Journal of Geochemical Exploration 145 (2014) 207–212

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Hydrothermal alteration mapping through multivariate logistic regression analysis of lithogeochemical data

Ahmad Reza Mokhtari *

Department of Mining Engineering, Isfahan University of Technology, Isfahan 8415683111, Iran

ARTICLE INFO

Article history:
Received 18 November 2013
Accepted 14 June 2014
Available online 22 June 2014

Keywords:
Logistic regression
Categorical variable
Hydrothermal alteration
Model fitting

ABSTRACT

Hydrothermal alterations play a key role in the formation of different mineralization types and zones of economic interests. In mineral exploration industry, the identification of such alterations is usually carried out by visual inspection of specimen through mineral assemblage study. In the following study, it is attempted to identify the alteration type mathematically, based on chemical analysis of the samples. However, modeling of alteration types due to its categorical nature requires special techniques that could handle such variables. The present research employs logistic regression analysis to propose a model for the classification of alteration types (argilic and propilitic) over part of the geological–geochemical dataset, collected at the Kuh Panj porphyry copper mineralization. Logistic regression is applied in a stepwise manner, and the final model has successfully classified the samples in the training dataset (90.50% of a correct classification). The model is also examined by the test dataset and it has concluded to an acceptable result, similar to training set with 90% of a correct classification in the differentiation of alteration types. The final model includes a constant and 9 explanatory variables, including: As, Cr, Cu, Fe, Na, Ni, Sc, Tl and Y. The Wald statistic has suggested that the selected variables are significant and the model itself is evaluated to be significant trough chi-square and Hosmer and Lemeshow tests (all in 5% of significance level).

Application of this technique can then effectively be used where powder samples are taken from an ore body. Additionally, training a model where microscopic studies are available can more precisely separate the alteration types, leading to lower cost of exploration program and consequently, more efficient programming at exploration, exploitation and mineral processing stages.

1. Introduction

The formation of metallic ore deposits from hydrothermal solutions originating from cooling magma also involves formation of zones of alteration minerals due to chemical changes in the mineralogy of host rocks (Barnes, 1997; Hedenquist and Lowenstern, 1994). Mapping of alteration zones is a major concern in mineral exploration, as it can define the strategy and orientation of an exploration program. Delineation of alteration zones in 3-D deposit modeling is also an important step for both exploration and exploitation purposes, since it can affect parameters that influence mine design. Mechanical properties of rocks are affected by hydrothermal alterations that, in turn influence mining and mineral processing programs.

Mapping of hydrothermal alteration is usually carried out by visual inspection of samples and identification of alteration minerals for alteration coding; however, when a non-core drilling program is conducted over an ore body, this task gets more difficult, and if powder samples are collected, the situation worsens, as it would be impossible to recognize the types of minerals. In addition, there might be situations where drill cores are just sampled for analysis but alteration logging is not completed over the samples. Moreover, when alteration identification is completed through microscopic studies, extension of the results to the larger sections of a deposit is quite difficult, as the number of thin sections is usually small and the distribution of alteration minerals cannot be visualized in hand specimens. However, the chemical composition of samples can be helpful to identify hydrothermal alteration zones if a proper mathematical model is developed and trained for the classification of different alteration types for an area.

However, construction of a predictive model for variables such as alteration type, which is non-metric in nature, is quite different from usual modeling method of metric variables like elemental concentrations. Modeling of metric variables by different techniques, like regression analysis, is widely studied in different fields; but once the prediction of categorical variables (e.g., alteration or rock type) is an option, the common ordinary regression modeling is not useful. In this case modeling and predictions require special techniques that can handle categorical variables. Among available techniques, logistic regression and discriminant function analysis have been widely used at

* Tel.: +98 9134105024.
E-mail address: ar.mokhtari@cc.iut.ac.ir.

http://dx.doi.org/10.1016/j.gexplo.2014.06.008
0375-6742/

different research areas (Agresti, 2002; Hair et al., 1998; Pohar et al., 2004; Reimann et al., 2008). The latter method requires the fulfillment of preliminary assumptions on independent variables (Hair et al., 1998) that are not usually met in datasets collected from geochemical and geological studies. This research aims on trying to construct a model to predict the hydrothermal alteration types based on geochemical and geological data. Logistic regression has been utilized on geochemical and geological data collected from surficial exploration activities over the Kuh Panj porphyry Cu mineralization located in Kerman province (Iran).

2. Logistic regression

Logistic regression is one of the most important statistical techniques developed for the analysis and classification of categorical variables (Agresti, 2002; Hair et al., 1998; Pohar et al., 2004). It is based on binomial probability theory in which maximum likelihood method is used to build the best predicting function. This function then can be employed to calculate the probability of a case as a member of predefined classes.

