Unsupervised segmentation of the Serra Dourada Granite, Goiás Tin Province, based on Airborne Geophysics

Guilherme Ferreira da Silva[1], Marcos Vinícius Ferreira[1], Iago Sousa Lima Costa[1], Lucy Takehara Chemale[1], Luiz Gustavo Pinto[1].

[1] Geological Survey of Brazil – CPRM

This paper was prepared for presentation during the 17th International Congress of the Brazilian Geophysical Society held in Rio de Janeiro, Brazil, 16-19 August 2021.

Abstract

Manual interpretation of high-dimensional geoscience data can represent a significant challenge since a large volume of data is increasingly overwhelming geoscientists. Dimensional reduction techniques, such as unsupervised segmentation, can reveal high-dimensional patterns within data that can provide meaningful insights to support the geoscientific data interpretation. In this work, we use unsupervised segmentation techniques, such as Principal Component Analysis (PCA) and Model-Based Clustering on radiometric and magnetic airborne geophysical data of the Serra Dourada Granite (SDG), located at the western portion of the Goiás Tin Province, center Brazil. The SDG has an association with greisen mineralization and a late-magmatic enrichment on Rare-Earth Elements (REE), with the formation of supergene deposits of REE on its southernmost portion. Due to the high similarity of the rock's aspects on the outcrop scale, the whole batholith lacks a more accurate mapping, and the division on sub-units remains a problem. We propose here an alternative approach to assist the batholith's cartography through a quantitative analysis over the airborne geophysical data: potassium (K; %), Thorium (eTh; ppm), Uranium (eU; ppm), Total Count (CT; µR/h), and the Total Gradient Amplitude (TGA; nT/m). The features are prepared using Centered Log-Ratio, opening the compositional data, and Min-Max Feature Scaling for variance equalization. Then, we applied PCA to decrease redundancy and create the new orthogonal basis, which improves the segmentation's performance. For cluster analysis, we apply a Model-Based Clustering approach, with the Bayesian Criteria Information function, to automatically identify the most appropriate number of clusters accordingly to a probabilistic determination. Our results allow the individualization of nine significant areas with local coherence and similar multivariate relations by looping through some data segmentation rounds. Additionally, we detail the cluster with the highest values of eTh as it also has relevance for REE prospection. Finally, we evaluate the proposed segmentation by comparing its response to lithogeochemistry data and identifying the multivariate patterns found in every clusters.

Introduction

The Goiás Tin Province (GTP) is a cluster of more than a dozen alkaline granites, mainly mineralized in tin with greisen association, and intrusive on metasedimentary rocks of Ticunzal Formation and Serra da Mesa Group (Figure 1). All bodies are dominantly biotite bearing-granites, sometimes muscovitized, with a very restricted hornblende facies (Marini and Botelho, 1986). Post-magmatic changes like albitization and greisenization are quite common and associated tin deposits occur in albitites, biotitites, greisenized granites, exo and endogreisen, pegmatites and quartz veins (Marini and Botelho, 1986).

The Serra Dourada Granite (SDG) is the largest granitic body of the GTP, with 55 km length and up to 13 km width (Santana et al., 2015). The batholith has an ellipsoidal shape, with a N-S predominant direction and is strongly deformed showing a brachyanticlinal structure with centrifugal foliation (Araujo-Filho et al., 2013). Despite recent efforts and its high outcrop rate, with good quality rock exposure, the batholith lacks a more detailed cartography. This is mainly due to the high similarity of the rocks on outcrop scale, predominantly biotite-granitoids, and any other effort non relying in systematic geochemistry failed to separate the granitic body.

