Journal of the Geological Survey of Brazil

Unveiling geological complexity in the Serra Dourada Granite using self-organizing maps and hierarchical clustering: Insights for REE prospecting in the Goiás Tin Province, Brasília Belt, Central Brazil

Guilherme Ferreira da Silva[1,*], Marcos Vinícius Ferreira[1], Lucy Takehara Chemale[2], Igor Vasconcelos Santana[3], Nilson Francisquini Botelho[4]

[1]Geological Survey of Brazil - SGB, Setor Bancário Norte, Quadra 02, Bloco H, Ed. Central Brasília, Brasília - DF, Brazil, CEP: 70040-904.
[2]Geological Survey of Brazil - SGB, Rua Banco da Província, 105, Porto Alegre - RS, Brazil, CEP: 90840-030.
[3]Universidade Federal de Minas Gerais - UFMG, Av. Pres. Antônio Carlos, 6627, Belo Horizonte - MG, Brazil, CEP: 31270-901.
[4]Universidade de Brasília - UnB, Campus Universitário Darcy Ribeiro, Asa Norte, Brasília - DF, Brazil, CEP: 70910-900.

Abstract

This study explores the use of Self-organizing maps (SOM) combined with hierarchical clustering to provide insights into the geological differentiation and mineral prospecting in the Serra Dourada Granite (SDG), part of the Goiás Tin Province, northern Brasília Belt. After some issues on the geological cartography of the SDG based on traditional approaches, such as the interpretation of outcrops and the limited geochemistry data, often struggle to capture the complexity of high-dimensional geophysical datasets. To address this, we apply unsupervised machine learning techniques to segment airborne radiometric data, providing a more nuanced understanding of the SDG internal structure. Using airborne gamma-ray data, we employed SOM for dimensionality reduction and data segmentation, supported by hierarchical clustering. This methodology enabled us to identify distinct geological units with greater accuracy and resolution than traditional methods such as Principal Component Analysis (PCA). The SOM-based approach retained the data's original topology and revealed fine-scale patterns within the dataset, distinguishing between areas affected by magmatic processes and those influenced by post-magmatic hydrothermalism and supergene leaching. The results indicate that some clusters are mainly associated with magmatic differentiation, characterized by average concentrations of potassium (K), equivalent thorium (eTh), and equivalent uranium (eU) and others show evidence of secondary processes, including hydrothermal alteration and weathering. Notably, Cluster 4 is spatially linked to REE-enriched plateaus and the Serra Verde Mine, reinforcing its significance for mineral exploration. The SOM model proved more effective than PCA at capturing non-linear relationships within the data. While PCA provided insights into the primary variance, it did not fully account for the complex geological processes at play. In contrast, the SOM model segmented the data into clusters that reflected both broad radiometric trends and localized variations, particularly in areas influenced by hydrothermalism and supergene processes. Our findings underscore the value of machine learning techniques, particularly SOM, in geoscientific data analysis. This approach provides a robust framework for integrating multivariate radiometric data, offering valuable insights for geological mapping and mineral exploration, especially in regions with complex geological histories. The methodology presented here can be adapted to other geological settings, enhancing the accuracy of subsurface mapping and identifying areas of economic interest, such as Rare Earth Element (REE) and other critical mineral deposits.

Article Information

Publication type: Research papers
Received 24 September 2024
Accepted 27 March 2025
Online pub. 2 April 2025
Editor: Carlos G. Asato

Keywords:
Unsupervised Segmentation
Compositional Data Analysis
Clustering Algorithms
Machine Learning
Dimensionality Reduction

*Corresponding author
Guilherme Ferreira da Silva
E-mail address: guilherme.ferreira@sgb.gov.br

1 - Introduction

Geological mapping is a technique that involves extrapolating information from explicit geological evidence to infer the continuity, geometry, and shape of rocks and mineralized bodies in a two-dimensional representation (Brimhall et al. 2005). While the interpretation of outcrops has traditionally been used for geological mapping, remote sensing products such as digital elevation maps, satellite images, and regional airborne geophysics have become increasingly relied upon in recent decades due to their representation of physical responses across a systematic sampling. However, combining different layers of information from remote sensing products into a meaningful map can be time-consuming

ISSN: 2595-1939
https://doi.org/10.29396/jgsb.2025.v8.n1.5
Open access at https://jgsb.sgb.gov.br

and confusing, and some information may be unconsidered due to the lack of synthesis capacity for high-dimensional geoscientific data (Cracknell and Reading 2014; Kuhn et al. 2018; Lehmann et al. 2023; Nagar et al. 2024).

This research aims to propose a methodology for aiding in geological cartography through a quantitative assessment of available airborne radiometry. Specifically, the study focuses on the segmentation of the Serra Dourada Granite, the largest of the Goiás Tin Province (GTP) batholiths, northern Brasília Belt, using machine learning algorithms to integrate multivariate radiometry (K, eTh, eU, and other products derived from these). The batholith lacks detailed cartography, primarily due to the high similarity of rocks on the outcrop scale and the failure of efforts relying on systematic geochemistry to separate the granitic body.

Geological mapping is essential in understanding the geological history and mineral endowment of a region, but the process can be time-consuming and labor-intensive, particularly in regions with complex geology. Recent advances in machine learning algorithms have shown promise in aiding the geological mapping process by integrating various geophysical datasets or for targeting mineral exploration (Bergen et al. 2019; Costa et al. 2020; Silva et al. 2022a; Silva et al. 2022b; Lawley et al. 2023, 2024; Prado et al. 2020; Ng et al. 2023; Parsa et al. 2024; Rodriguez-Galiano et al. 2015). The paper explores the use of machine learning algorithms to integrate different layers of radiometric data to identify lithological boundaries and structural features in the case study of a granitic batholith in a region with complex geology.

The results demonstrate that machine learning algorithms can effectively integrate multivariate data to aid in the geological mapping process, significantly reducing the time and effort required while improving the accuracy and resolution of the resulting maps. This paper highlights the potential of using quantitative approaches in geoscience, including machine learning algorithms, for aiding the task of geological mapping and provide valuable insights for future studies aiming to integrate geophysical datasets for geological mapping purposes.

2 - Geological Settings

2.1 - Goiás Tin Province

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, northern Brasília Belt (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), the biggest batholith of the Goiás Tin Province, is a roughly 8-shaped massif with batholitic dimensions, with approximately 55 km length and 13 km width at its broader section. Along with other correlated oval-shaped massifs, namely, Serra do Encosto, Serra da Mesa and Serra Branca Granites, encompass a series of REE,Sn-mineralized granitoids that crop out in mid north Goiás State, Brazil. The massifs, especially the SDG, stand out in the region as high topographic terranes, contrasting with the surrounding geological units of the Serra da Mesa Group (not to be confused with the Serra da Mesa Granite). 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).

The granites of the Goiás Tin Province contrast sharply with the surrounding rocks in all evaluated radiometric maps, displaying K (%), eTh, and eU TC values that are tens to several hundred times higher than those of the adjacent lithologies (Carvalhêdo et al. 2025; Carvalhêdo et al. 2020). As a result, in ternary maps, these granites appear as predominantly white ellipsoids with no internal contrast.

This radiometric pattern can be attributed to the felsic and evolved nature of these granites, which are typically enriched in potassium-bearing minerals such as K-feldspar and biotite, as well as accessory phases that concentrate thorium and uranium, such as monazite, xenothime, and zircon. In contrast, the surrounding metasedimentary and metavolcanic rocks, which often consist of lower-t-high-grade metamorphic assemblages and mafic to intermediate compositions, tend to exhibit much lower K, eTh, and eU values. Depending on their specific mineralogical compositions, these units are expected to appear as darker or more muted tones in ternary radiometric maps. Additionally, within the Goiás Tin Province, pegmatites spatially associated with these granites may also show localized radiometric anomalies, particularly where they contain significant concentrations of radioactive accessory minerals. However, the qualitative use of radiometric data has yielded limited practical results in distinguishing the internal subdivisions of these granites. This limitation underscores the need for a new quantitative approach, which this study aims to introduce.

2.2 - Geochronology and Geological Evolution

The earliest geochronological studies carried out in the SDG dates back to the early 1980's, when Reis Neto (1980, 1983) and Macambira (1983) reported ages obtained by means of the Rb-Sr method. Since those pioneer studies, the authors seem to agree on the difficulties faced, due to the geological evolution of the granites in the realm of the Tocantins Province and Brasília Fold Belt, where complex deformational, orogenic and tectono-metamorphic events took place. They stress that such events have opened the original isotopic systems of rocks and/or minerals, promoting scattering of samples away from the best-fit isochrones.

