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

Geochemistry

Geochemistry helps one to determine:

- The **physical conditions** under which the rocks formed.
- The **chemical distribution** or **redistribution** of elements over geologic time [1].

Area of Interest

- **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].

Northern Peninsular Ranges batholith (PRB) in southern California

Contribution

- To **compare**:
  - Our previous geochemical interpretation of the **Californian northern Peninsular Ranges Batholith** based on **Principal Component Analysis** (PCA) and **Geographic Information Systems** (GIS).
  - The results from **machine learning** (**K-Means**) based on a larger data set with almost **800 samples** that comes from a **larger area in southern California**.

Our Previous Work

1. In our previous work [5], we identified multivariate outliers using **Mahalanobis distance** [9], and excluded.

2. Then **four components, identified** by **PCA**, were mapped with **GIS** to observe their **spatial distribution**.

Our Previous Work (Cont.)

- **PCA** is a statistical method based on the **variance between variables** where **high-dimensional data** is transformed into **low dimensional data** [7].
  - Reduce **40 geochemical variables** to **4 components**.
- **GIS** is a system designed to capture, store, manipulate, analyze, manage, and present all types of **spatial or geographical data.**
  - We **approximated the values** of the **discrete sample points** over the whole study region to recreate the **continuous geochemical variation** that was discretely sampled in the field [8].

Our Previous Work (Cont.)

- Four components were identified:
  - **Compatible:** compatible (and negatively correlated incompatible) elements indicate extent of differentiation as typified by SiO_2 (*Silicon dioxide*).
  - **High Field Strength (HFS):** HFS elements indicate crustal contamination as typified by Sri (*Initial* $^{87}Sr/^{86}Sr$ *ratios*).
  - **Heavy Rare Earth (HRE):** HRE elements indicate source depth as typified by the Gd/Yb (*Gadolinium/Ytterbium*) ratio.
  - **Large Ion Lithophile (LIL) elements:** LIL elements indicate alkalinity as typified by the K_2O/SiO_2 (*Potassium oxide/Silicon dioxide*) ratio.

Geochemical Analysis by Means of Machine Learning

WEKA was used to carry out the geochemical analysis of the southern California granitic rocks [14].

- Free tool
- Written in Java
- Large number of data analysis techniques
- Facilitates data visualization

Geochemical Analysis by Means of Machine Learning (Cont.)

- **Clustering** to group the set of samples by geochemical factor (SiO_2, Sri, Gd/Yb, and K_2O/SiO_2).
    - Samples in the same cluster are more similar to each other than to those in other clusters.
    - **K-Means:** aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

