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

indicate crustal contamination as typified by Sr_i; 3) HRE elements indicate source depth as typified by the Gd/Yb ratio; and 4) LIL elements indicate alkalinity as typified by the K_2O/SiO_2 ratio. Note that concentrations for major elements are usually expressed as percent major oxide. Also, Sr_i is a calculated $^{87}Sr/^{86}Sr$ ratio.

Our goal in this paper is to analyze the geochemical data of the southern California granitic rocks using machine learning. Machine learning is a branch of Artificial Intelligence, which studies agents or programs that learn or evolve based on experience to perform a particular task better [10].

There are many machine learning methods for data analysis. One of the most popular is Simple K-Means [11]. Simple K-Means is a clustering technique with a relatively simple implementation. The goal of clustering is to partition a set of objects, which have associated multidimensional vectors of attributes in homogeneous groups (i.e., the "K"); such that patterns in each group are similar.

There are four steps to describe the functionality of Simple K-Means [12, 13]: 1) a set of objects to be partitioned, the number of groups, and each group's centroid are defined; 2) for each object in the data set, the nearest centroid is determined, and the object is added to the group related to that centroid; 3) for each generated group, the centroid is recalculated; and 4) multiple convergence conditions are used. The most common ones are the following: converge when a number of iterations has been reached, converge when there is no exchange of objects among the groups, or converge when the difference among centroids in two iterations is smaller than a given threshold. If the convergence condition is not satisfied, steps two, three, and four are repeated.

3 Geochemical analysis by means of machine learning

In this study, WEKA was used to carry out the geochemical analysis of the southern California granitic rocks [14]. WEKA is a free tool written in Java that has a large number of data analysis techniques, such as preprocessing and clustering. It also facilitates data visualization.

In this section we present the comparison between our previous results with PCA and GIS, and our present results with Simple K-Means for the following geochemical factors: SiO_2, Sr_i, Gd/Yb, and K_2O/SiO_2.

3.1 SiO_2 analysis

Through PCA and GIS, we found that the extent of differentiation is more uniformly high or low in the East and more intermediate in the West.

A trend surface analysis interpolation map of SiO_2 shows the spatial distribution to be high in the far West, low in the West Central, and moderately high in the East. Figure 1 shows the distribution of this oxide. Red areas represent a high concentration of SiO_2 and blue areas show a low concentration. The other colors indicate intermediate concentrations.

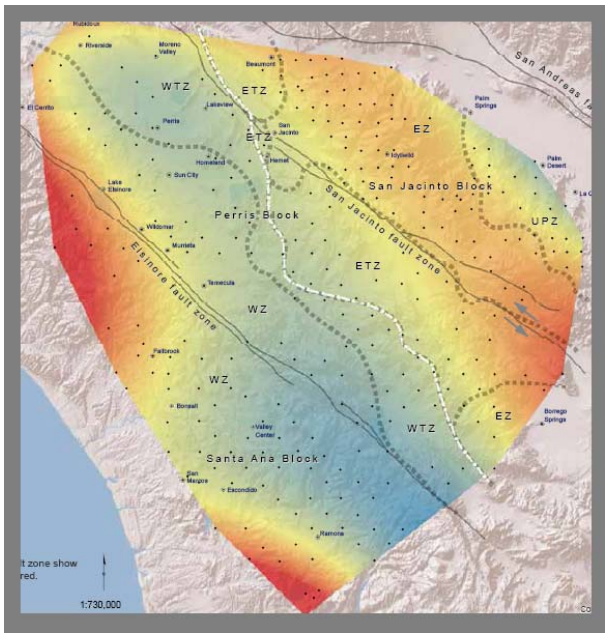

Figure 1. Spatial distribution of SiO_2. The zones in red have a concentration above 70%. The zones in blue have a concentration below 60%

For the same SiO_2 oxide from the larger data set, the results with Simple K-Means can be seen in Table 1. In our experiments, on the one hand we found that within a cluster the sum of squared errors decreases as the number of clusters increases. On the other hand, we found that if a very large number of clusters is generated, then some of them will have a very small number of samples. This fact can produce inconsistent results. We argue that it is important to have a balance between the error and the average number of clusters. In our case, we realized that four clusters gave us the best balance when analyzing SiO_2 and the other geochemical variables.