Nevertheless, Reis Neto (1983) carried out analyses in the SDG – as well as in the Serra Branca and Serra da Mesa Granites – using the Rb-Sr method, which yielded ages of approximately 1479 Ma with initial $^{87}Sr/^{86}Sr$ = 0.790, for the SDG. In an attempt to produce more robust data, that author combined the measurements of both massifs in one single isochron diagram, having obtained 1465 Ma, with $^{87}Sr/^{86}Sr$ = 0.792. However, he emphasizes that those numbers are not totally reliable because tectono-thermal events and greisenization processes may have caused chemical disequilibrium and opening of the isotopic systems.

Still using the Rb-Sr method in four samples of the SDG, Macambira (1983) obtained ages ranging from 1870 – 1259 Ma, with initial $^{87}Sr/^{86}Sr$ = 0.710. Combining his own results with those from (Reis Neto 1980) (n=7), the numbers are 1653 ± 179 Ma, with $^{87}Sr/^{86}Sr$ = 0.700. When the two samples with

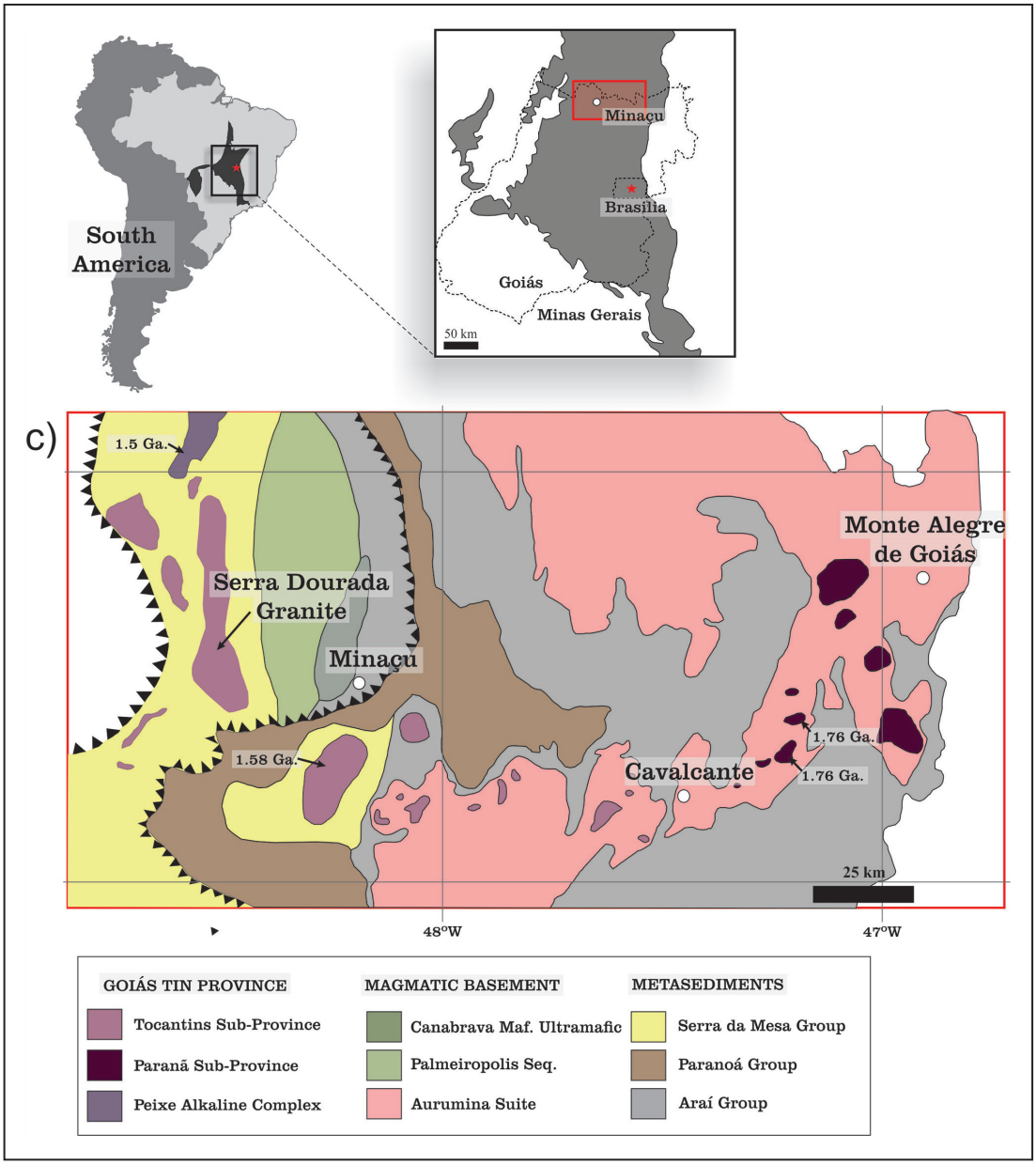

Figure 1: a) localization of the Tocantins Province in Central Brazil; b) Study area in the central portion of the Brasília Belt; c) Simplified Geological Map of the Goias Tin Province and the placement of the granites from the Tocantins and Paranã Sub-Provinces.

the highest deviation were disregarded, Macambira (1983) obtained ages of 1441 ± 105 Ma, with initial $^{87}Sr/^{86}Sr$ = 0.775, and when he considered only the two oldest samples, the values are 1885 Ma, with $^{87}Sr/^{86}Sr$ = 0.701. According to him, the latter age is the most reliable because it was yielded by a grayish granite sample, indicating that it did not go through late-magmatic processes that otherwise oxidize the Fe from feldspars, imprinting them a pinkish-red color.

The Rb-Sr dating technique requires the isotopic system to remain undisturbed, which is hardly fulfilled by granites such as the SDG and others related, given their complex evolution involving regionpal tectono-metamorphic events. This is partly overcome making use of other methods such as Pb-Pb and U-Pb, such as Reis Neto (1983), Pimentel et al. (1991) and Rossi et al. (1992) did. The former, combining data from Serra Branca and Serra da Mesa, reported Pb-Pb age of 1658 ± 44 Ma for those granites. Pimentel et al. (1991), studying the Sucuri and Soledade granites (Rio Paranã Subprovince) obtained U-Pb zircon ages of 1767 ± 10 Ma and 1769 ± 2 Ma, respectively. For the Serra da Mesa granite, that author obtained ages approximately 150 Ma younger, which gives about 1618 Ma, calculated by a simple subtraction. For the same massif (Serra da Mesa), Rossi et al. (1992) obtained Pb-Pb ages of 1580 ± 20 Ma. Those Pb-Pb and U-Pb ages are still considered today as the most reliable ones and, even not being carried out directly on the SDG, are considered to represent the age of the Rio Tocantins Subprovince as a whole.

In addition to the datings aforementioned, Reis Neto (1983) also reports K-Ar ages in biotites from the SDG, which yielded 535 Ma. This is similar to 571 ± 24 Ma obtained by Teixeira (2002) using the U-Pb method in monazite and to 530 Ma reported by Hasui and Almeida (1970), in biotite from the Serra da Mesa granite. Those Neoproterozoic ages are attributed to the Brasiliano Orogeny, the last major deformational event to act upon the granites.

2.3 - Tin and Rare Earth Mineralization

The REE and Tin mineralized deposits of these granites are associated with late-to-post-magmatic alteration greisen and pegmatite veins. The latter ones cut granites and the embedded granite metasediments (Marini and Botelho 1986). These greisen occurrences are of great importance in the metallogenesis of tin (Marini and Botelho 1986).

The Serra Dourada Granite has been studied as a potential region for REE deposits similar to those in Southeastern China (Pinto-Ward 2017; Santana et al. 2015; Santana and Botelho 2022). The unaltered granites of GTP show high content of ∑REE and REE pattern similar to alkaline granite, enriched more than 100 times when normalized to chondrite (Marini and Botelho 1986). The REE content increased during the hydrothermal process, such as albitization and biotitization, in the SDG controlled by primary and secondary REE minerals (Teixeira and Botelho 2006). This residual enrichment of REE in the altered rocks was attributed to the alkaline fluids during the hydrothermal process, which not mobilized the REE. In the SDG lateritic profile, the average REE enrichment is higher than the parent rock concentration, contributing tohe Serra Verde deposit (Pinto-Ward 2017). And other granite bodies of this province could have the similar potential of this deposit type (Costa Filho 2020; Vieira et al. 2019; Zapata and Botelho 2018).