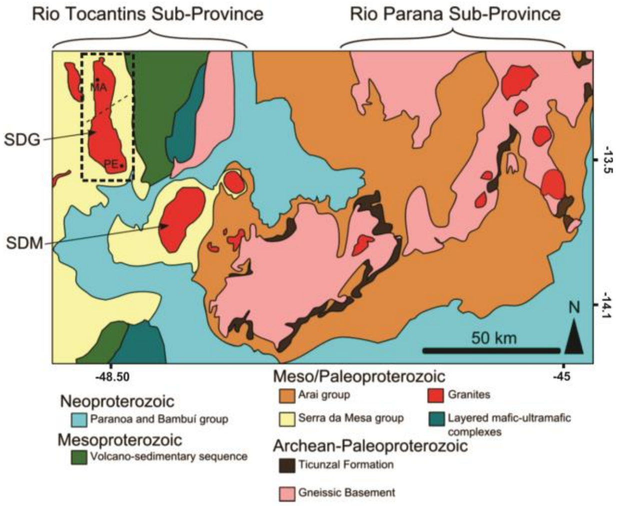

Figure 1: Simplified geological map of the Goiás Tin Province. The SDG is located on the western portion of the GTP (Pimentel and Botelho, 2001; Pinto-Ward, 2018).

Thus, this work presents a quantitative approach to help in the SDG cartography, by using an unsupervised segmentation based on airborne geophysical data. The results presented here can be interpreted as a “pre-field investigation stage”, since the clusters show strong relation and spatial and statistical coherence between radioelements.

Data Source

For the total study area, shown in Figure 1, aerogeophysical data were integrated using the Gridknit methodology of Oasis Montaj. Three projects were used in this integration, they were: “Complemento do Tocantins”, “Arco Magmático de Mara Rosa” and “Paleo-Neoproterozoico do Nordeste de Goiás”. The first acquired by the Geological Survey of Brazil and the others by the state government of Goiás. These projects have spacing between flight lines of 500m and nominsl flight height of 100m. Subsequently, the data was cropped only for the detailed geological body.

The lithogeochemistry data used here for cluster validation was obtained from Polo and Diener (2017). The analysis is partly available at the Geological Survey of Brazil’s (CPRM) database (http://geosgb.cprm.gov.br/).

Data Preparation

The feature data was prepared in order to optimize the processing steps accordingly to some criteria. Radiometric data was first corrected, to remove any negative value founded in the conversion of measured signal to equivalent concentration. To this step, a constant was added to data, in order to level correct negative values without impacting severely the distribution parameters.

As data has different units and scale values, it is needed to normalize the distributions in order keep the same values of ranges, putting the variances in the same order of magnitude. It is specially needed for applying such techniques as PCA, used in this work and described in the following sections. For this purpose, we applied a min-max feature scaling in all input data, for equalizing the range in a 0 to 1 distribution (Figure 2a).

As three of the four channels of Radiometric data corresponds to equivalent concentration of chemical elements (K, eTh and eU), these features can be considered compositional (Aitchison, 1986), as they are ideally positives and interfere on each other, as the ideal concentration of any chemical composes is equal to 100%. This effect causes major problems on traditional multivariate analysis (Aitchison, 2008), and one alternative to mitigate this effect is to open the data by applying log-ratio transformations. For this specific work, we applied a Centered Log-Ratio (CLR) transformation (Figure 2b). To simplify the processing step, we chose to apply the CLR transformation on the other features, as it helps to attenuate the effects of outliers on data distribution, allowing a better perception of the inner variation of background.

For the last step of data preparation, we evaluate the relation between the features and applied PCA. The PCA is a multivariate method based on some algebraical principles (e.g., cross product and inner product of matri-