Table 1. WEKA results for percent SiO_2

Cluster #	Number of samples	Oxide concentration
0	104	54.4%
1	294	63.4%
2	181	73.4%
3	192	68.0%

With the data in Table 1, it was possible to generate Figure 2. The horizontal axis indicates longitude and the vertical axis latitude. Cluster 0 is in blue, Cluster 1 is in yellow, Cluster 2 is in red, and Cluster 3 is in green.

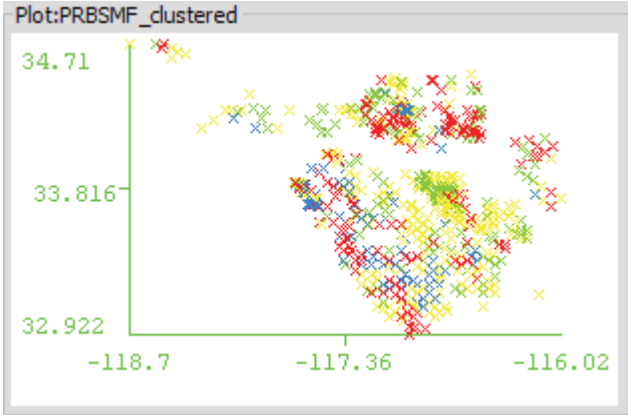

Figure 2. Cluster assignment visualization for SiO_2. Cluster 0 is in blue, Cluster 1 is in yellow, Cluster 2 is in red, and Cluster 3 is in green

A similarity can be observed between the concentration of elements in Figure 1 and the lower half of Figure 2. For instance, Cluster 2 (which is in red) has a high concentration of SiO_2 (73.4%); whereas, cluster 0 (which is in blue) has a low concentration of this oxide (54.4%). These results reflect a similarity with the results in the map of Figure 1.

3.2 Sr_i analysis

The analysis using PCA and GIS on the one hand shows a low Sr_i in the West and an increasing Sr_i to the East. Higher values indicate greater crustal contamination. Figure 3 was generated using kriging interpolation. The blue color represents a low value of Sr_i, whereas the red color shows a high value. Table 2 shows the results with Simple K-Means for this element.

Table 2. WEKA results for Sr_i

Cluster #	Number of samples	Isotope ratio
0	135	0.7091
1	358	0.7068
2	31	0.7126
3	243	0.7042

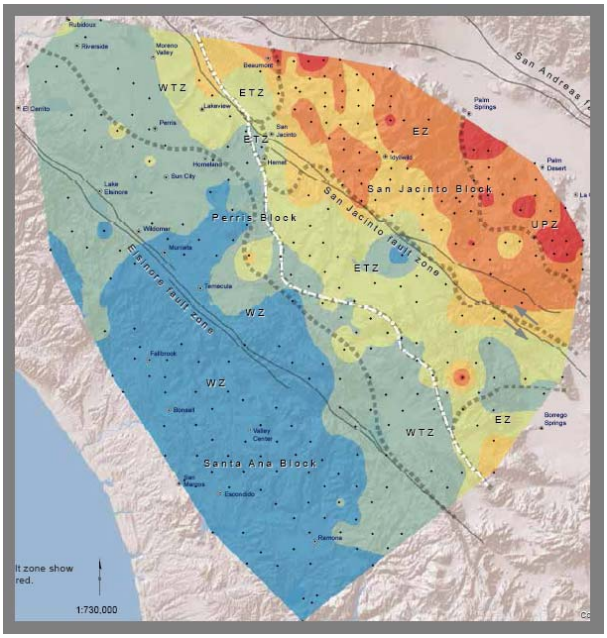

Figure 3. Spatial distribution of Sr_i. The zones in red have a value greater than 0.707 for this variable. The zones in blue have a value less than 0.705

The visual description of the concentration and distribution of Sr_i is presented in Figure 4. Cluster 0 is in yellow, Cluster 1 is in green, Cluster 2 is in red, and Cluster 3 is in blue.