The REE deposit in the SDG is mainly concentrated in the saprolite horizon and strongly enriched in the clay layer at the base of this horizon. The REE enrichment is 1,5 to 10 times compared to the parent bedrock (Pinto-Ward 2017). In addition to the SDG, other granite bodies of the GTP in the Rio Paranã subprovince were studied as potential REE deposits (Costa Filho 2020; Vieira et al. 2019; Zapata and Botelho 2018)

3 - Materials and Methods

3.1 - Data Source

For the total study area, shown in Figure 1, airborne radiometric data available by the Geological Survey of Brazil 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” (Alves et al. 2022; Silva and Alves 2021). The first acquired by the Geological Survey of Brazil and the others by the state government of Goiás. These airborne radiometric data have spacing between flight lines of 500m and nominal flight height of 100m. Data was interpoled to a cell size of 125m. Subsequently, the data was cropped for comprinsing only data corresponding to the Serra Dourada Granite, with a total of 29,675 samples with 6 variables.

3.2 - Data preparation and dimensionality reduction

In addition to traditional channels of Airborne Gamma-Ray data, i.e., Potassium (K), equivalent Thorium (eTh), equivalent Uranium (eU) and Total Count (TC), we calculated three more variables: Uranium Anomalous (Ud) and Potassium Anomalous (K), based on the methodology proposed by Saunders et al. (1993) and the K/eTh ratio.The feature data was prepared to optimize the processing steps according to some criteria. Radiometric data was first corrected to remove any negative value founded in the conversion of measured signal to equivalent concentration. A constant was added to this step to level correct negative values without severely impacting the distribution parameters. No influence of vegetation was observed in the dataset (Silva and Graça 2018) taking as case study an area in the center of the Rondônia State, Amazonia, northern Brazil, where wooded and deforested areas are frequently juxtaposed. The control of the wooded areas is made using Landsat satellite images, by the calculation of the Normalized Difference Vegetation Index (NDVI, as the batholith remained entirely covered by native vegetation at the time of the airborne survey.

As data have different units and scale values, it is necessary to normalize the distributions in order to keep the same values of ranges, putting the variances in the same order of magnitude. It is specially needed for applying such techniques as Principal Component Analysis (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 concentrations of chemical elements (Potassium, Thorium, and Uranium), these features can be considered compositional (Aitchison 1982, 2008), as they are ideally positive and interfere on each other, as the ideal concentration of any chemical composes is equal to 100%. For this specific work, we used a Centered Log-Ratio (CLR) transformation (Figure 2b).

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 algebraical principles (e.g., cross product and inner product of matrices) that creates mechanisms to rotate 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 dimensionality reduction (Grunsky and Arne 2020), as the first components explain more of data variance, the last ones is many times negligible. Despite generating 6 components at the end (Figure 2c), for this work, only the first four components components corresponds to 99.6% of the data variance and were considered to this work.

3.3 - Data segmentation

In this work, we ally Self Organing Maps and Hierarchical Clustering for data segmentation, i.e. the agroup similar data together and discriminate between different feature signatures based on dissimilarity.

Self-organizing maps (SOM) are a type of unsupervised neural network algorithm that can be used for data visualization and clustering. According to Kohonen (1998), SOM use a competitive learning process to identify patterns in the input data and organize them onto a low-dimensional map. The algorithm works by creating a set of nodes (or units), each of which represents a different location on the map, and then iteratively through different epochs adjusting the weights associated with these nodes to minimize the difference between the input data and the node weights (Kohonen 1998). As a result, similar input vectors are mapped

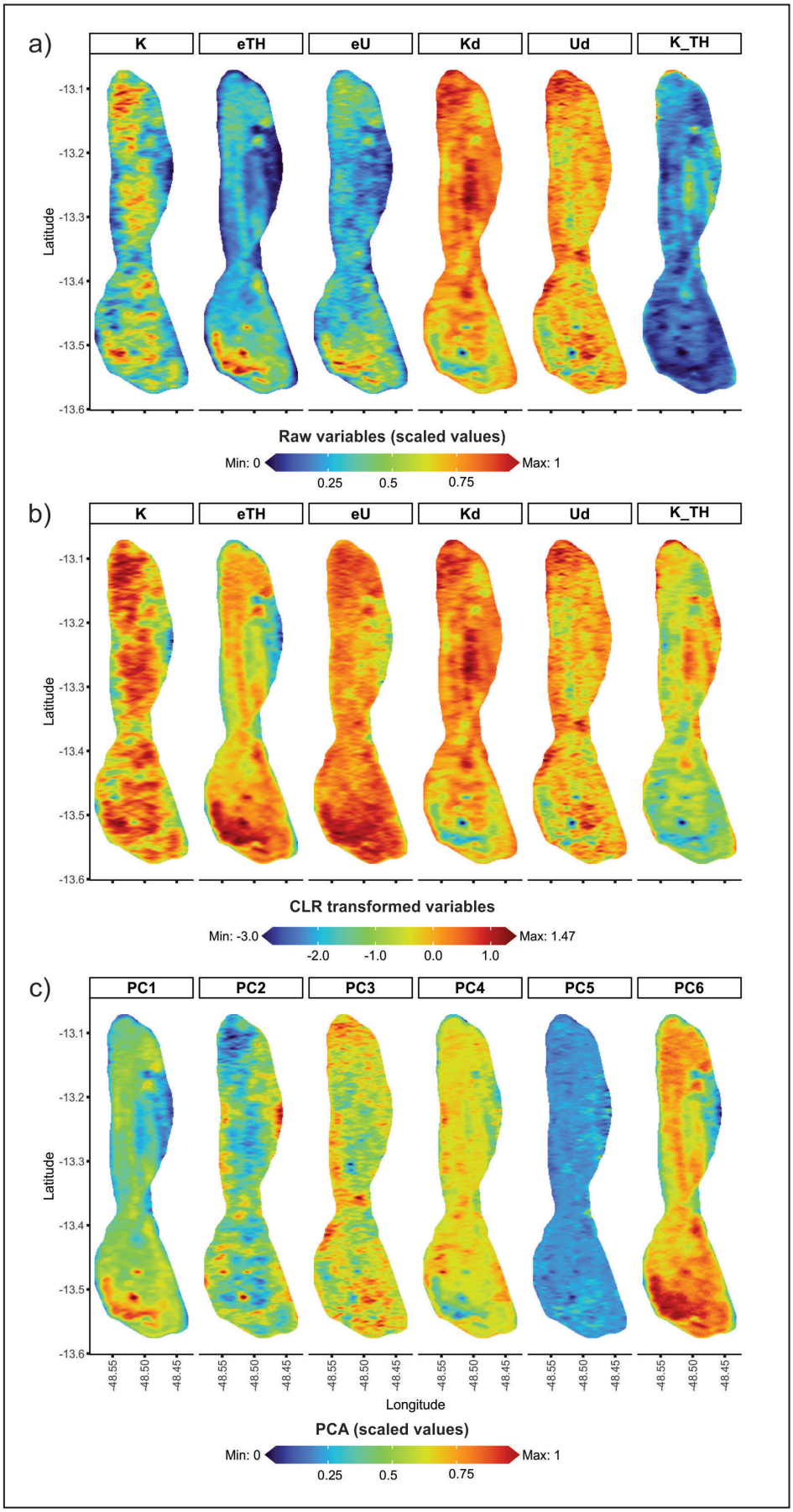

Figure 2: Representation of Input Features in Maps. a) raw data rescaled to a 0-1 distribution; b) Input features after a CLR transformation; c) Maps of PC1 to PC6 rescaled to represent internal variation.

to nearby nodes on the map, while dissimilar vectors are mapped to more distant nodes (Vesanto and Alhoniemi 2000). This organization of the input data onto the map provides a visualization of the patterns in the data that can be useful for understanding complex relationships between variables (Kaski et al. 1998). Additionally, clustering of the input data can be achieved by grouping together nodes on the map that have similar weights, thus identifying groups of similar data points (Vesanto and Alhoniemi 2000). SOM have been applied in various fields, including image analysis, natural language processing (Kohonen 1998) and data mining (Vesanto and Alhoniemi 2000), and geoscience, due to their ability to reveal complex multivariate patterns within data (Bação et al. 2005; Carneiro et al. 2012; Chudasama et al. 2022; Kebonye et al. 2021; Torppa et al. 2019).

As pointed out by Vesanto and Alhoniemi (2000), SOM has been shown to be effective in visualizing complex relationships between geological and geophysical data in a more intuitive and interpretable way than traditional statistical methods. Similarly, Filippi et al. (2010) employed SOM in the classification of remote sensing data for lithological mapping, demonstrating its effectiveness in identifying lithological units and mapping geological structures. Additionally, SOM has been applied in mineral exploration, such as in the work of Chudasama et al. (2022), where it was used to identify mineralization zones based on geophysical data and geology interpreted layers. These studies demonstrate the potential of SOM in aiding geoscientific data analysis and interpretation.