SiO_2 (Silicon Dioxide) Analysis

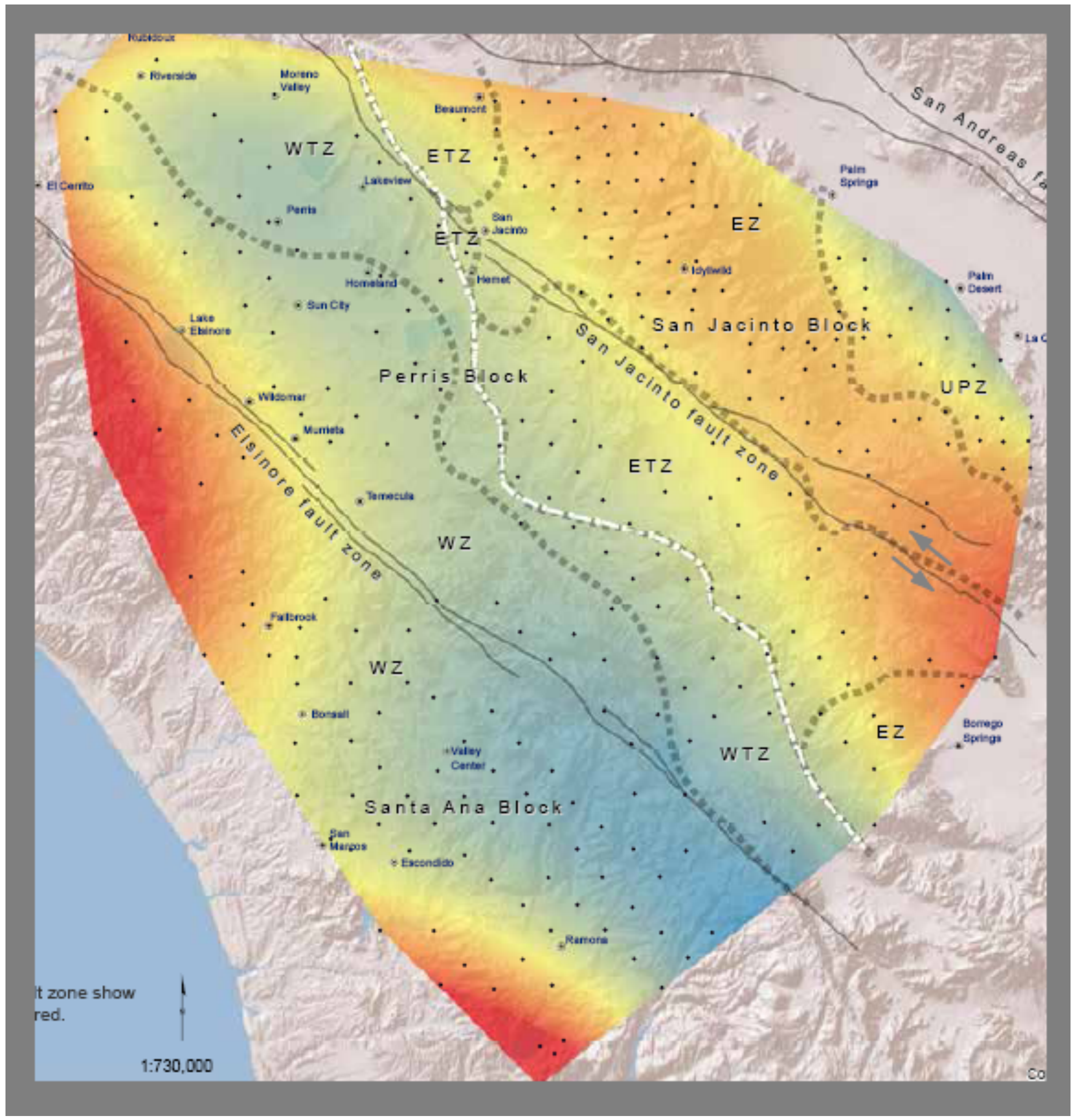

Figure. 1. Spatial distribution and concentration of SiO_2. The zones in red have a concentration above 70%. The zones in blue have a concentration below 60%

