogeneous red dot data.

Table 5. Curve fitting formula coefficients

p1	-4.227 (-8.649, 0.1942)
p2	10.85 (6.664, 15.04)
p3	0.1588 (-0.3619, 0.6795)

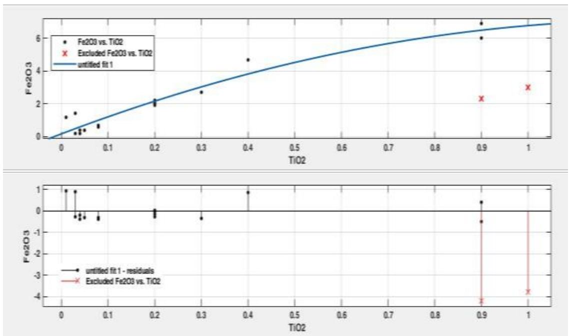

Fig. 5. The relationship of the variable Fe2O3 with TIO2

Modeling of binomial linear dual regression for the MnO variable

In this model, the fit of changes of one variable to another is measured to obtain a more accurate prediction. So fitting the curves of the two variables to each other, the formula f (x) = p1 * x ^ 2 + p2 * x + p3 with the coefficients of Table 6 and 95% correlation, the root mean square standard error (RMSE: 0.04976), the degree of freedom of the correlation square (Adjusted R-square: 0.9505), the sum of the error squares (SSE: 1.485) and the correlation square (R-square: 0.9595). The relationship between the two variables is shown in Figure 6, which is fitted by binomial modeling. In this study, the blue line is the best type of regression without two variables, which are drawn by removing heterogeneous red dot data.

Table 6. Curve fitting formula coefficients

p1	12.5 (8.706, 16.29)
p2	-4.606 (-7.96, -1.252)
p3	0.8498 (0.3876, 1.312)

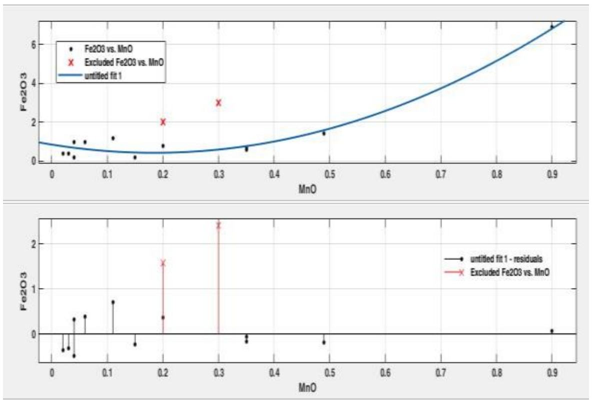

Fig. 6. The relationship of the variable Fe2O3 with MnO

Conclusion

According to the performed fits, TiO2 and Fe2O3 in univariate fitting, even with the removal of heterogeneous data, is not reliable due to high error (SSE: 3.70), but in multivariate regression, a good relationship can be considered because of low error (0.028389) and ratio. They have an acceptable P / T.

In predicting the SrO variable, Na2O was used for univariate fitting and also Al2O3 in multivariate fitting, in both cases the curve fitting is acceptable due to its low error (SSE: 0.02228). But in multiple regression, another variable with more known data should be used for better fit.

In MnO fitting compared to Fe2O3 by removing heterogeneous data the fitting error (SSE: 1.485) decreases and the correlation increases (R-square: 0.9595) but in multivariate regression better performance was observed with high P / T ratio and low error.

Finally, according to the obtained results and analyzes, it can be estimated that up to the curve fitting method with polynomial functions, the data are completely accurate and uniform, and for the elements that need to be predicted. Is more responsive cannot be responsive and should be deleted heterogeneous data in this case should use MLR correlation methods and data of another variable that can be related to other existing elements in predicting the data Got help.

References

Cerar Sonja , Mezga. K , Zibret G , Urbanc J, (2018), Comparison of prediction methods for oxygen-18 isotope composition in shallow groundwater. Science of the Total Environment 631:358-368.

Bauluz, B., Mayayo, M. J., Fernandez-Nieto, C. and Gonzalez Lopez, J. M., 2000- Geochemistry of Precambrian and Paleozoic siliciclastic rocks from the Iberian Range (NE Spain) : Implications for source-area weathering, sorting, provenance, and tectonic setting. Chemical Geology, v. 168, p. 135–150.

Das, B. K., Al-Mikhlaf, A. S. and Kaur, P., 2006- Geochemistry of Mansar Lake sediments, Jammu, India: Implication for source-area weathering, provenance, and tectonic setting. Journal of Asian Earth Sciences, v. 26, p. 649–668.

Rencher, Alvin C.; Christensen, William F. (2012), "Chapter 10, Multivariate regression – Section 10.1, Introduction", Methods of Multivariate Analysis, Wiley Series in Probability and Statistics, 709 (3rd ed.), John Wiley & Sons, p. 19, ISBN 978-1-118-39167-9, archived from the original on 15 June 2019, retrieved 25 September 2018.