Hierarchical clustering is especially useful for geoscience data segmentation, where the goal is frequently to identify meaningful groups within complex, multivariate datasets. Unlike methods that require a predetermined number of clusters, Hierarchical Clustering offers a more adaptable approach by allowing for the exploration of clustering structure at various levels of granularity. This feature is especially useful in geological mapping and mineral prospecting, where natural divisions in the data may not be immediately apparent. By combining Hierarchical Clustering with techniques like Self-organizing maps (SOM), we can improve the identification of coherent geological units while revealing subtle patterns that other clustering methods may miss.

For this work, we implemented SOM segmentation with different arrays of nodes and epochs in order to identify the best fit. After selecting the best performing model, we clustered the nodes based on their inputs for each variable through a visual estimation of the dissimilarity matrix and then transferred this assignment to the original data. In all SOM models, we used hexagonal toroidal topology with Euclidean distance and a Gaussian Neighborhood Function (please refer to Data Availability session to access the full specifications).

Table 1: Statistical summary of airborne gamma ray data for selected features. Abbreviations: K (Potassium), eTh (equivalent Thorium), eU (equivalent Uranium), TC (Total Count), Kd (Anomalous Potassium), and Ud (Anomalous Uranium)

Statistics	K (%)	eTh (ppm)	eU (ppm)	TC	Kd	Ud
Min	0.3274	3.777	0.07418	2.424	-6.5641	-6.08414
1st Quartile	1.649	30.006	2.3474	13.529	-0.5413	-0.49619
Median	2.2333	42.869	3.17032	17.76	0.1653	0.05227
Mean	2.2837	44.995	3.39466	18.569	0	0
3rd Quartile	2.892	54.453	4.23792	22.435	0.6867	0.56553
Max.	5.0667	152.62	10.95191	54.492	2.7726	3.06004

4 - Results and Discussions

4.1 - Dimensionality reduction

In this study, Principal Component Analysis (PCA) was employed as a dimensionality reduction technique. The first four principal components (PC1 to PC4) account for approximately 99% of the total data variance (Figure 3), indicating that the contributions of PC5 and PC6 are nearly negligible. Therefore, these higher-order components were not considered in subsequent analyses.

The solely PCA is valuable to provide insights to the interpreter. The variables eU and eTh are highly correlated and their representing vectors in Figure 3 usually points to same direction, implying that also in PCA space this correlation is true, as expected for igneous rocks. However, some features are better explained in determined PC, as example, eU and Uk are better represented in the positive values of PC3, while K, K/eTh and Kd are represented in the negative values of the same component.

PCA is inherently a linear technique and may not fully capture complex, non-linear relationships present in the data. In our analysis, we observed that the fourth principal component (PC4) exhibits non-linear relationships compared to the other three principal components (PC1 to PC3). This observation suggests that traditional linear methods like PCA may not entirely capture the underlying data behavior. Despite its limitations, PCA still serves as a practical preliminary step for data exploration and visualization, reducing dimensionality and highlighting key features before applying more advanced non-linear techniques.

4.2 - SOM model

The SOM model iterations on the training length (rlen) and the number of units across different grid arrays allowed the scanning of the compositional variation on the SDG using different parameters (Figure 4). This iteration also provides valuables insights about how SOM works on retaining the data original topology and how an excess of nodes can leeds to overfit in noisy data.

Models with lower length of training (rlen = 10 epochs) are easier to compute, but often are more simplistic than the others, and thus, they tend to fail in capturate all the inner variations of the data structure. On the other hand a longer length of training (rlen = 1000 epochs) are computational onerous, and not necessarily it comes with improvements on segmentation. In the same way, the Figure 4 shows that increment of the number of units can also leads to unsatisfying clustering, with the aspects of output models for this arrange being more noisy than the others with a looser array.

Therefore, we considered for this work a intermediate model of a grid array of 24 x 32 units treined for 300 epochs the best match for the purpose of this work, in the way that this is not an overtrained model and the used array of units is able for capturing the data internal structuring without apparent noise effect.

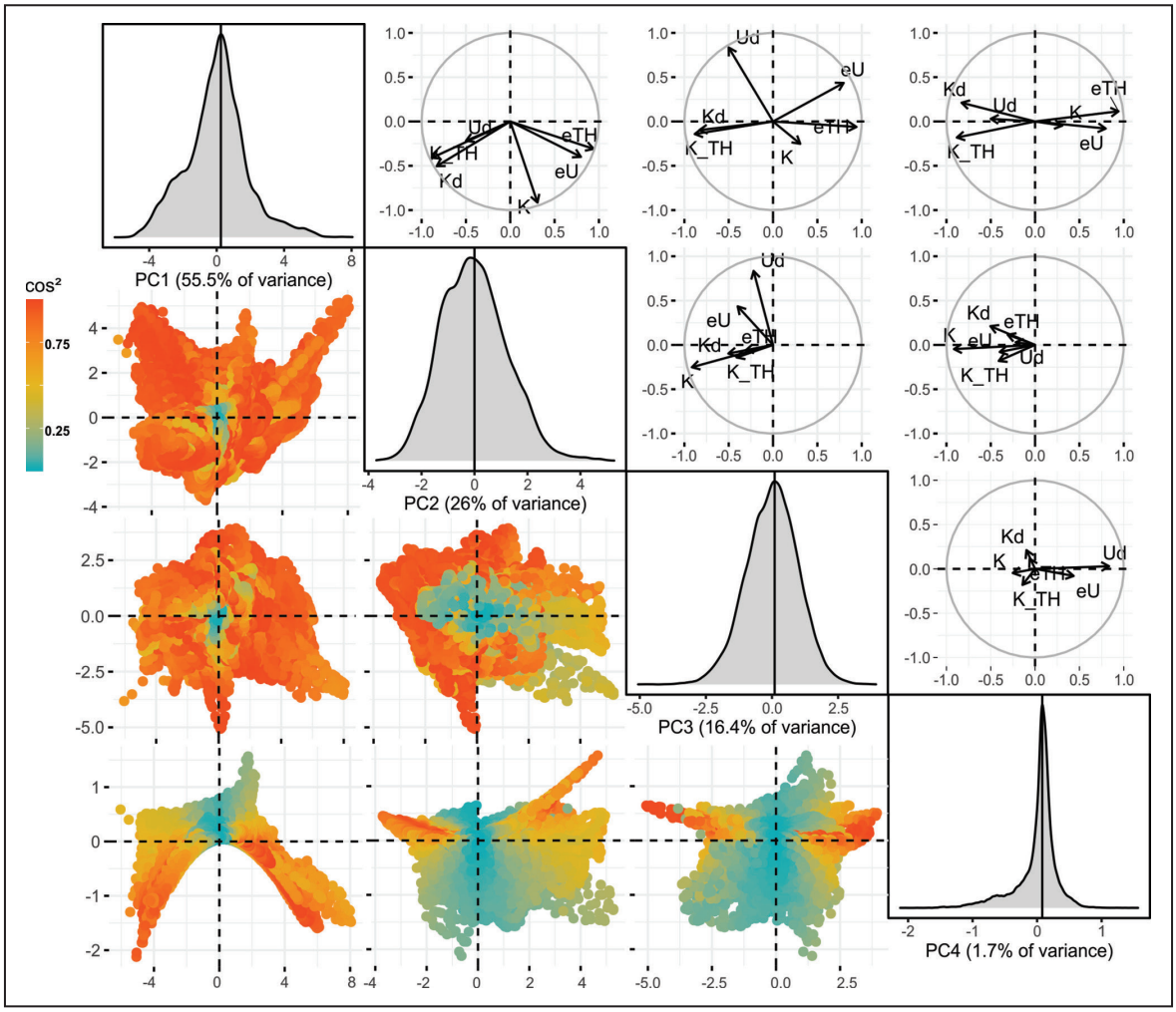

Figure 3: The matrix illustrates three different ways of analyzing the PCA output. The main diagonal displays the distribution of values for each principal component across the entire dataset. The upper diagonal matrix shows bivariate plots for all possible combinations of principal components, with vectors indicating the correlation of the input features with each axis. The lower diagonal matrix represents the data distribution for each pairwise combination of principal components, with each point color-coded to indicate the goodness of fit to the represented plane, as measured by the square cosine method.

4.3 - Data segmentation

After training the SOM model, we analyzed the SOM units using a dissimilarity matrix to visually estimate the optimal number of clusters (Figure 5a). The dissimilarity matrix, created by ranking and sorting the Euclidean distances between paired units, allowed us to observe regions of similarity and dissimilarity across the map. Clusters were identified as groups of units with relatively small distances, as indicated by the blue rectangles in Figure 5a. These regions of lower distances near the diagonal of the matrix suggest coherent groupings within the dataset, providing insight into the internal structure of the SOM.

