Decision Tree Analysis Report

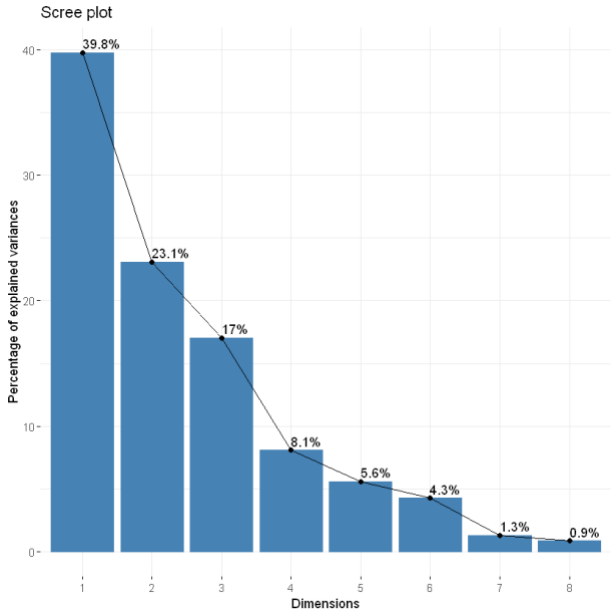
11/22/22 | Miro Manestar

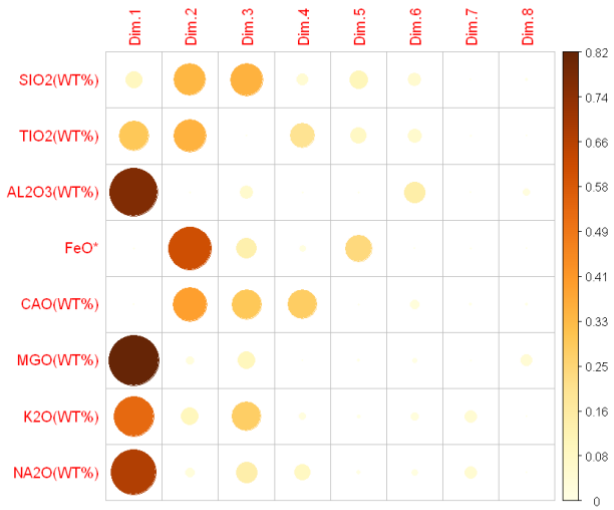
Section 1 – PCA Report

Analysis was done using the primary elements from the Hawaii dataset. Samples with missing data points for any of these 8 variables was excluded. This reduced the total number of usable samples from 12,995 to 2,909 for the purposes of this analysis. No imputation was used at any point during the experiment.

The used elements were as follows:

SiO_2	**Na_2O**
AlO_3	**K_2O**
FeO*	**TiO_2**
MgO	**CaO**





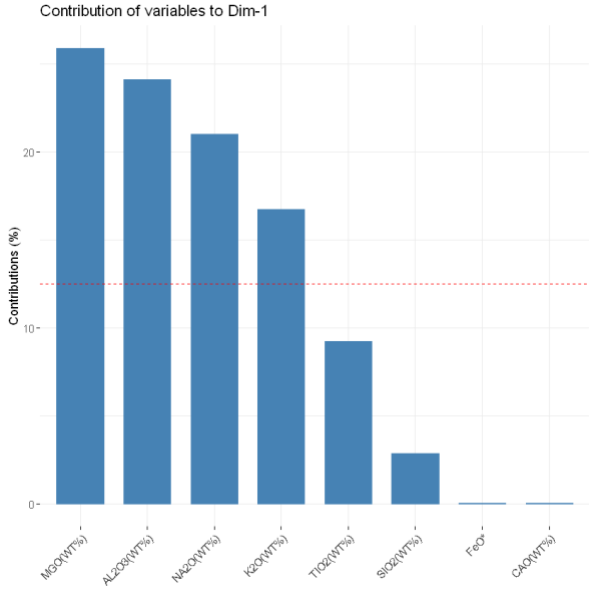
The removal of the elements P_2O_5 and MnO from the dataset only changed the contribution of the first three principal components by 2% on average when compared to the last report, indicating that these two elements were not very useful for determining unique features within the data. Given the drop off in representation by dimension 4, only 3 dimensions were chosen for the subsequent PCA and decision tree analysis. The PCA analysis was modified to include both the variables and individuals plotted, with the clustering done via a kmeans algorithm on the first three dimensions.