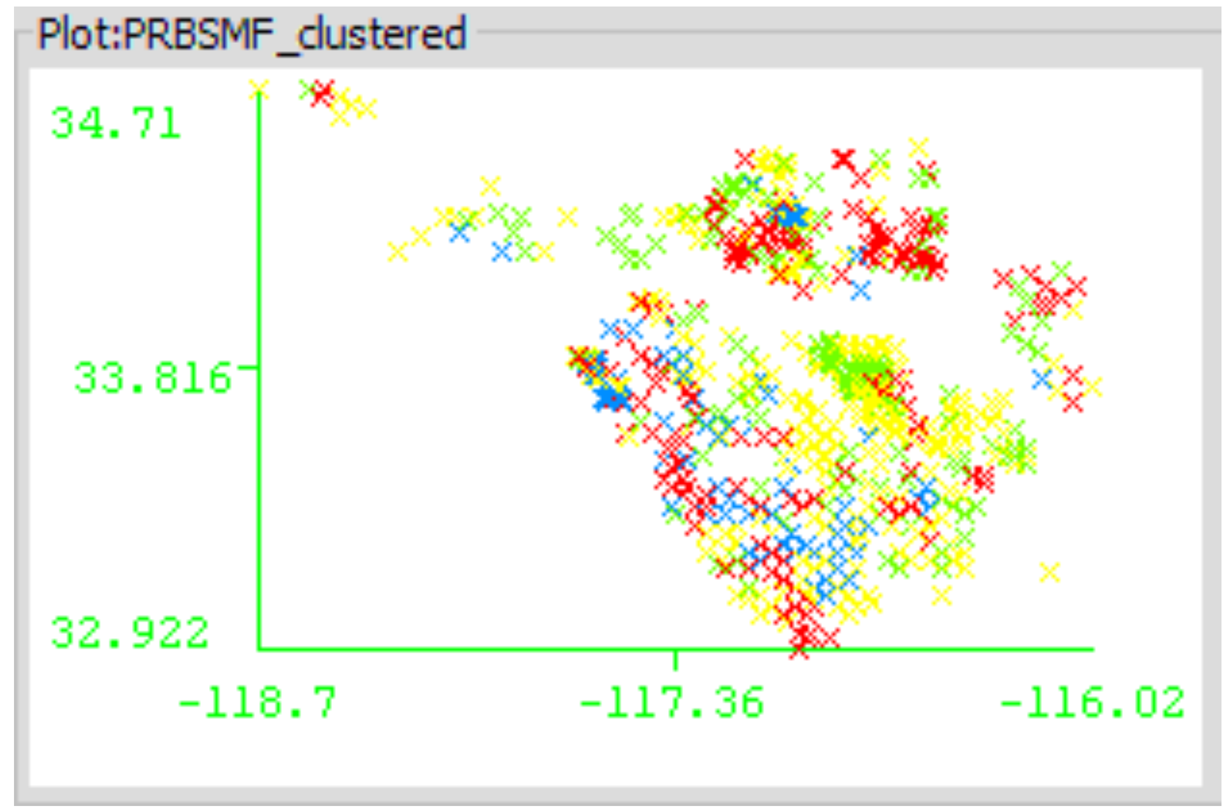

Figure 2. Cluster assignment visualization for SiO2.
Cluster 0 is in blue, **Cluster 1 is in yellow**, **Cluster 2 is in red**, and **Cluster 3 is in green**