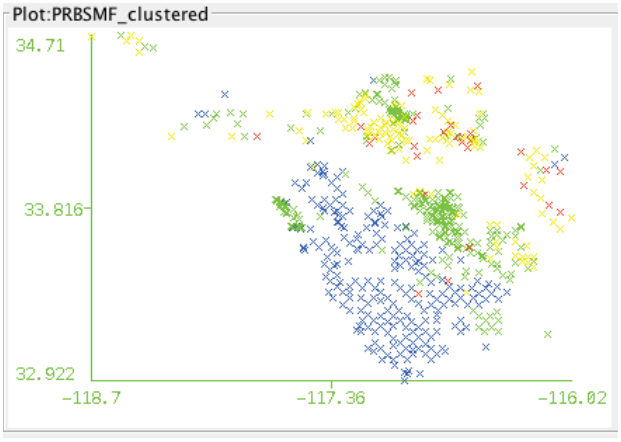

Figure 4. Cluster assignment visualization for Sr_i. Cluster 0 is in yellow, Cluster 1 is in green, Cluster 2 is in red, and Cluster 3 is in blue

Cluster 3 has very low values similar to what is found in Figure 3. Likewise, Cluster 1 has higher values, also similar to what is found in Figure 3. The results reflect a similarity with the results in the map of Figure 3.

3.3 Gd/Yb analysis

According to the experiments with PCA and GIS, Gd/Yb ratios are related to magma source depth (see Figure 5). In this map, the West is uniformly low indicating a shallow magma source depth.

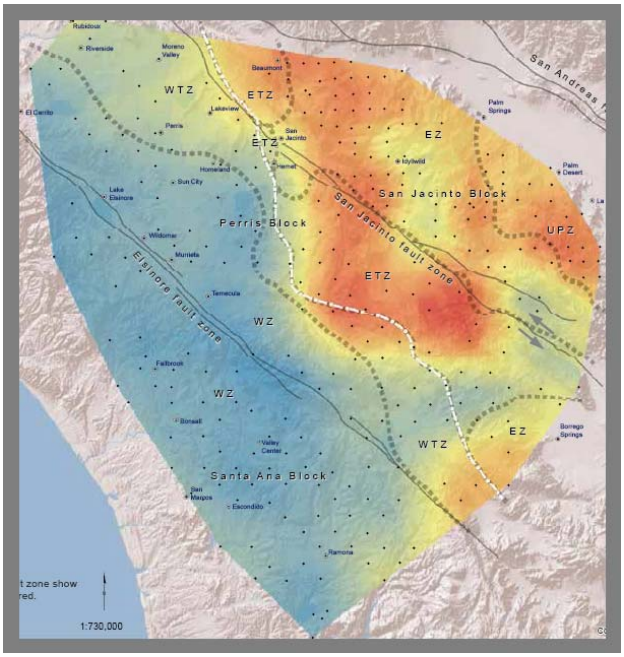

Figure 5. Spatial distribution of Gd/Yb. The zones in red have a high concentration above 2 for this ratio. The zones in blue have a low concentration below 2 for this ratio

The results with Simple K-Means for Gd/Yb are shown in Table 3.

Table 3. WEKA results for Gd/Yb

Cluster #	Number of samples	Element ratios
0	461	2.4
1	96	3.6
2	119	1.8
3	95	1.3

Figure 6 was generated according to the data in Table 3. Cluster 0 is in yellow, Cluster 1 is in red, Cluster 2 is in blue, and Cluster 3 is in green. The bottom half of this map is similar to the one shown in Figure 5. Specifically, Cluster 1 (which is in red) has the highest ratio. In contrast, Cluster 3 (which is in blue) has the lowest ratio.