Once the ideal number of clusters was established, we applied a hierarchical clustering approach to further refine the segmentation. The resulting dendrogram (Figure 5b) clearly shows a distinction between dominant clusters, such as Clusters 1 to 3, 5, and 7, which represent the majority of the dataset, and less frequent clusters, like Clusters 4, 6, and 8. The hierarchical clustering process enabled us to visualize how data points are grouped at different levels of similarity, with clusters merging at higher levels of the dendrogram, reflecting increasing dissimilarity.

Visualizing the segmented SOM units through Kohonen Maps (Figure 5c to j) provided further insights into the spatial distribution and internal structure of the clusters. Each map reflects how the different clusters are represented across the SOM grid, with clear differentiation between various zones. This internal structure mirrors the original data architecture, allowing us to interpret these clusters as reflecting distinct geological processes affecting the Serra Dourada Granite (SDG).

The Kohonen Maps also revealed subtle patterns in the distribution of radiometric properties, suggesting that the internal structure of the SDG is more complex than previously thought. The identified clusters highlight areas of potential mineralogical interest, such as zones with higher thorium or potassium content, which could be associated with REE (Rare Earth Element) mineralization or other economic mineral deposits. This detailed segmentation not only enhances our understanding of the compositional variation within the SDG but also provides a framework for more targeted geological mapping and mineral exploration., representing variations on the granite surface. These variations may be attributed to magmatic, hydrothermal, or supergenic processes that have affected the granitic rock.

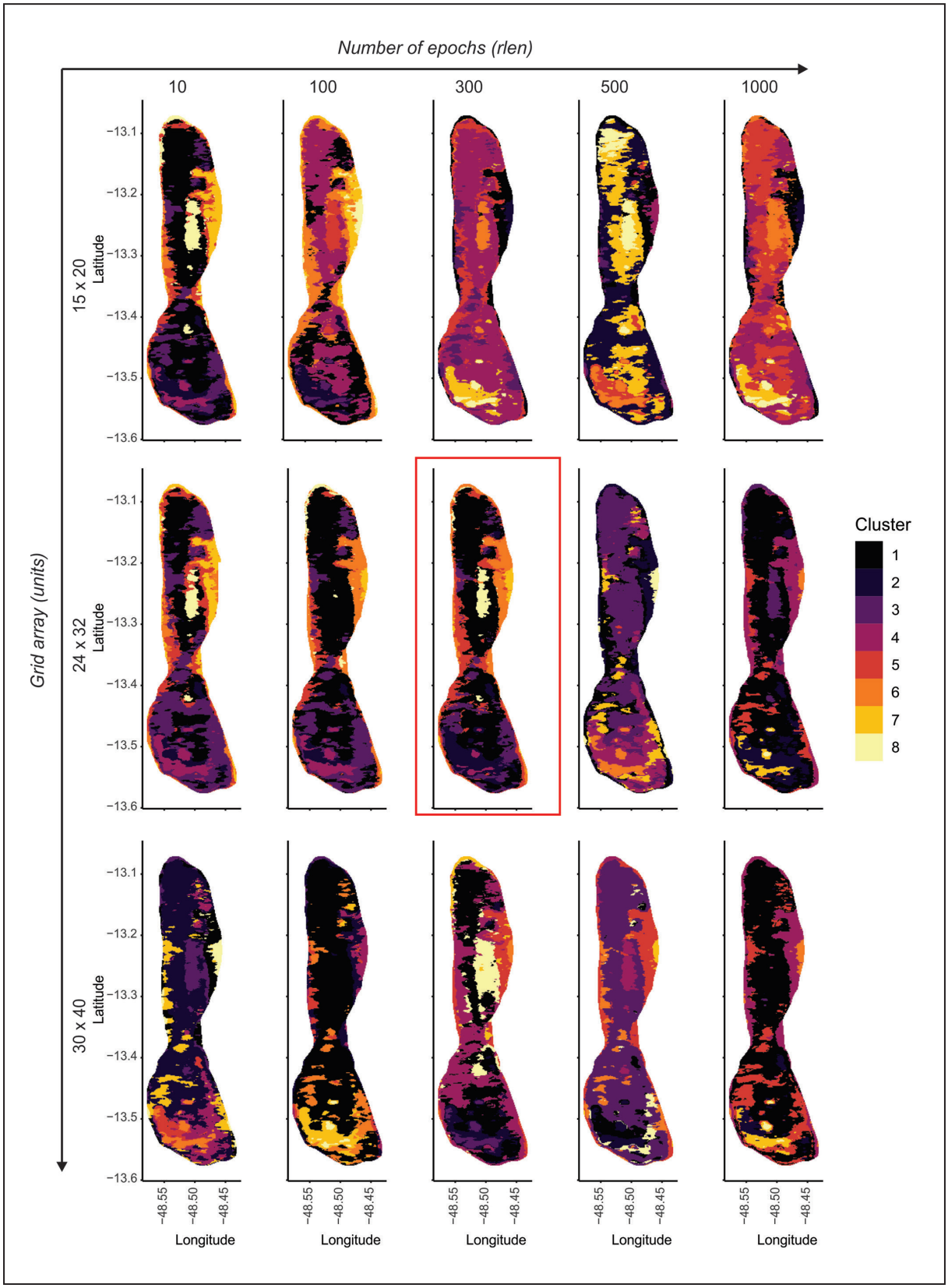

Figure 4: Different iterations on SOM varying the grid array and the length of training with number of clusters fixed (8 groups). The optimum model regarding the balance of noise and rlen is outlined in red.

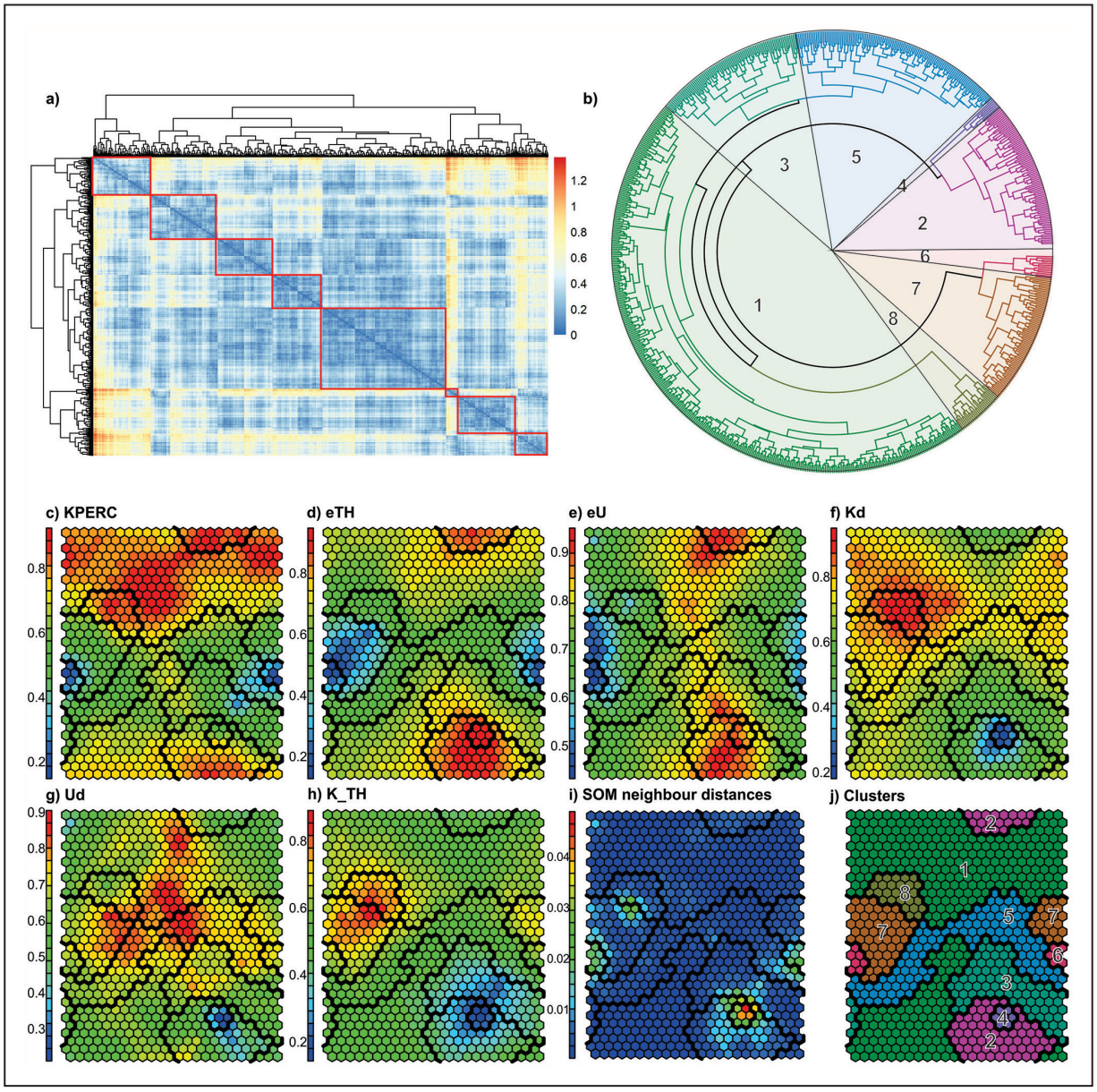

Figure 5: Clustering representation across differente methods: a) dissimilarity matrix for each SOM node calculated by Euclidean Distance; b) Hierarchical dendogram segmented according to the dissimilarity for the eight selected clusters; c to h) Kohonen Maps with 24 x 32 array showing the mean value for each feature; i) Kohonen Map represented by the mean euclidean distance from neighbours; j) final cluster assignment on Kohonen maps for the designed model.