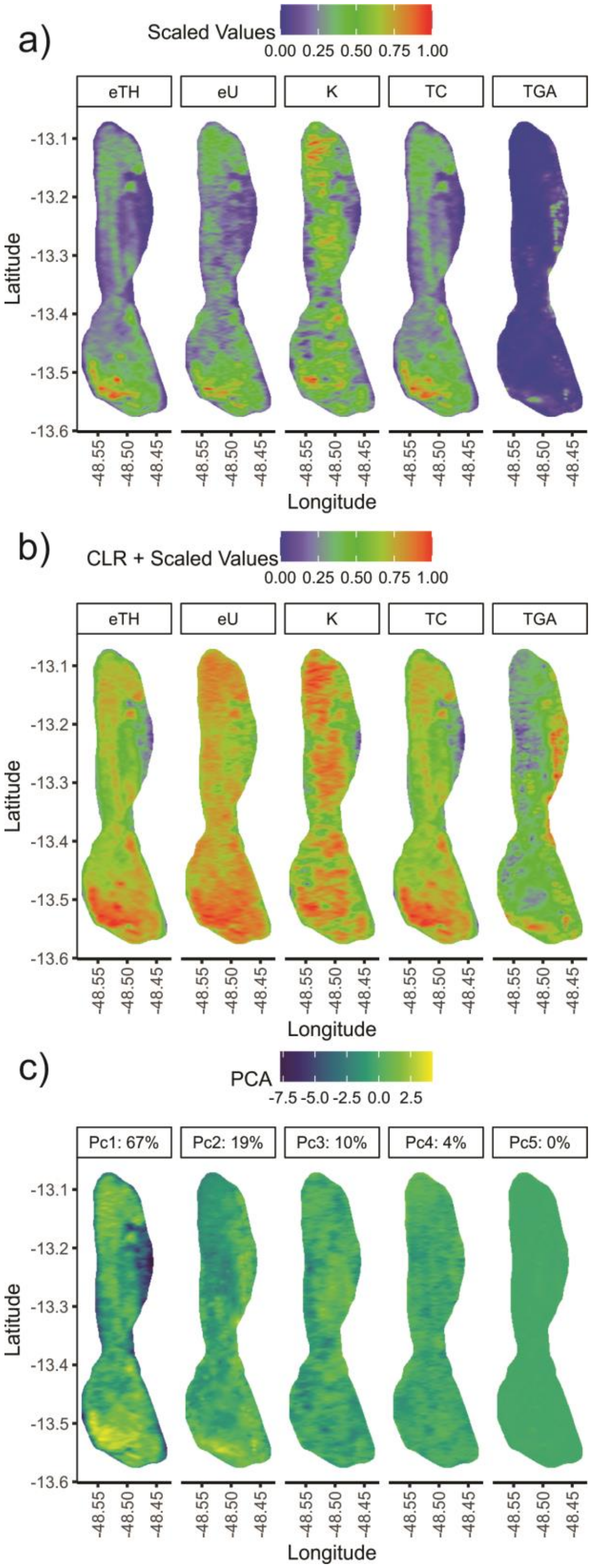

Figure 2: a) Maps of input features with a min-max scaling to a 0-1 distribution; b) Map of input features transformed to centered log-ratio followed by a min-max feature scaling; c) Maps of PCA and percentage of explained variance by component.

-ces) that creates mechanisms that rotates the vectorial space and allows the definition of new features that are orthogonal to each other (eigenvectors or components) and have a particular length (eigenvalues associated with the variance of data) that represents the data with minimum redundance and vector optimization. PCA is commonly used to dimensional reduction (Grunsky and Arne, 2020), as the first components explain more of data variance, the last ones is many times negligible. For this work, all the five components were considered to the clustering processing (Figure 2c).

Unsupervised segmentation

The Model-Based Clustering algorithm (MBC), also known as Gaussian Mixed Models, does not require any hyperparameter imputation or prior information of groups or the distributions. The algorithm assumes that the data are generated by a mixture of underlying probabilities distributions in which each component represents a different group or cluster (Fraley and Raftery, 2002; Landau and Chis Ster, 2010).

The data partition is determined by the Expectation-Maximization function (EM) for maximum likelihood, with initial values defined from agglomerative Hierarchical Clustering (Fraley and Raftery, 2002). After first adjustments, several distribution models are compared using the Bayesian information criterion (BIC) applied to multiple models simultaneously.

The BIC is based on Bayesian factors, which is the posterior odds for one model against other assuming neither is favored a priori (Fraley and Raftery, 2002). The BIC is independent of how the different models are built, changing some constraints about the clusters' shape, volume, and orientation. After comparing the possible models through different numbers of clusters, from 1 to n, the first local maximum is the best suitable model.

Results

For this work, we ran the BIC criteria iteratively, evaluating the results of the clustering by the analysis of bivariate plots. We divided the clustering step into two rounds of segmentation, applying the second round on some clusters that have showed more complexity or are of particular interest. The BIC evaluation is defined to run from 1 to 9 clusters, in order to show the general patterns of the data (Figure 3).