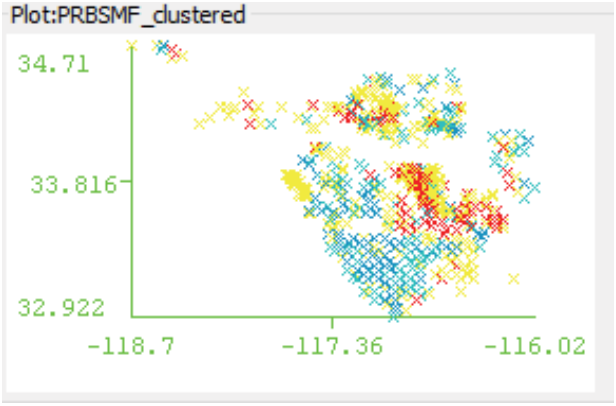

Figure 6. Cluster assignment visualization for Gd/Yb. Cluster 0 is in yellow, Cluster 1 is in red, Cluster 2 is in blue, and Cluster 3 is in green

3.4 K_2O/SiO_2 analysis

According to our study based on PCA and GIS, the map in Figure 7 shows the distribution of K_2O/SiO_2, which indicates alkalinity. The red color represents a larger ratio and the blue color represents a lower ratio. The yellow and orange colors represent intermediate ratios.

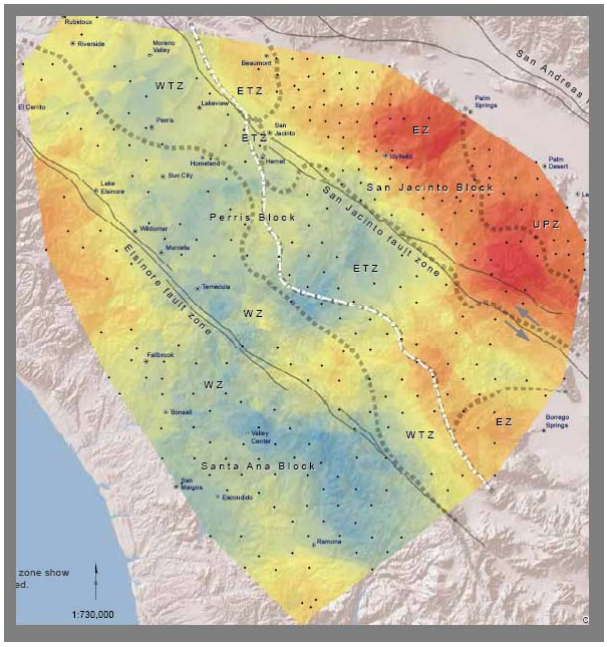

Figure 7. Spatial distribution of K_2O/SiO_2. The zones in red have a high ratio above 0.3. The zones in blue have a low ratio below 0.3

The results with Simple K-Means for K_2O/SiO_2 are shown in Table 4.

Table 4. WEKA results for K_2O/SiO_2

Cluster #	Number of samples	Ratio values
0	277	0.045
1	81	0.007
2	164	0.066
3	249	0.029

The spatial distribution of K_2O/SiO_2 is presented in Figure 8. Cluster 0 is in yellow, Cluster 1 is in blue, Cluster 2 is in red, and Cluster 3 is in orange. High ratios, which are represented in red, are in Cluster 2. In contrast, low ratios in Cluster 1 are in blue. Figure 7 and 8 show similar results.

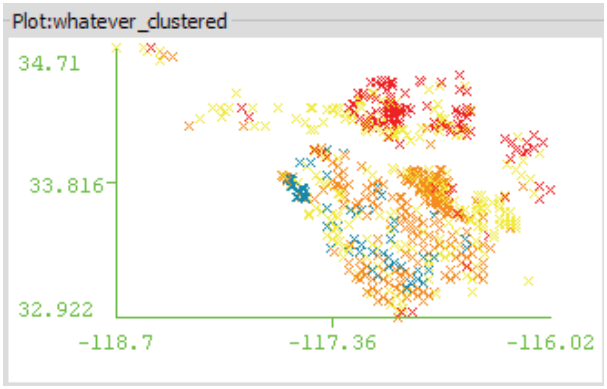

Figure 8. Cluster assignment visualization for K_2O/SiO_2. Cluster 0 is in yellow, Cluster 1 is in blue, Cluster 2 is in red, and Cluster 3 is in orange

4 Related work

Increasingly larger geochemical data sets are becoming available from the geology literature. The purpose for the current research project is to determine what new information can be gleaned from these data sets using statistical analysis, geospatial analysis, and machine learning techniques.