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%

Sri (Initial $^{87}Sr/^{86}Sr$ ratios) Analysis

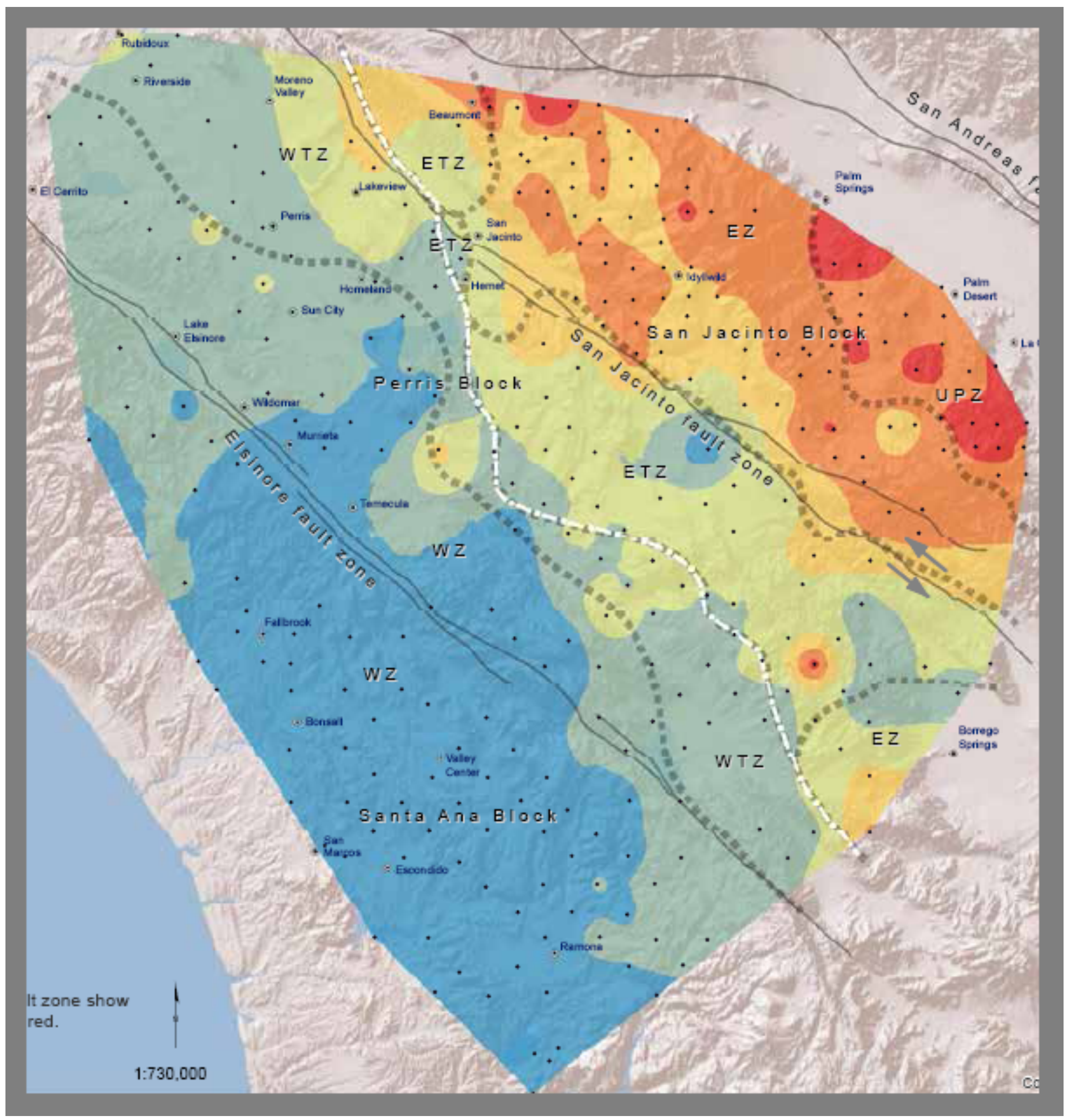

Figure 3. Spatial distribution and concentration of Sri. The zones in red have a value greater than 0.707 for this variable. The zones in blue have a value less than 0.705

