Interpreting the Geochemistry of Southern California Granitic Rocks using Machine Learning

Germán H. Alférez[1], Jocksan Rodríguez[1], Benjamin Clausen[2], and Lance Pompe[2]
[1]Global Software Lab, Facultad de Ingeniería, Universidad de Montemorelos, Montemorelos, N.L., Mexico
[2]Geoscience Research Institute, Department of Earth and Biological Sciences, Loma Linda University, Loma Linda, CA, USA

Abstract – *Extensive geochemical analyses have been done on granitic rocks in southern California. Almost forty elements were measured for each of several hundred samples. In our previous work, we analyzed the geochemical components of these rocks using two methods, namely Principal Component Analysis (PCA) and Geographic Information Systems (GIS). In this paper, machine learning is used to validate the results previously obtained. We describe an evaluation in which it was found that the results obtained with machine learning are similar to the results obtained by means of PCA and GIS.*

Keywords: Machine Learning, Principle Component Analysis, Geographic Information Systems, Geochemistry.

1 Introduction

A combination of disciplines such as geochemistry and computer science can provide a powerful tool for conducting a thorough study of rocks of interest. Geochemistry helps one to determine the physical conditions under which the rocks formed and the chemical distribution or redistribution of elements over geologic time [1]. Here we are studying the Cretaceous batholithic rocks in southern California [2], which were emplaced in a plate tectonic subduction zone. A batholith (or large granitic body) covers more than one hundred square kilometers in the crust [3, 4].

In previous work [5], we used two approaches to understand the statistical and spatial geochemistry variation of part of the aforementioned area: Principal Component Analysis (PCA) and Geographic Information Systems (GIS). In that data analysis, we used 287 samples from a large systematically collected granitoid geochemical data set [6].

In this work, our contribution is to compare our previous geochemical interpretation of the Californian northern Peninsular Ranges Batholith based on PCA and GIS, and the results from machine learning based on a larger data set with almost 800 samples that comes from a larger area in southern California. This data set includes the 287 samples used for PCA and GIS [6]. We decided to use a larger data set for machine learning analysis to get results as accurate as possible according to our most exhaustive and updated data space.

We believe that our results are of interest to geologists because they demonstrate that analysis of geochemical data with PCA and GIS, as well as machine learning, can elucidate plate tectonic environments. Specifically, in this study we used the Simple K-Means method of machine learning.

This paper is organized as follows. Section II presents the basis of our approach. Section III presents the geochemical analysis by means of machine learning. Section IV presents related work. Finally, Section V presents conclusions and future work.

2 Basis of our approach

In order to understand our approach, it is important to describe the underlying concepts. First, on the one hand PCA is a statistical method based on the variance between variables where high-dimensional data is transformed into low dimensional data. This method can be used to detect coherent patterns [7]. On the other hand, GIS is a way to approximate the values of the discrete sample points over the whole study region, attempting to recreate the continuous geochemical variation that was discretely sampled in the field [8].

In our previous work [5], multivariate outliers were identified using Mahalanobis distance [9], and excluded. Then four components identified by PCA were mapped with GIS to observe their spatial distribution. Bivariate plots relating the component variable to the distance from the transition zone between oceanic and continental crust were used to better understand the trends.

Data were analyzed using PCA with IBM SPSS. Using this method, we were able to reduce 40 geochemical variables to 4 components, which are approximately related to the compatible, High Field Strength (HFS), Heavy Rare Earth (HRE), and Large Ion Lithophile (LIL) elements. The 4 components were interpreted as follows: 1) compatible (and negatively correlated incompatible) elements indicate extent of differentiation as typified by SiO_2; 2) HFS elements