Early geochemical clustering done by Pearce et al. [15] was able to discriminate between granitic-type rocks from different plate tectonic environments just by using pairs of trace elements displayed on bivariate plots. Sr_i values have been used to discriminate between granitic magma sources from the Earth's mantle and the Earth's crust [16, 17]. The ratio between light and heavy rare earth elements has been used to discriminate between granitic magma from shallow and deep sources [18]. Instead of using only two or three elements to group the data into clusters, this research is asking whether it is possible to use PCA, GIS, and machine learning to group large geochemical data sets more effectively and to find new patterns.

Grunsky et al. [19] has been able to classify volcanic rocks into three types using their major element geochemistry. Grunsky and Smee [20] have used PCA and digital topography to visualize, classify, and interpret the geochemistry of 1665 soil samples based on 27 elements. Grunsky [21] used thousands of observations with as many as fifty elements for process identification and pattern discovery using multivariate data analysis and geospatial analysis. Templ et al. [22] used cluster analysis on geochemical data to group samples from northern Europe. Classic books on geostatistical analysis of compositional data include Aitchison [23] and Pawlowsky-Glahn and Olea [24].

Machine learning approaches have shown promising results when applied to complex geological problems involving big data sets. For example, Lüdtke et al. [25] used a supervised machine learning technique to automatically analyze large quantities of spatially referenced seafloor video mosaics of mud volcanoes. Classification accuracy and speed varied between four commonly applied machine leaning classifiers, namely support vector machines, K-nearest-neighbors classifier, C4.5 decision trees, and the naïve Bayes classifier. Classification rates of up to 98.86% were achieved on the full data set with support vector machines when cross-validated with the training data. An average error rate of 1.52% was found when testing the system over a reference data set covering 60% of the investigation area.

Some of the most recent machine learning techniques have been used in discriminating tsunami deposits in Japan [26], predicting acid mine drainage [27], and prospecting for minerals [28, 29].

5 Conclusions and future work

In this paper, we presented an approach to carry out geochemical analysis by means of machine leaning. Specifically, we have focused our analysis on Simple K-Means. We demonstrated that the results with PCA and GIS are similar to the results found with Simple K-Means. This is an important finding because geologists will be able to: 1) use machine learning to validate what they find with statistical tools; or 2) use machine learning to obtain fast results with easily available tools, such as WEKA.

In the future we would like to explore other ways to use machine learning to analyze geochemical data and geological events. For instance, Could we predict possible earthquakes by means of generating forecasts based on historical data?

6 References

[1] F. Albarède, 2003. Geochemistry: An Introduction. Cambridge University Press.

[2] A. K. Baird and A. T. Miesch, 1984. Batholithic rocks of southern California—a model for the

petrochemical nature of their source materials. U. S. Geological Survey Professional Paper 1284.

[3] W. Pitcher, 1993. The Nature and Origin of Granite. Blackie Academic & Professional, pp. 193-217.

[4] A. Hall, 1987. Igneous Petrology. Longman Scientific & Technical, pp. 91-93.

[5] L. Pompe, B. L. Clausen, and D. M. Morton, December 17, 2014. Interpreting the Geochemistry of the northern Peninsular Ranges Batholith Using Principle Component Analysis and Spatial Interpolation. Abstract V33B-4852 presented at 2014 Fall Meeting, AGU, San Francisco, CA, 15-19 Dec. https://agu.confex.com/agu/fm14/meetingapp.cgi#Paper/13737

[6] D. M., Morton, et al., 2014. Framework and petrogenesis of the northern Peninsular Ranges batholith, southern California. In: Peninsular Ranges batholith, Baja California and southern California. D. M. Morton and F. K. Miller, eds. GSA Memoir 211, pp. 61-143.