4.4 - Cluster Interpretation

Based on the median values of each cluster for the analyzed variables (Figure 6), Clusters 1, 3, and 5 exhibit average concentrations of K, eTh, and eU, with medium to low levels of Ud, Kd, and the K/eTh ratio. In contrast, Clusters 2, 4, and 6 to 8 display a more irregular pattern, with relative enrichment in certain variables. Specifically, Clusters 2 and 4 are significantly enriched in eTh and eU but show lower concentrations of K, Kd, Ud, and the K/eTh ratio. This signature suggests secondary processes, such as post-magmatic hydrothermal activity (e.g., albitization) and/or leaching of mobile elements due to lateritic weathering, which could explain the loss of potassium content. A similar interpretation applies to Cluster 6, which shows an intermediate composition for all radioelements.

In fact, Clusters 2 and 4 are spatially correlated with areas identified as hydrothermally alterad by Pinto-Ward (2017), specially with Cluster 4 corresponding to plateaus related to leached and laterized domains associated with the Serra da Verde Mine and other REE Ionic Clay occurences (Figure 7).

Therefore, Clusters 1, 3, and 5 may represent areas where magmatic processes have played a significant role, given their prevalence and distinct multivariate signatures. These clusters suggest magmatic differentiation within the granite body, with variations in mineral composition that could justify differences in radioelements. In contrast, Clusters 4, 6, and 8, being less frequent and spatially confined, are likely associated with hydrothermal or supergene processes. These processes may have altered the original granitic composition through fluid circulation or weathering, resulting in localized enrichment or depletion of specific elements.

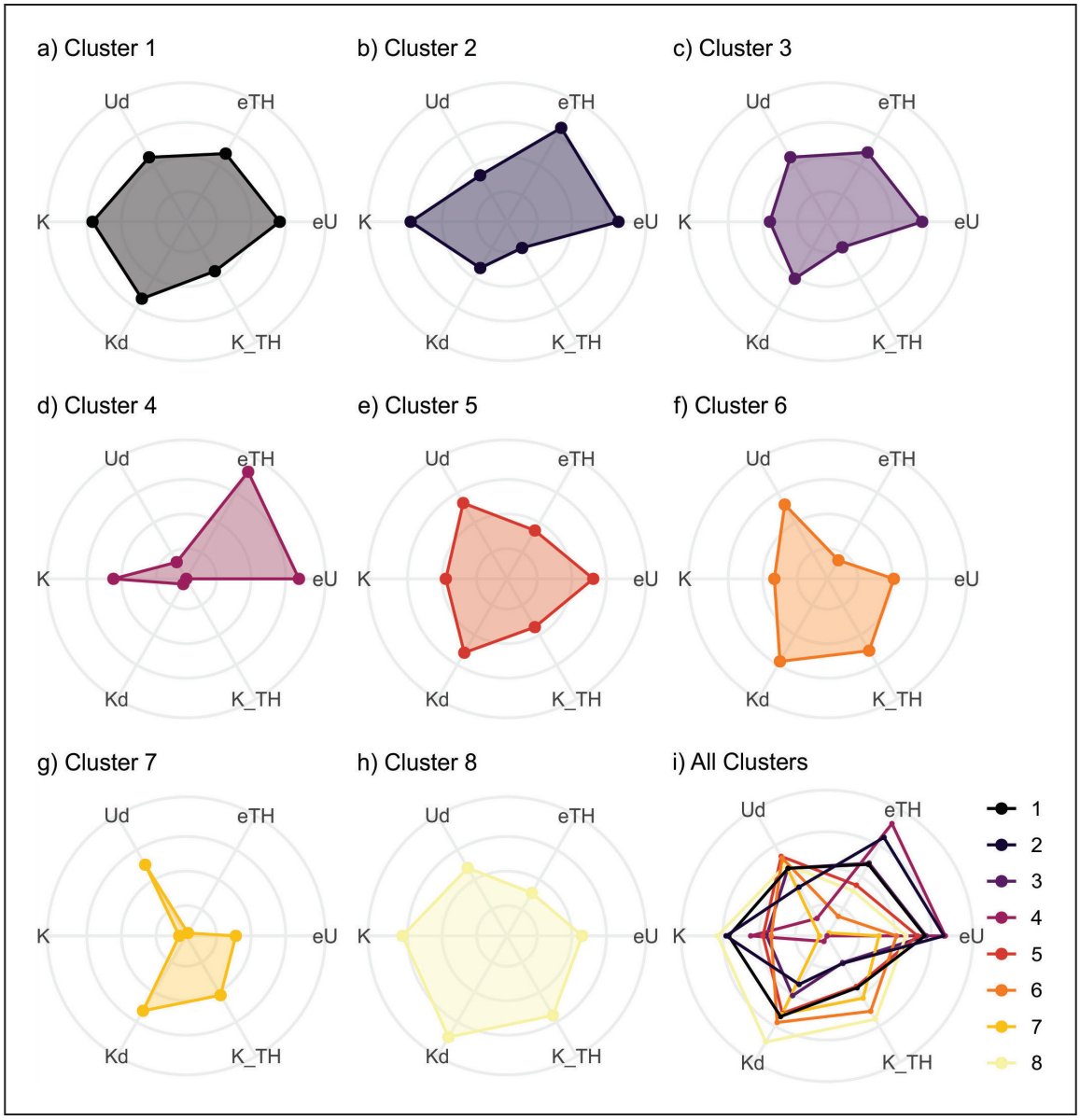

Figure 6: a to h)radar plot for cluster signature, showing the median value of each variable for all eight selected clusters/ i) All clusters grouped in the same plot.

Cluster 7, characterized by the lowest concentrations of eTh, K, and eU, likely represents alluvial and/or elluvial sedimentation or even schists and quartzites of the Serra da Mesa Group surrounding the SDG intrusion (Figure 7). This interpretation is supported by the spatial distribution of Cluster 7, which is predominantly located along the border of the SDG body.

4.5 - Comparison with PCA

The SOM combined with hierarchical clustering proved more efficient at differentiating the internal structure of the SDG compared to the first four principal components (PCs) of PCA. While the first four PCs capture the major variances in the dataset, their linear nature limits the ability to fully resolve the complex, non-linear relationships within the data. The clusters derived from the SOM, on the other hand, were able to segment the dataset more effectively, capturing both large-scale patterns and subtle, localized variations.

For instance, while PC1 highlights broad radiometric trends, the SOM-based clusters, particularly Clusters 1, 2, and 5, align more closely with distinct compositional zones, offering better spatial resolution. Additionally, PCs like PC3 and PC4 reveal only partial insights into potassium and thorium variations, whereas the SOM clusters capture finer details, such as areas impacted by hydrothermal or supergenic processes (e.g., Clusters 4 and 8), that PCA alone cannot fully distinguish. This demonstrates the enhanced capability of the SOM + hierarchical clustering approach in revealing the complex geological signature of the SDG.