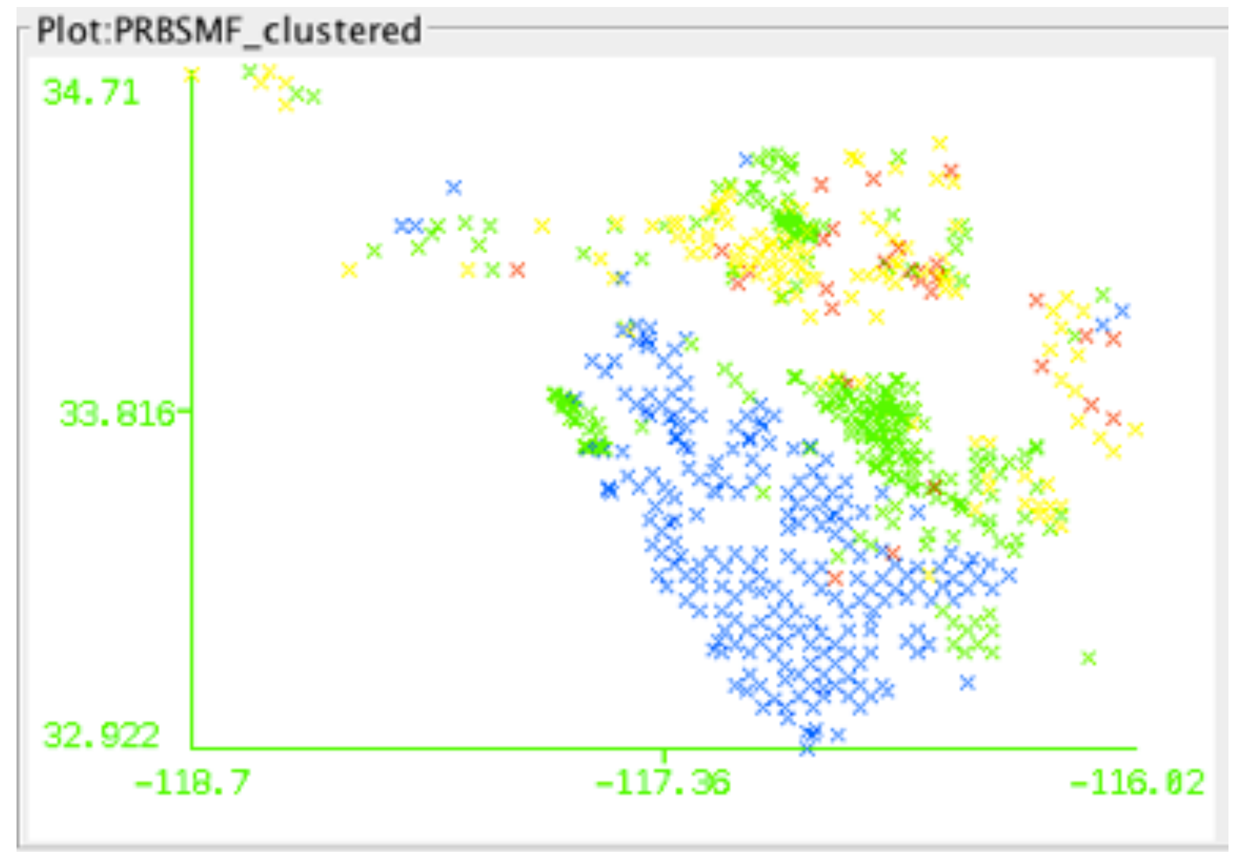

Figure 4. Cluster assignment visualization for Sri. **Cluster 0 is in yellow**, **Cluster 1 is in green**, **Cluster 2 is in red**, and **Cluster 3 is in blue**