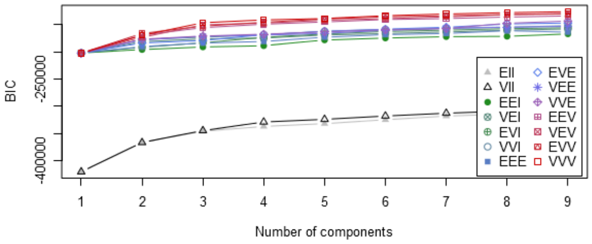

Figure 3: BIC curve for the first round of segmentation showing that data has nine or more suitable clusters. The winner model (VVV) has variable shape (spherical or ellipsoidal), varying volume, and varying direction of major axis.

The analysis based on PCA showed relative complexity, with low separability for all clusters when analyzing just two components at each time. Also, the relation between PC5 to the others returned non-linear patterns that are supposed to be product of the influence of the interaction of radiometric features with TGA (Figure 4).

The statistical analysis of each feature inside all clusters helps the clarify the influence of every feature for the data segmentation (Figure 5).

In the bivariate eTh (ppm) vs eU (ppm) analysis for each cluster (Figure 6), we noticed that the clusters from 1 to 5, including the numbers 8 and 9, have a similar structure, agreeing with a good linear correlation of those two radioelements and showing also moderate to low values of TC. In other hand, clusters 6 and 7 show the highest values of radioelements and a more complex structure, indicating that they are both suitable to a second round of segmentation.

Observing the First-Round clusters' map (Figure 7 a), it is possible to verify that the data segmentation has a high spatial coherence. In general, the points of a given group tend to be surrounded by points of the same classification. Also, some of the patterns found are consistent with relief features mapped in the batholith, such as valleys and alluviums.

The map of the Second-Round (Figure 7 b) clustering shows the areas with highest values of eTh (ppm) and eU (ppm). The cluster d shows negative correlation between K (%) and eU (ppm) and could spot areas of albitization. The other groups are gathered at the southern and eastern portion of the batholith. Together, these subclusters represents around 25% of batholiths area, and may be the most preferred zones to REE prospection, due to the radiometric responses.

The REE contents on lithogeochemistry analysis corroborate that the clusters with the highest eTh values (Clusters 6 and 7) present the most interesting values for REE (Figure 7 c). Despite the REE concentrations are remarkably high (3.000 ppm on average), the highest values are scattered over the southernmost portion of the batholith, over the highlighted area delimitated by the cluster analysis.

Discussion and Conclusions

As we are working mainly with airborne geophysical data of felsic rocks, it is expected that the values of eU and eTh obey a simple linear correlation, which may or may not vary direct proportionally according to the mineralogy inner properties. Th and U have approximately the same ionic charge and ratio, thus they tend occupy similar sites in the crystalline structure of minerals, thus favoring a good correlation on quantitative analysis.

On the other hand, as Th and U are considered “incompatible elements”, they tend to remain on magma until the last stages of crystallization, showing a relative enrichment of concentration for those rocks crystalized in early stages, and then, high values of Th and U may indicate areas of the batholith where the magma is more evolved. Then, the analysis of the response of these two features was used to identify clusters of interests.

Finally, we highlight here some final considerations:

- Generally, all clusters showed two distinct patterns: they presented good linear correlation between eTh and eU, or they showed a more complex structure, indicating that a new round of clustering is acceptable.

- The clusters have good spatial correlation, as they do not tend to be scattered across the batholith, showing some local coherence.

- The features, K, TC and TGA have some importance on the unsupervised model, but they tend impact less than the other two features. The exception is the TGA, which have direct influence to detect clusters 1 and 4, that have respectively low and high magnetic areas.

- At the southernmost portion of the batholith, the clusters 6 and 7 forms a “waning moon shape”. This pattern is remarkably close to a map presented by Pinto-War (2018) that was drawn based on a systematical geochemistry survey.

- In the second round of segmentation, some areas showed a negative correlation between K (%) and eU (ppm). These areas may indicate some hydrothermal process, as albitization.