5 - Conclusions

This study presents a novel approach to geological mapping and mineral prospecting of the Serra Dourada Granite (SDG) using Self-organizing maps (SOM) combined with hierarchical clustering. The proposed methodology successfully integrates airborne radiometric data and provides an enhanced framework

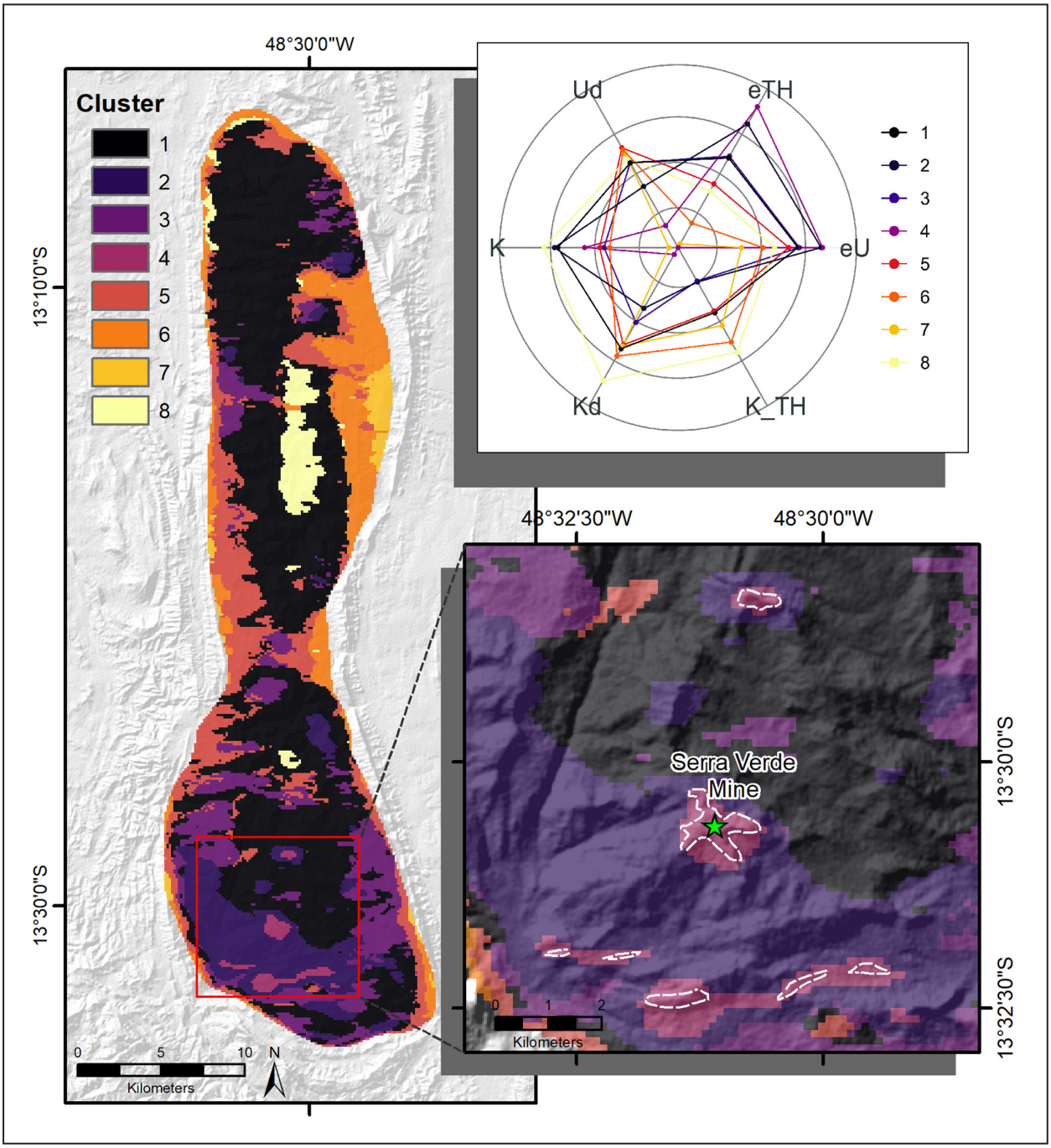

Figure 7: Detailed map of the Serra Dourada Granite (SDG) with the proposed subdivision based on the SOM model shows the distribution and signature of each cluster. The detailed map highlights the predominance of Clusters 1 (black), 2 (dark purple), and 4 (light purple) in the southern portion of the SDG batholith. Notably, Cluster 4 is associated with the plateaus linked to Rare Earth Element (REE) mineralization, which aligns with the location of the Serra Verde Mine.

for data segmentation compared to traditional linear methods, such as Principal Component Analysis (PCA). By leveraging unsupervised machine learning techniques, we were able to overcome the limitations posed by the complex geology of the SDG, allowing us to identify meaningful geological units with greater accuracy and resolution.

Our findings indicate that the SOM-based clustering approach is particularly effective in capturing both large-scale and localized variations within the SDG. Clusters 1, 3, and 5 are probably associated with mainly magmatic processes, showing average concentrations of potassium (K), equivalent thorium (eTh), and equivalent uranium (eU), whereas Clusters 2, 4, and 6 to 8 exhibit signatures of secondary processes such as post-magmatic hydrothermalism and weathering, contributing to the enrichment or depletion of certain radioelements. Notably, Cluster 4 correlates with REE-enriched plateaus and is spatially associated with the Serra Verde Mine, highlighting the cluster segmentation's relevance for mineral exploration. The location and correlated signatures of Cluster 4 can also serve as an exploration guide for REE deposits in the SDG and surrounding bodies within the Goiás Tin Province, as it likely represents the superposition of post-magmatic hydrothermal processes and supergene leaching that resulted in the "Serra Verde type" ionic clay mineralizations.

Furthermore, the SOM method demonstrated superior performance compared to PCA in differentiating non-linear relationships within the dataset. While PCA provided valuable insights into the primary variance in the data, it was unable to fully capture the complexity of the geological processes reflected in the multivariate dataset. The SOM model, on the other hand, effectively retained the data's original topology and revealed fine-scale compositional patterns, including regions influenced by hydrothermal and supergene processes.

The results of this study underscore the potential of machine learning techniques, particularly SOM, in geoscientific data analysis. By offering a more robust and nuanced understanding of the SDG's internal structure, this approach can serve as a valuable tool for future geological mapping efforts and mineral exploration initiatives, particularly in regions with similar complex geology. Additionally, the methodology can be adapted and applied to other geological settings to enhance the accuracy of subsurface mapping and identify areas of economic interest, such as zones of REE and other critical mineral deposits.

Acknowledgments

We acknowledge the Geological Survey of Brazil for supplying the data. Some of the authors are employees of the Geological Survey of Brazil and this work was conducted as part of their duties. No funding was received from external sources outside the Brazilian Government through the Geological Survey of Brazil activities. We thank the generosity of Colin Farquharson during reviewing this manuscript, as well as the handling editors Carlos Gabriel Asato and Evandro Luiz Klein. Additionally, the authors acknowledge the use of Large Language Models for grammar checking and enhancing the writing quality.

Data availability

Data and code used for this work can be found at the following GitHub repository <https://github.com/gferrsilva/SOMSerraDourada>. The script was written in R language, version 4.3.1. All the dependancies are described on the repository.

Authorship credits

Author	A	B	C	D	E	F
GFS	■	■	■	■	■	
MVF	■	■	■	■	■	
LTC	■		■	■	■	■
IVS	■		■	■	■	
NFB	■		■		■	■

A - Study design/ Conceptualization **B** - Investigation/ Data acquisition
C - Data Interpretation/ Validation **D** - Writing
E - Review/Editing **F** - Supervision/Project administration

References

Aitchison, J. 1982. The statistical analysis of compositional data. Journal of the Royal Statistical Society, 44(2), 139-160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Aitchison, J. 2008. The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. In: Compositional Data Analysis Workshop, 3, 1–28.

Alves F.M., Silva E.R., Silva A.B. 2022. Atlas aerogeofísico do estado de Goiás. Goiânia, SGB-CPRM. Available online at: https://rigeo.sgb.gov.br/handle/doc/23325 / (accessed on 31 March 2025).

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., Silva V.S., Borges W., Carmelo A.C., Almeida T. 2013. Geologia e características estruturais do Projeto Mata Azul (GO), Faixa Brasília Setentrional. In: Simpósio de Geologia do Centro Oeste, 13, 1–5.