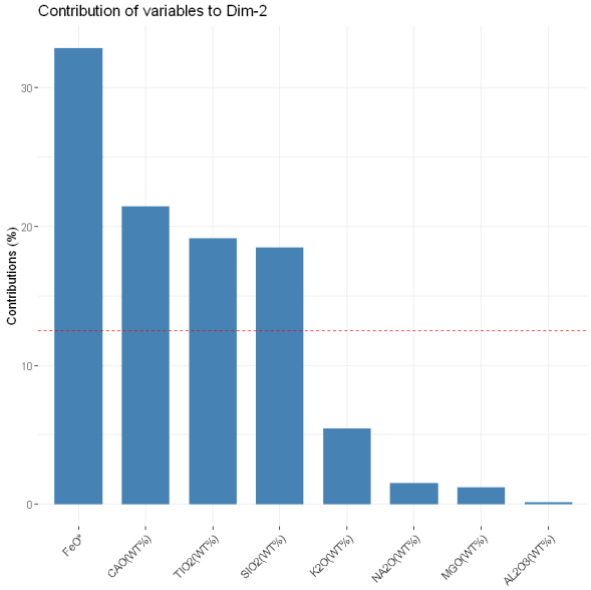
As an aside, clustering based upon which island the sample came from. Examples of this is provided only for dimension 1 as comparisons against subsequent dimensions did not appear to

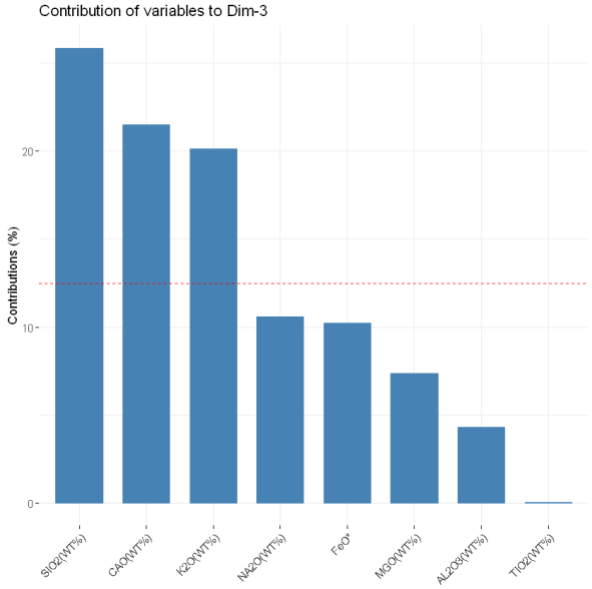
yield any interesting results but was included for posterity. Grouping by volcanoes was also attempted and did not produce better results. This lack of clear grouping would indicate creating a decision tree with either of these classification groups would result in an inaccurate model. So far, this is exactly what has occurred.

The results of the decision tree are gone into more detail in section two after the contribution and PCA analysis.