- Despite it still lacks some field validation, this process has proved to return an interesting product for aiding in a more accurate cartography, or even spotting zones of anomalous patterns (hydrothermal zones?) across the whole batholith area. Yet, the lithogeochemistry data available on the public database of the Geological Survey of Brazil supports both the area potential for exploration and the accuracy of the technique.

Acknowledgments

This work was developed with support of the Geological Survey of Brazil – CPRM. All the quantitative analysis was done in the R language processing, with use of the Tidyverse collection of tools (Wickham, 2014). The clustering analysis was done using the mclust R package (Scrucca et al., 2016). All the data and code used in this work is available at the following Github repository: https://github.com/gferrsilva/GoiasTinProvince.

References

Aitchison, J., 2008. The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. CoDaWork 2008 1–28.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Stat. Anal. Compos. Data 44, 139–177. https://doi.org/10.1007/978-94-009-4109-0

Araujo-Filho, J.O., Silva, G.F., Ferreira, V.N., Prado, E.M.G., Lima, E.A.M., Braga, A.A., Zedes, A.L., Toledo, C.L.B. de, Silva, V.S., Borges, W., Carmelo, A.C., Almeida, T. de, 2013. Geologia e características estruturais do Projeto Mata Azul (GO), Faixa Brasília Setentrional., in: Anais Do XIII Simpósio de Geologia Do Centro Oesteeologia Do Centro Oeste. Sociedade Brasileira de Geociências, Cuiabá, MT, pp. 1–5.

Fraley, C., Raftery, A.E., 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. J. Am. Stat. Assoc. 97, 611–631. https://doi.org/10.1198/016214502760047131

Grunsky, E.C., Arne, D., 2020. Mineral-Resource Prediction Using Advanced Data Analytics and Machine Learning of the QUEST-South Stream-Sediment Geochemical Data , Southwestern British Columbia (Parts of NTS 082 , 092), in: Geoscience BC Summary of Acitivities 2019: Minerals, Geoscience BC, Report 2020-01. Geoscience BC, Vancouver, BC, pp. 55–76.

Landau, S., Chis Ster, I., 2010. Cluster Analysis: Overview, in: International Encyclopedia of Education. Elsevier, pp. 72–83. https://doi.org/10.1016/B978-0-08-044894-7.01315-4

Marini, J.O., Botelho, N.F., 1986. A PROVÍNCIA DE GRANITOS ESTANÍFEROS DE GOIÁS. Rev. Bras. Geociências 16, 119–131.

Pimentel, M.M., Botelho, N.F., 2001. Sr and Nd isotopic characteristics of 1.77-1.58 Ga rift-related granites and volcanics of the Goiás tin province, Central Brazil. An. Acad. Bras. Cienc. 73, 262–276. https://doi.org/10.1590/s0001-37652001000200010

Pinto-Ward, 2018. Controls on the enrichment of the Serra Verde rare earth deposit, Brazil. Imperial College of London.

Polo, H.J. de O., Diener, F.S., 2017. Geologia e recursos minerais da Folha Mata Azul: SD.22-X-D-II. Projeto Nordeste de Goiás, escala 1:100.000. Estado do Goiás e Tocantins., Programa Geologia do Brasil, Levantamentos Geológicos Básicos. Serviço Geológico do Brasil - CPRM, Goiânia, GO.

Santana, I. V., Wall, F., Botelho, N.F., 2015. Occurrence and behavior of monazite-(Ce) and xenotime-(Y) in detrital and saprolitic environments related to the Serra Dourada granite, Goiás/Tocantins State, Brazil: Potential for REE deposits. J. Geochemical Explor. 155, 1–13. https://doi.org/10.1016/j.gexplo.2015.03.007

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R J. 8, 289–317. https://doi.org/10.32614/rj-2016-021

Wickham, H., 2014. Tidy Data. J. Stat. Softw. 59. https://doi.org/10.18637/jss.v059.i10