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

Gd/Yb (Gadolinium/Ytterbium) Analysis

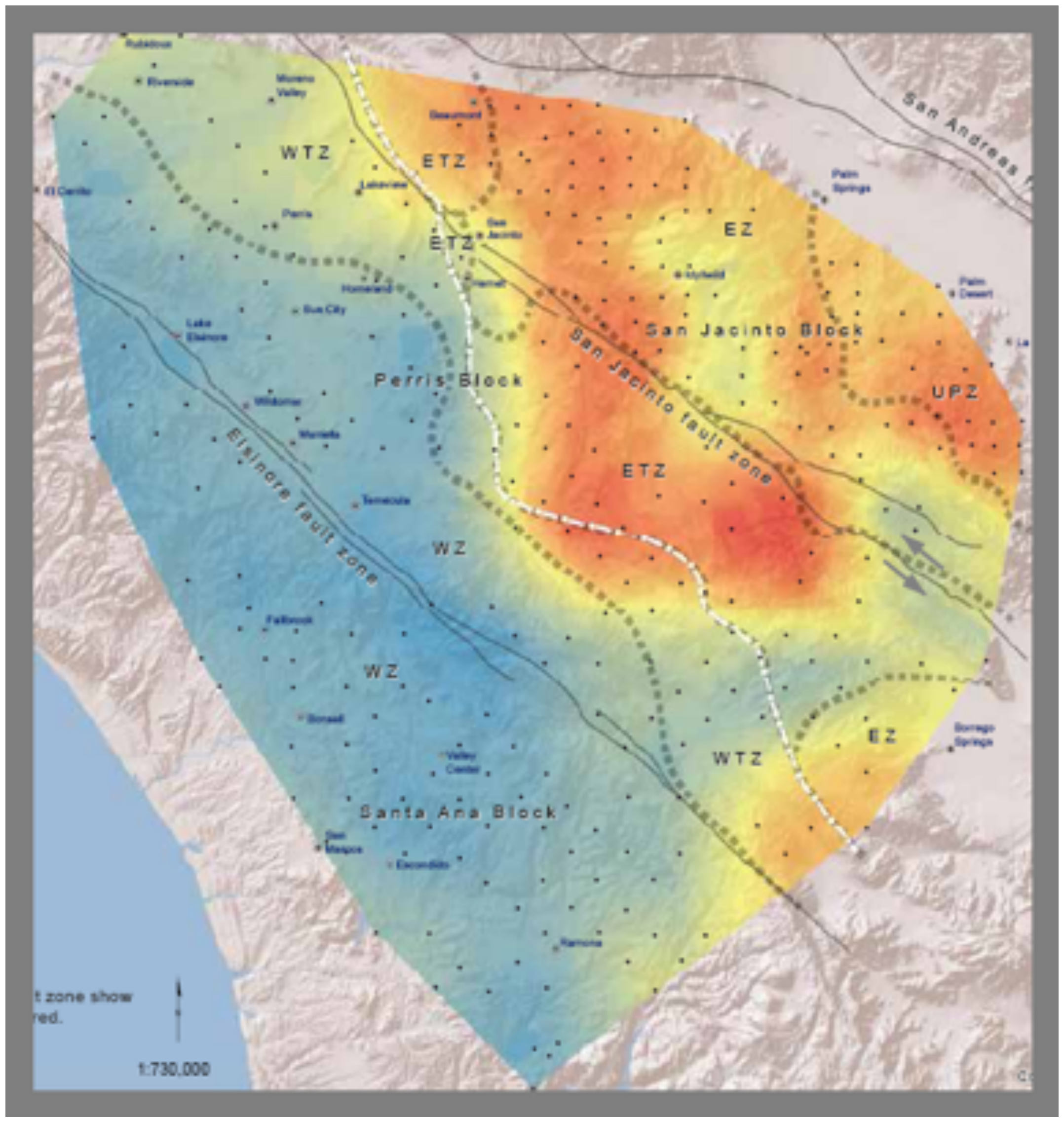

Figure. 5. Spatial distribution and concentration ratios of Gd/Yb. The zones in red have a high concentration above 2 for this ratio. The zones in blue have a low concentration below 2 for this ratio

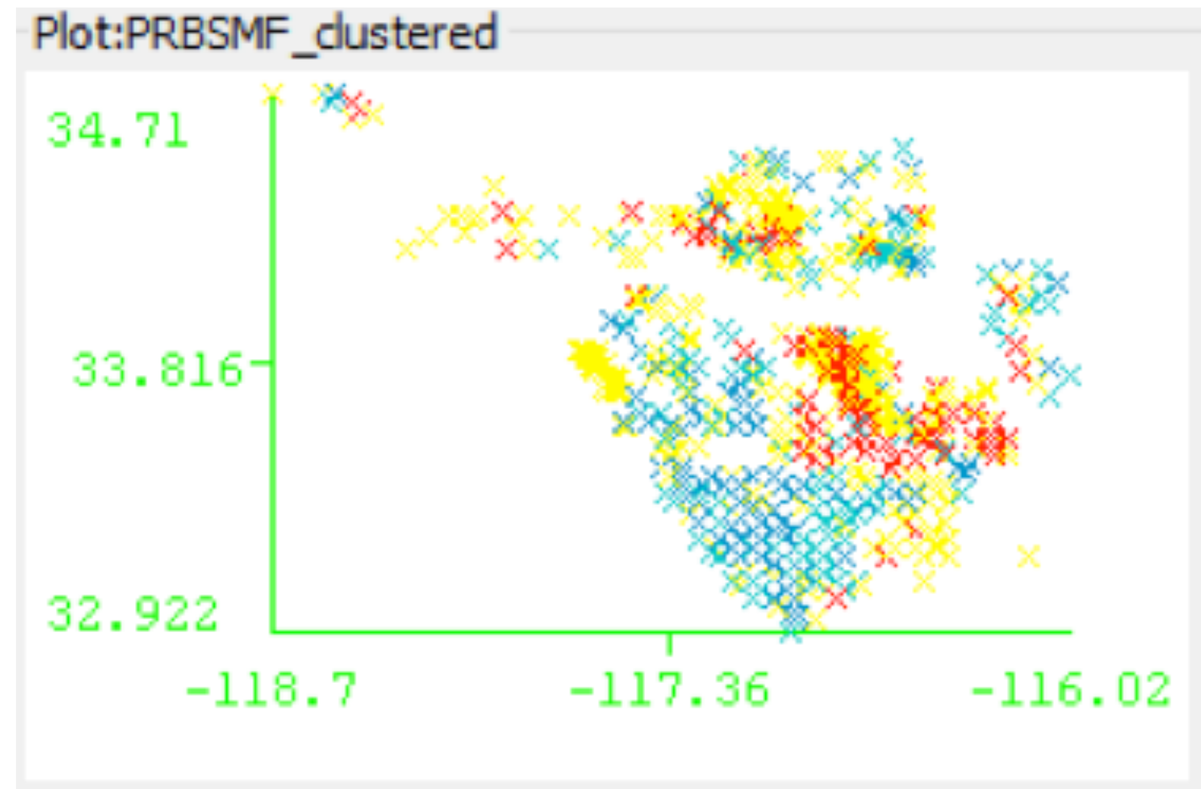

Figure. 6. Cluster assignment visualization for Gd/Yb. **Cluster 0 is in yellow**, **Cluster 1 is in red**, **Cluster 2 is in blue**, and **Cluster 3 is in green**