Contributions

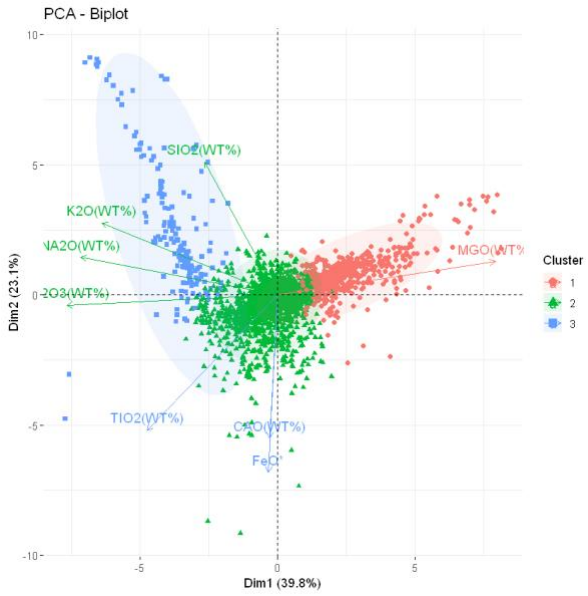




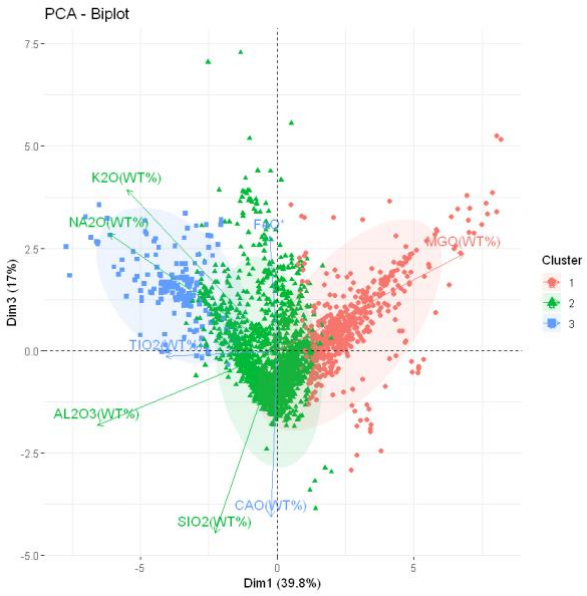


Dimension 1

D1 vs D2

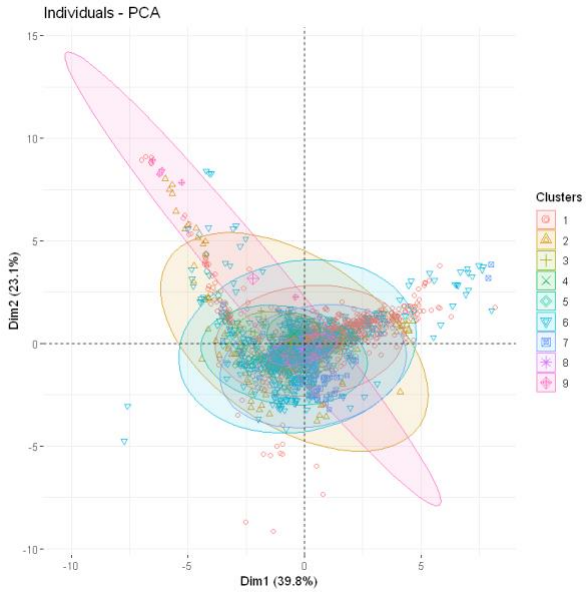


D1 vs D3

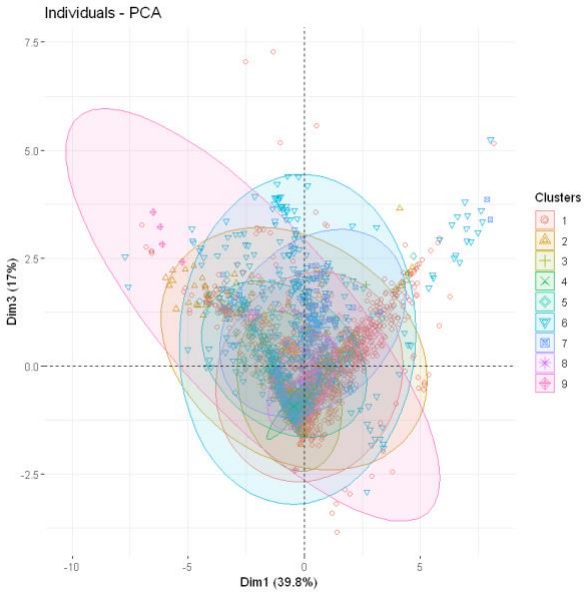


Dimension 1 – Cluster by Island

D1 vs D2

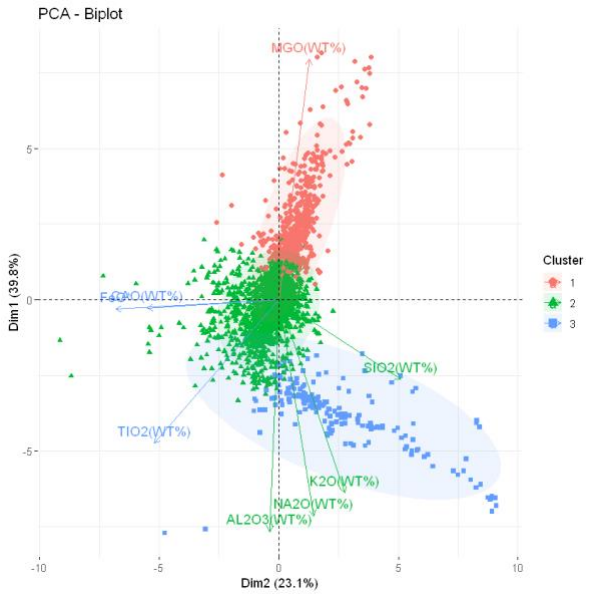


D1 vs D3

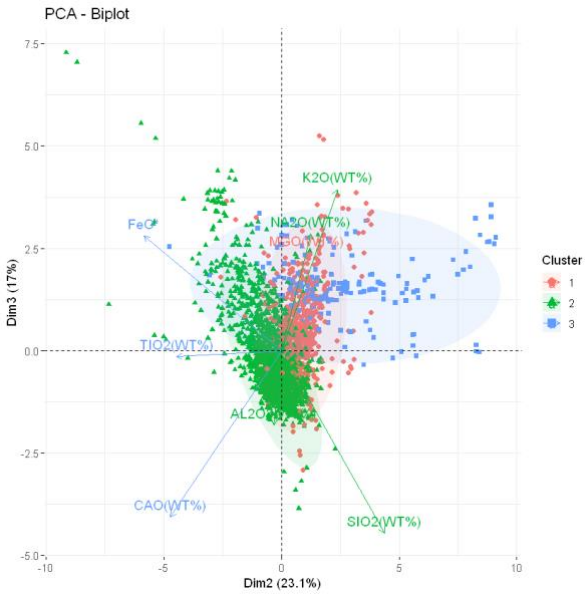


Dimension 2

D2 vs D1

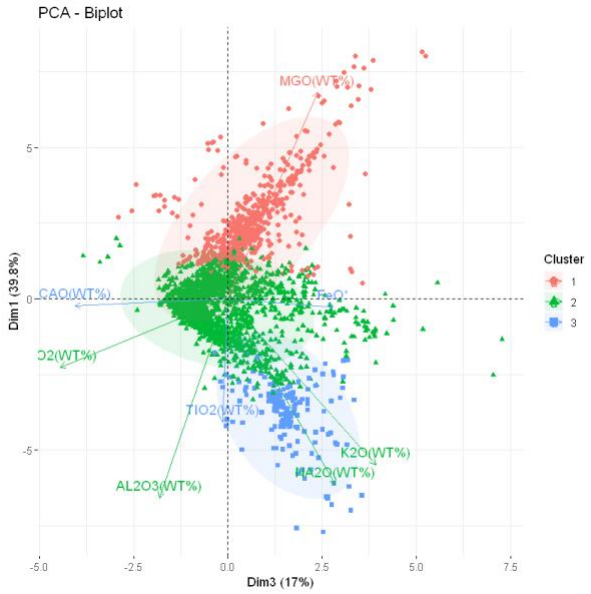


D2 vs D3

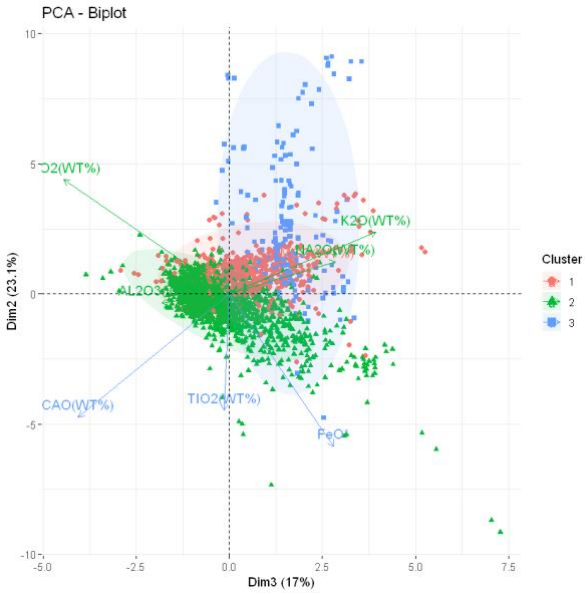


Dimension 3

D3 vs D1



D3 vs D2



Section 2 – Decision Tree Analysis

As predicted by the lack of distinct features between samples clustered in the PCA analysis based on which island they came from, the decision tree used on the same dataset intended to classify samples into either an island or a volcano produced inaccurate models. Given the disappointing results of the Island and Volcano classifier, classification based on the "Kea Loa Trend" was done due to its 4 classes as opposed to the 9 classified islands or 17 classified volcanoes. The resulting trees have been attached as separate PDFs in this report as well.

Island Tree

	precision	recall	f1-score	support
1.0	0.85	0.87	0.86	377.0
3.0	0.42	0.44	0.43	36.0
7.0	0.13	0.5	0.21	4.0
9.0	0.0	0.0	0.0	0.0