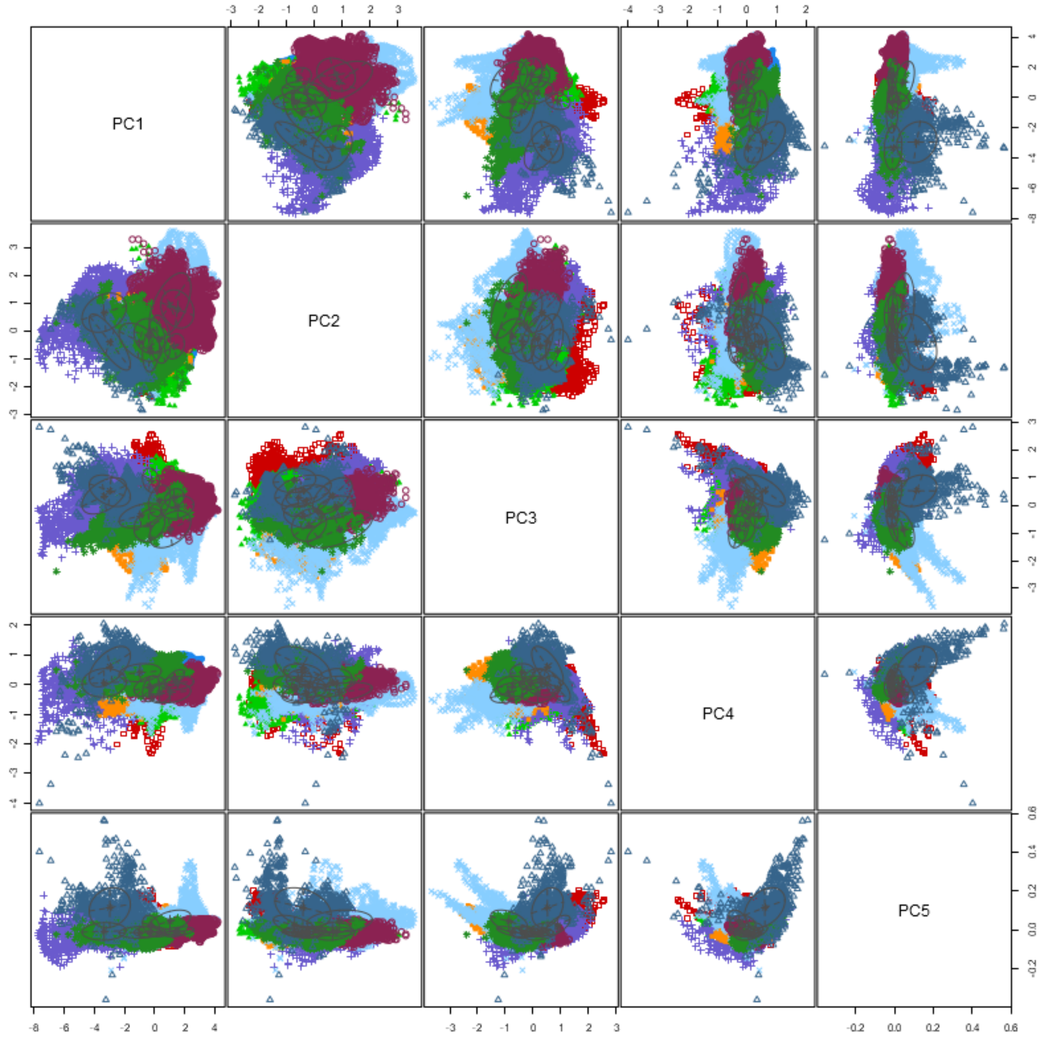

Figure 4: cluster segmentation and adjustment of the 9 VVV ellipsoids for all Principal Components combinations