Logistic regression has been widely used in different research areas. However, in geological context, it spans application from target selection for mineral exploration to forecasting geological hazards. Logistic regression has been used successfully to separate two sand environments (costal dunes and modern beach sands) based on distribution parameters of particles (Vincent, 1986). Sahoo and Pandalai (1999) have used logistic regression for gold exploration in Archaean schist (India) by integrating structural and geochemical data. This technique is proposed and used for data-driven predictive modeling of mineral potential (Agterberg et al., 1993; Carranza, 2009; Pan and Harris, 2000). Carranza and Hale (2003) utilized this method to delineate areas favorable for gold mineralization in the Baguio district of the Philippines. Lithology, proximity to curve-linear geological features, and gold occurrences were employed as variables in their study for predictive mapping. Logistic regression was also applied by Carranza et al. (2008), to multiple sets of deposit occurrence favorability scores of univariate geoscience spatial data as independent variables and binary deposit occurrence scores as dependent variable, to analyze and select a set of spatially coherent deposit-type locations for data-driven modeling of mineral potential. Logistic regression has also been employed to forecast the probability of geological hazard occurrence (Jing et al., 2007). In addition, Gross and Low (2013) have employed logistic regression to produce a probability map for the prediction of locations with high arsenic content in groundwater in Pennsylvania and in three intrastate regions (USA).

2.1. Model fitting

Assume a binary response variable (Y), as a function of metric variable (X), representing an event that can take two possible situations: event occurred (Y = 1) and event that did not occur (Y = 0). This situation will generate a plot with two separated classes for response variable. However, if for every X value proportion (or probability) of occurrence of the event is calculated; then, the plot of probabilities against independent variable (X) resembles an S-shaped curve as shown in Fig. 1 (Bewick et al., 2005; Hair et al., 1998).

If P denotes the probability of Y = 1, the logistic or logit function will be:

$$\text{logit(P)} = \ln\left(\frac{P}{1-P}\right) = C + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (1)$$

where C is the intercept and $\beta_1, \beta_2 \ldots \beta_n$ are the slopes of the independent variables or the loadings of the equation, representing the contributions of the covariates to the probability of dependent variable occurrence. So, although P varies from 0 to 1, this transformation

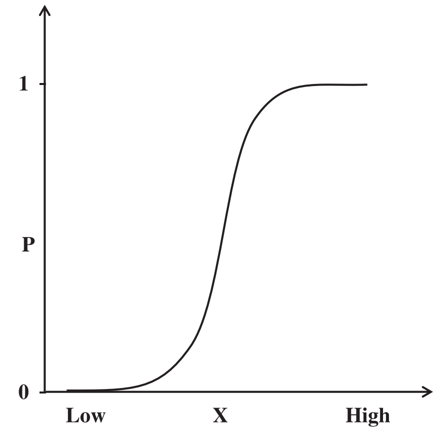

Fig. 1. S-shaped curve displaying the probability of event of interest (P) as a function of an explanatory variable (X).

extends the range of logit (P) from $-\infty$ to $+\infty$, which is asymmetric around 0. Thus, the probability of the target event is obtained from the following equation:

$$P = \frac{e^{C+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n}}{1+e^{C+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n}} \qquad (2)$$

The probabilities of memberships are calculated and they are compared to the cut-off probability value, which is usually considered to be 0.5, and samples with higher P values are classified as group one (Y = 1), otherwise they belong to the other group (Y = 0) (Sahoo and Pandalai, 1999).

2.2. Evaluation of model

Different statistics are proposed to evaluate model efficiency and the significance of explanatory variables in the final model. Inference on the overall fit of the model and its reliability can be carried out by two statistics:

1- Log-likelihood ratio test: −2 log-likelihood (−2LL) represents the fitness of the model to the observed values. This value denotes the smallest deviance (the fit of observed values to the expected values) for the best fitting model through the maximum likelihood algorithm. To carry out the log-likelihood test, the difference between −2LL for the model containing only the constant and −2LL of the model containing explanatory variables are calculated. This will define a new variable having χ^2 distribution.
2- Hosmer and Lemeshow test: a parameter that is used for the evaluation of model by comparing the observed values versus predicted values is the log-likelihood ratio; (Hosmer and Lemeshow, 1989). The model displaying the highest log-likelihood ratio is considered as the most significant model.

In addition to the above tests, two coefficients of determination called pseudo-R^2, are used in the regression analysis of metric variables. They are known as “Cox and Snell” and “Nagelkerke” statistics (R^2) presented in Eqs. (3) and (4), respectively (Menard, 2000; Nagelkerke, 1991).

$$R^2 = 1 - \left\{\frac{L(M_C)}{L\left(M_\beta\right)}\right\}^{\frac{2}{N}} \qquad (3)$$

$$R^2 = \frac{1-\left\{\frac{L(M_C)}{L(M_\beta)}\right\}^{\frac{2}{N}}}{1-L\left(M_\beta\right)^{\frac{2}{N}}} \qquad (4)$$