4.0	0.57	0.5	0.53	34.0
2.0	0.65	0.57	0.6	97.0
5.0	0.65	0.55	0.59	20.0
8.0	0.38	0.25	0.3	12.0
6.0	1.0	0.5	0.67	2.0
accuracy			0.74	582.0
macro avg	0.52	0.46	0.47	582.0
weighted avg	0.75	0.74	0.74	582.0

Kea Loa Trend Tree

	precision	recall	f1-score	support
2.0	0.0	0.0	0.0	1.0
1.0	0.87	0.84	0.85	317.0
3.0	0.78	0.79	0.79	200.0
0.0	0.49	0.55	0.51	64.0
accuracy			0.79	582.0
macro avg	0.53	0.54	0.54	582.0
weighted avg	0.79	0.79	0.79	582.0

Volcano Tree

	precision	recall	f1-score	support
1.0	0.45	0.5	0.48	10.0
2.0	0.53	0.62	0.57	39.0
3.0	0.79	0.79	0.79	160.0
4.0	0.38	0.26	0.31	19.0
5.0	0.81	0.77	0.79	73.0
0.0	0.64	0.67	0.65	79.0
8.0	0.56	0.62	0.59	24.0
14.0	0.0	0.0	0.0	5.0
16.0	0.33	0.25	0.29	8.0
9.0	0.0	0.0	0.0	0.0
6.0	0.35	0.58	0.44	19.0
7.0	0.56	0.42	0.48	12.0
10.0	0.74	0.56	0.63	70.0
11.0	0.36	0.39	0.38	31.0
15.0	0.44	0.41	0.42	17.0
12.0	0.47	0.5	0.48	14.0
13.0	0.5	0.5	0.5	2.0
accuracy			0.63	582.0
macro avg	0.47	0.46	0.46	582.0
weighted avg	0.65	0.63	0.64	582.0

Please note that the data was shuffled when placed into the tree to increase quality of learning.

The Island model, for the most part, is only useful for determining whether a sample is part of Island 1. Coincidentally, this is also the class with the most samples at 377. A similar trend occurs in the volcano model, in which the model is good at predicting whether a sample is in a particular class only for a handful of the 17 classes. Finally, the Kea Local Trend model is only good for determining whether a given sample is part of 2 out of 4 possible classes. This outcome altogether hints at a lack of distinctive features between the different classes in all three models in most cases. An interesting experiment might be to see if there exists a relationship between the classes themselves.

There does appear to be a correlation between how accurate the model for a given class is and how many samples that class has. For examples, classes 0 and 2 in the Kea Loa Trend model only has 1 sample for class 2 and 64 samples for class 0. Subsequently, it might be a good idea to remove these classes from the decision tree model altogether. In fact, further experimentation with decision trees might involve pruning classes with too few samples as those classes are less likely to generate usable models.

This was only a first in-depth look at classification trees with relation to the geochemical data. Of particular interest going forward is what classes might be more useful for this dataset.