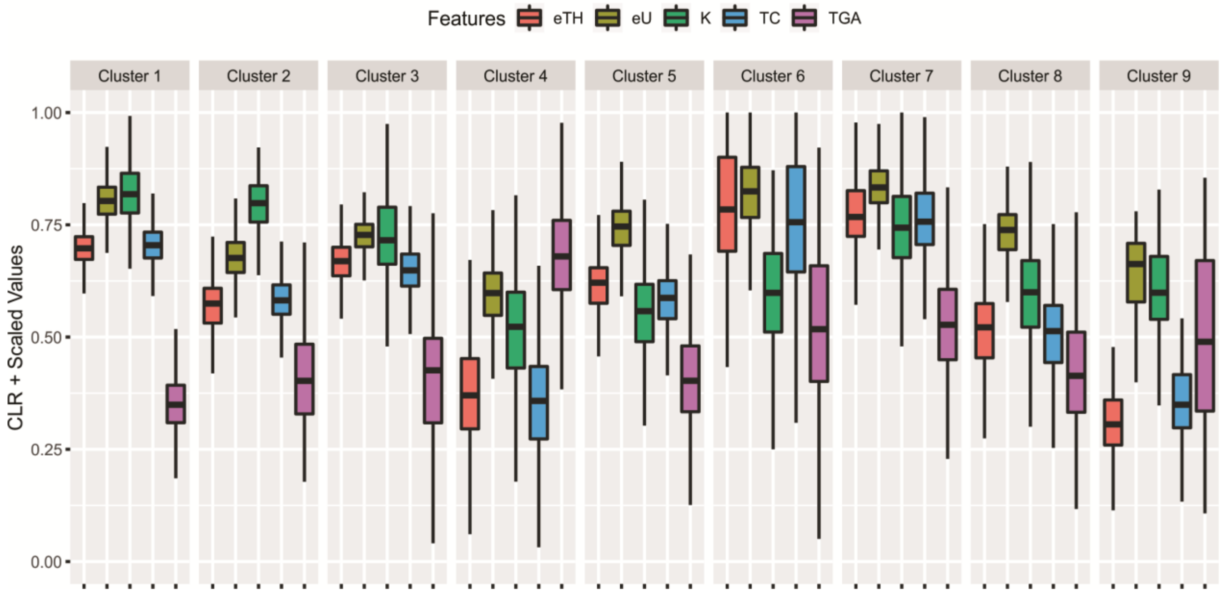

Figure 5: Boxplot of features classified according to the Cluster segmentation.

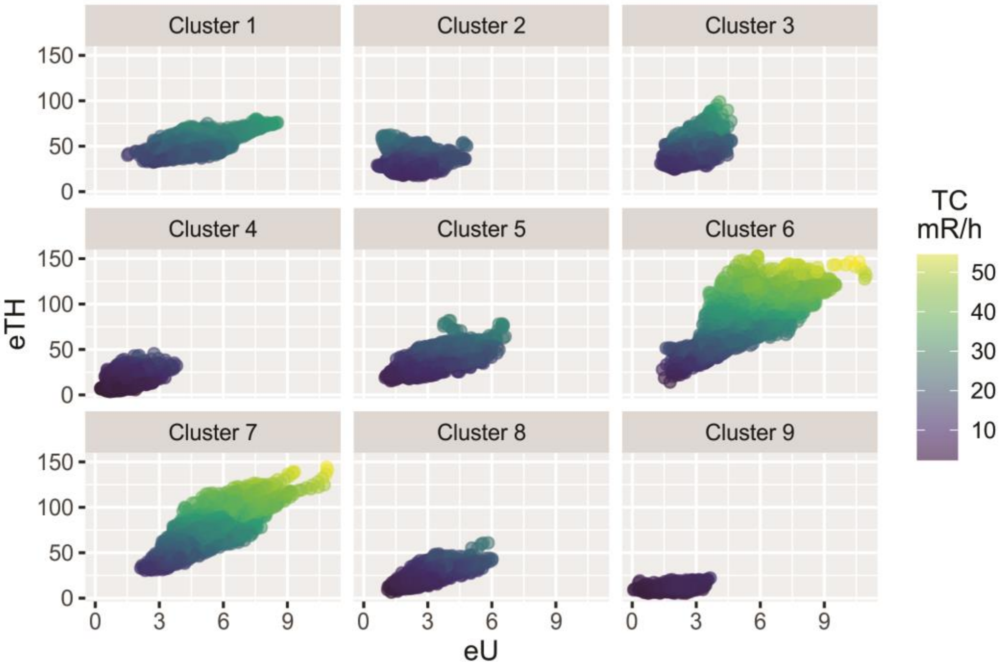

Figure 6: Biplot of eTh (ppm) vs eU (ppm) colored according to Total Count. The clusters 6 and 7 present the highest values of eTh and eU concentration, also showing more structural complexity.