[7] I. Jolliffe, 2002. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed. Springer, NY, XXIX, pp. 487.

[8] E. C. Grunsky, 2010. The interpretation of geochemical survey data. Geochemistry: Exploration, Environment Analysis 10, pp. 27-74.

[9] P.C. Mahalanobis, 1936. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2 (1), pp 49–55.

[10] T. Mitchell, 1997. Machine Learning, McGraw Hill.

[11] J. MacQueen, 1967. Some Methods for Classification and Analysis of Multivariate Observations. Proc. 5th Berkeley Symp. Math. Statistics and Probability, 1, pp. 28 -297.

[12] T. Kanungo, N. S., Netanyahu, A.Y., and Wu, 2002. An Efficient K-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7.

[13] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y., Wu, (2002). A Local Search Approximation Algorithm for K-Means Clustering. Proc. 18th Annual ACM Symposium on Computational Geometry (SoCG'02), pp. 10-18.

[14] I. H. Witten and E. Frank, 2000. Data Mining. Morgan Kaufmann Publishers.

[15] J. A. Pearce, et al., 1984. Trace element discrimination diagrams for the tectonic interpretation of granitic rocks. Journal of Petrology 25, pp. 956-983.

[16] R. W. Kistler and Z. E. Peterman, 1973. Variations in Sr, Rb, K, Na, and initial Sr87/Sr86 in Mesozoic granitic rocks and intruded wall rocks in central California. Geological Society of America Bulletin 84, pp. 3489-3511.

[17] V. E. Langenheim, et al., 2004. Geophysical and isotopic mapping of preexisting crustal structures that influenced the location and development of the San Jacinto fault zone, southern California. Geological Society of America Bulletin 116(9/10), pp. 1143-1157.

[18] L. P. Gromet and L. T. Silver, 1987. REE variations across the peninsular ranges batholith: implications for batholithic petrogenesis and crustal growth in magmatic arcs. Journal of Petrology 28, pp. 75-125.

[19] E. C. Grunsky, et al., 1992. Characterization and statistical classification of Archean volcanic rocks of the Superior Province using major element geochemistry. In: Geology of Ontario. P. C. Thurston, H. R. Williams, R. H. Sutcliffe, and G. M. Stott, eds. Ontario Geological Survey, Special Volume 4, Part 2, pp. 1397-1438.

[20] E. C. Grunsky and B. W. Smee, 1999. The differentiation of soil types and mineralization from multi-element geochemistry using multivariate methods and digital topography. Journal of Geochemical Exploration 67, pp. 287-299.

[21] E. C. Grunsky, 2010. The interpretation of geochemical survey data. Geochemistry: Exploration, Environment Analysis 10, pp. 27-74.

[22] M. Templ, et al., 2008. Cluster analysis applied to regional geochemical data: Problems and possibilities. Applied Geochemistry 23, pp. 2198-2213.

[23] J. Aitchison, 1986. The statistical analysis of compositional data. New York, Chapman and Hall.

[24] V. Pawlowsky-Glahn and R. A. Olea, 2004. Geostatistical analysis of compositional data, Oxford University Press.

[25] A. Lüdtke, K. Jerosch, O. Herzog, and M. Schlüter, 2012. Development of a machine learning technique for automatic analysis of seafloor image data: Case example, Pogonophora coverage at mud volcanoes. Computers and Geosciences 39, pp. 120-128.

[26] T. Kuwatani, et al., 2014. Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits. Scientific Reports 4; doi: 10.1038/srep07077

[27] G. D. Betrie, et al., 2013. Predicting copper concentrations in acid mine drainage: a comparative analysis of five machine learning techniques. Environmental Monitoring and Assessment 185, pp. 4171-4182.

[28] R. Zuo, and E. J. M. Carranza (2011). Support vector machine: A tool for mapping mineral prospectivity. Computers and Geosciences 37, pp. 1967-1975.

[29] M. Abedi, et al., 2012. Support vector machine for multi-classification of mineral prospectivity areas. Computers and Geosciences 46, pp. 272-283.