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

K_2O/SiO_2 (Potassium Oxide/Silicon Dioxide) Analysis

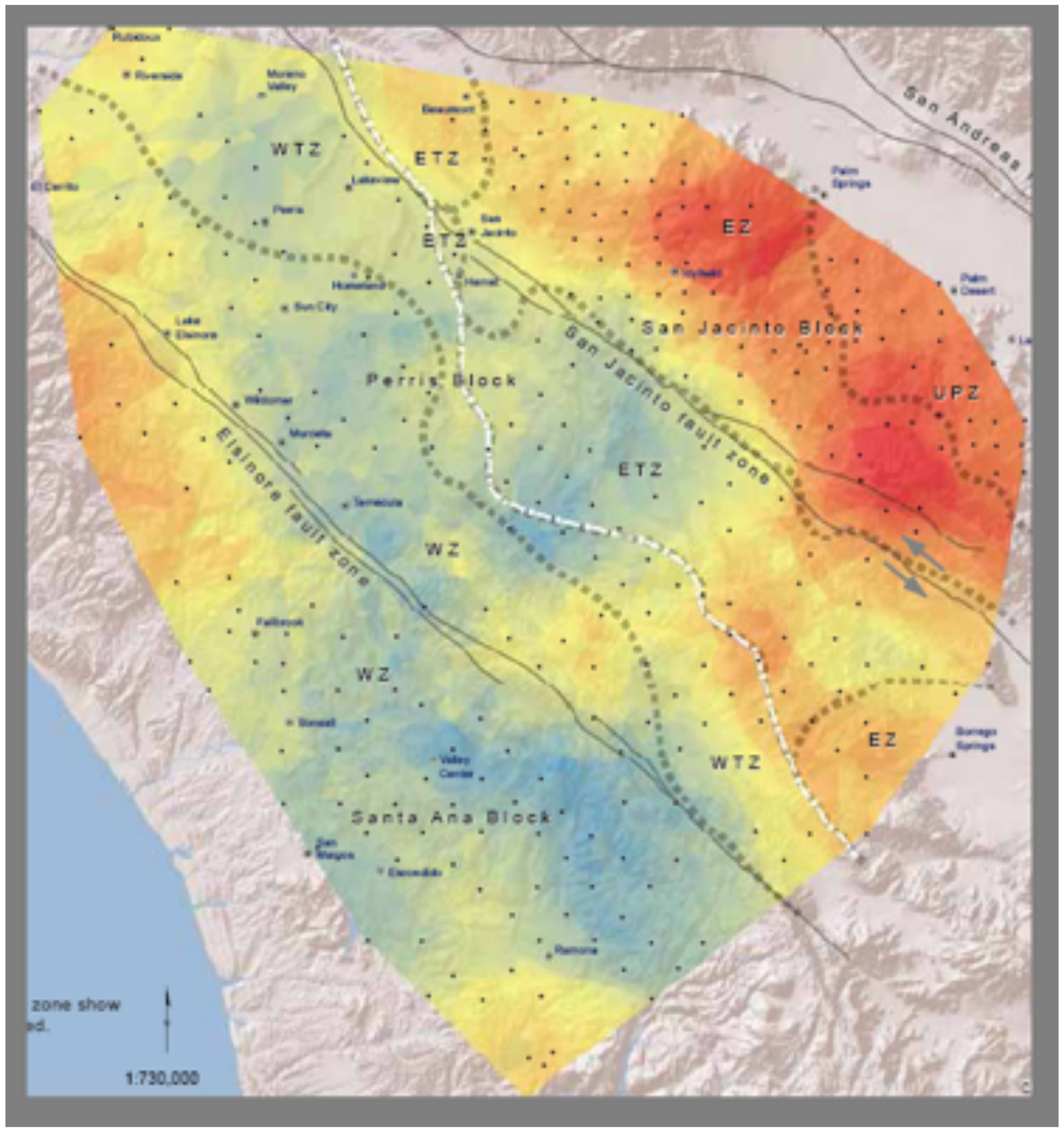

Figure 7. Spatial distribution and concentration of K_2O/SiO_2. The zones in red have a high ratio above 0.03. The zones in blue have a low ratio below 0.03