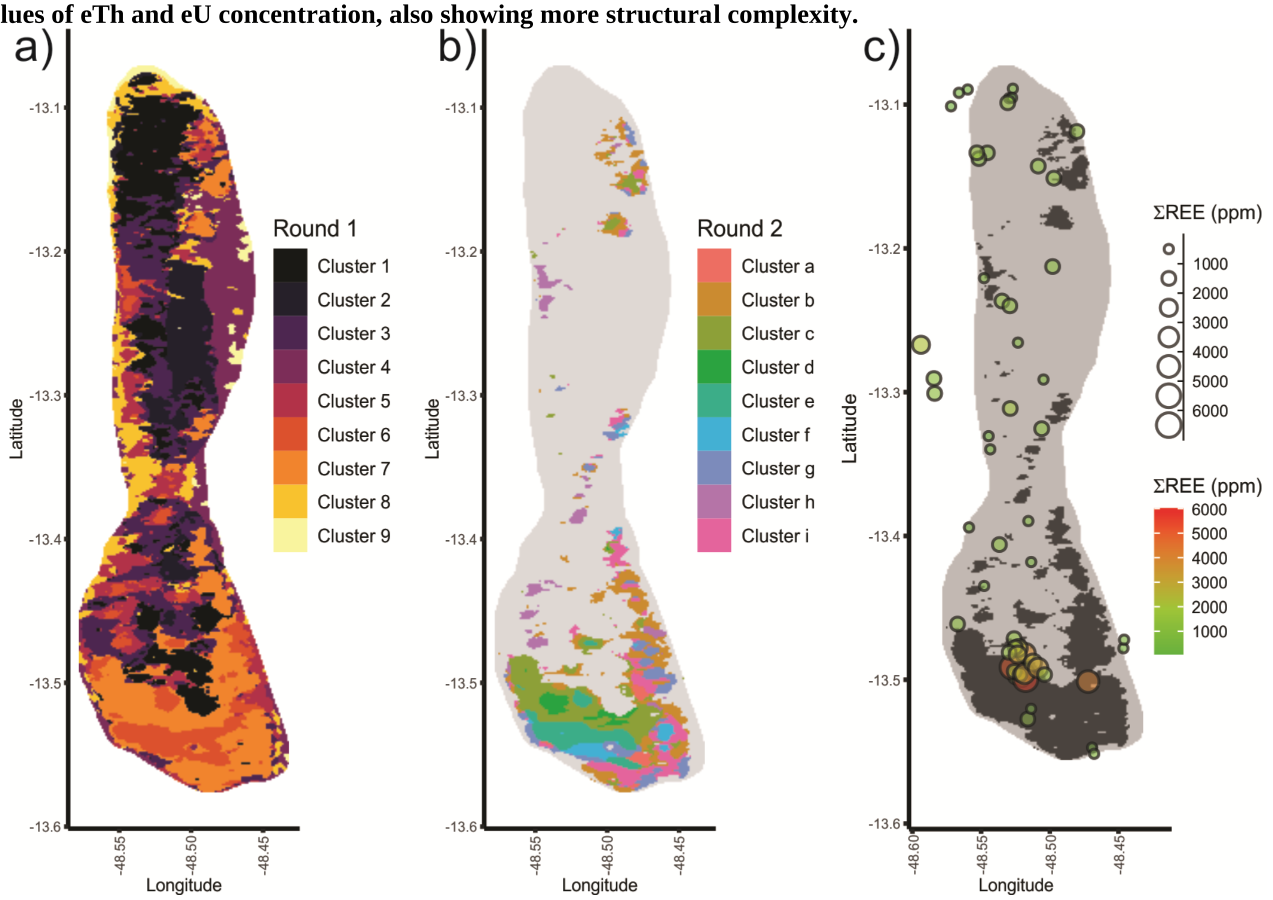

Figure 7: Unsupervised classification of SDG for a) Round 1, b) Round 2, and c) Lithogeochemistry data colored by the sum of all REE content (ΣREE, ppm), the highest concentrations are over the Clusters6 and 7 area (dark gray), despite there are relative REE-rich samples scattered all over the batholith.