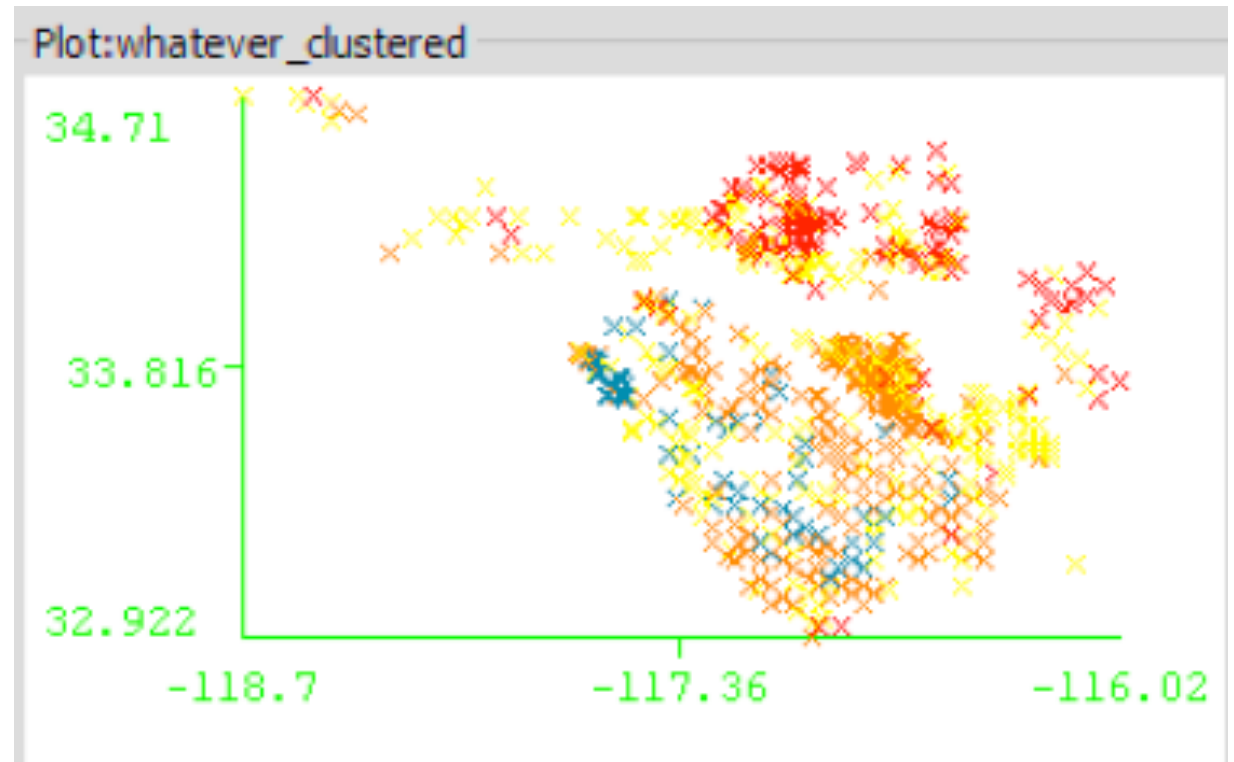

Figure. 8. Cluster assignment visualization for K_2O/SiO_2. **Cluster 0 is in yellow**, **Cluster 1 is in blue**, **Cluster 2 is in red**, and **Cluster 3 is in orange**

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

Related Work

Instead of using only two or three elements to group the data into clusters [15-18], this research used PCA, GIS, and machine learning:

- To group large geochemical data sets more effectively
- To find new patterns

Discri-mination

Related Work (Cont.)

- Some of the most recent **machine learning** techniques have been used in:
  - Analyzing large quantities of spatially referenced seafloor video mosaics of mud volcanoes [25]
  - Discriminating tsunami deposits in Japan [26]
  - Predicting acid mine drainage [27]
  - Prospecting for minerals [28, 29]

Machine Learning

Conclusions

- An approach to carry out **geochemical analysis** by means of **machine leaning**.
    - **K- Means**.
- **We demonstrated that the results with PCA and GIS are similar to the results found with 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.

Future Work

- 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?

Thank you!