Bação F., Lobo V., Painho M. 2005. The self-organizing map, the Geo-SOM, and relevant variants for geosciences. Computers & Geosciences, 31(2), 155–163. https://doi.org/10.1016/j.cageo.2004.06.013

Bergen K.J., Johnson P.A., Hoop M.V., Beroza G.C. 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science, 363(6433). https://doi.org/10.1126/science.aau0323

Brimhall G.H., Dilles J.H., Proffett J.M. 2005. The role of geologic mapping in mineral exploration. In: Doggett M.D, Parry J.R. Wealth creation in the minerals industry. Society of Economic Geologists, p. 221–241. https://doi.org/10.5382/SP.12.11

Carneiro C.C., Fraser S.J., Crósta A.P., Silva A.M., Barros C.E.M. 2012. Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon. Geophysics, 77(4). https://doi.org/10.1190/geo2011-0302.1

Carvalhêdo A.L., Carmelo A.C., Botelho N.F. 2020. Geophysical-geological model of the Pedra Branca massif in the Goiás Tin Province, Brazil. Journal of South American Earth Sciences, 101, 102593. https://doi.org/10.1016/j.jsames.2020.102593

Carvalhêdo A.L., Carmelo A.C., Lima J.P.D., Botelho N.F., Chornobay A. 2025. Investigation of radiogenic heat production in granites of the Goiás Tin Province, Central Brazil. Geothermics, 125, 103183. https://doi.org/10.1016/j.geothermics.2024.103183

Chudasama B., Torppa J., Nykänen V., Kinnunen J. 2022. Target-scale prospectivity modeling for gold mineralization within the Rajapalot Au-Co project area in northern Fennoscandian Shield, Finland. Part 2: Application of self-organizing maps and artificial neural networks for exploration targeting. Ore Geology Reviews, 147, 104936. https://doi.org/10.1016/j.oregeorev.2022.104936

Costa Filho D.S. 2020. Caracterização mineralógica e proveniência de Monazita-(Ce), Xenotima-(Y) e Zircão de Placer na Província Estanífera de Goiás: estão estes minerais relacionados com o granito tipo-A Serra Dourada? MSc Dissertation, Universidade de Brasília, Brasília, 49 p. Available online at: http://repositorio2.unb.br/handle/10482/39482 / (accessed on 31 March 2025).

Costa I.S.L., Serafim I.C.C.O., Tavares F.M., Polo H.J.O. 2020. Uranium anomalies detection through Random Forest regression. Exploration Geophysics, 51, 555–569. https://doi.org/10.1080/08123985.2020.1725387

Cracknell M.J., Reading A.M. 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Computers & Geosciences, 63, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008

Filippi A., Dobreva I., Klein A.G., Jensen J.R. 2010. Self-Organizing Map-based Applications in Remote Sensing. In: Matsopoulos G.K. (ed.) Self-Organizing Maps. InTech. https://doi.org/10.5772/9163

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). Geoscience BC Report 2020-06. Available online at: https://cdn.geosciencebc.com/project_data/GBCReport2020-06/GBCR%202020-06%20Mineral-Resource%20Prediction%20Using%20Advanced%20Data%20Analysis%20and%20Machine%20Learning%20revised%20November%2013%202020.pdf / (accessed on 31 March 2025).

Hasui Y., Almeida F.F.M. 1970. Geocronologia do centro-oeste brasileiro. Boletim da Sociedade Brasileira de Geologia, 19(1). Available online at: http://boletim.siteoficial.ws/pdf/1970/19_1-5-25.pdf / (accessed on 31 March 2025).

Kaski S., Honkela T., Lagus K., Kohonen T. 1998. WEBSOM - Self-organizing maps of document collections. Neurocomputing, 21(1-3), 101–117. https://doi.org/10.1016/S0925-2312(98)00039-3

Kebonye N.M., Eze P.N., John K., Gholizadeh A., Dajčl J., Drábek O., Němeček K., Borůvka L. 2021. Self-organizing map artificial neural networks and sequential Gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils. Journal of Geochemical Exploration, 222, 106680. https://doi.org/10.1016/j.gexplo.2020.106680

Kohonen T. 1998. The self-organizing map. Neurocomputing, 21(1-3), 1–6. https://doi.org/10.1016/S0925-2312(98)00030-7

Kuhn S., Cracknell M.J., Reading A.M. 2018. Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. Geophysics, 83, B183–B193. https://doi.org/10.1190/geo2017-0590.1

Lawley C.J.M., Gadd M.G., Parsa M., Lederer G.W., Graham G.E., Ford A. 2023. Applications of natural language processing to geoscience text data and prospectivity modeling. Natural Resources Research, 32, 1503–1527. https://doi.org/10.1007/s11053-023-10216-1

Lawley C.J.M., Haynes M., Chudasama B., Goodenough K., Eerola T., Golev A., Zhang S.E., Park J., Lèbre E. 2024. Geospatial data and deep learning expose ESG risks to critical raw materials supply: the case of lithium. Earth Science, Systems and Society, 4. https://doi.org/10.3389/esss.2024.10109

Lehmann J., Brower A.M., Owen-Smith T.M., Bybee G.M., Hayes B. 2023. Landsat 8 and Alos DEM geological mapping reveals the architecture of the giant Mesoproterozoic Kunene Complex anorthosite suite (Angola/Namibia). Geoscience Frontiers, 14(5), 101620. https://doi.org/10.1016/j.gsf.2023.101620

Macambira M.J.B. 1983. Ambiente geológico e mineralizações associadas ao granito Serra Dourada (extremidade meridional) Goiás. MSc Dissertation, Universidade Federal do Pará, 132 p. Available online at: https://repositorio.ufpa.br/jspui/handle/2011/14907 / (accessed on 31 March 2025).

Marini J.O., Botelho N.F. 1986. A província de granitos estaníferos de Goiás. Revista Brasileira de Geociências, 16(1), 119–131. https://doi.org/10.25249/0375-7536.1986119131

Nagar S., Farahbakhsh E., Awange J., Chandra R. 2024. Remote sensing framework for geological mapping via stacked autoencoders and clustering. Advances in Space Research, 74(10), 4502–4516. https://doi.org/10.1016/j.asr.2024.09.013

Ng W., Minasny B., McBratney A., De Caritat P., Wilford J. 2023. Digital soil mapping of lithium in Australia. Earth System Science Data, 15(6), 2465–2482. https://doi.org/10.5194/essd-15-2465-2023

Parsa M., Lawley C.J.M., Cumani R., Schetselaar E., Harris J., Lentz D.R., Zhang S.E., Bourdeau J.E. 2024. Predictive modeling of Canadian carbonatite-hosted REE +/− Nb deposits. Natural Resources Research, 33, 1941-1965. https://doi.org/10.1007/s11053-024-10369-7

Pimentel M.M., Heaman L., Fuck R.A., Marini O.J. 1991. U-Pb zircon geochronology of Precambrian tin-bearing continental-type acid magmatism in central Brazil. Precambrian Research, 52(3-4), 321–335. https://doi.org/10.1016/0301-9268(91)90086-P

Pinto-Ward C. 2017. Controls on the enrichment of the Serra Verde rare earth deposit, Brazil. PhD Thesis, Imperial College of London, London, 442 p. https://doi.org/10.25560/78794

Prado E.M.G., Souza Filho C.R., Carranza E.J.M., Motta J.G. 2020. Modeling of Cu-Au prospectivity in the Carajás mineral province (Brazil) through machine learning : Dealing with imbalanced training data. Ore Geology Reviews, 124, 103611. https://doi.org/10.1016/j.oregeorev.2020.103611

Reis Neto J.M. 1980. Geocronologia dos granitos da região Centro-oeste. Seminários Gerais, Instinto de Geociências da USP, São Paulo, 106.

Reis Neto J.M. 1983. Evolução geotectônica da Bacia do Alto Tocantins, Goiás. MSc Dissertation, Universidade de São Paulo, São Paulo, 98 p. https://doi.org/10.11606/D.44.1983.tde-11092015-121945

Rodriguez-Galiano V., Sanchez-Castillo M., Chica-Olmo M., Chica-Rivas M. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest , regression trees and support vector machines. Ore Geology Reviews, 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rossi P., Andrade G.F., Cocherie A. 1992. The 1.58 Ga A-type granite of Serra da Mesa (GO): an example of NYF fertile granite pegmatite. In: Congresso Brasileiro de Geologia, 37, 389–390. Availabe online at: http://www.sbgeo.org.br/home/pages/44#Anais_de_Congressos_Brasileiros_de_Geologia / (accessed on 31 March 2025).

Santana I.V., Botelho N.F. 2022. REE residence, behaviour and recovery from a weathering profile related to the Serra Dourada Granite, Goiás/Tocantins States, Brazil. Ore Geology Reviews, 143, 104751. https://doi.org/10.1016/j.oregeorev.2022.104751

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. Journal of Geochemical Exploration, 155, 1–13. https://doi.org/10.1016/j.gexplo.2015.03.007

Saunders D.F., Burson R.K., Branch F.J., Thompson K. 1993. Relation of thorium-normalized surface and aerial radiometric data to subsurface petroleum accumulations. Geophysics, 58, 1417–1427. https://doi.org/10.1190/1.1443357

Silva A.B., Alves F.M. 2021. Atlas aerogeofísico do estado do Tocantins. Goiânia, CPRM. Available online at: https://rigeo.sgb.gov.br/handle/doc/22566 / (accessed on 31 March 2025).

Silva G.F., Larizzatti J.H., Silva A.D.R., Lopes C.G., Klein E.L., Uchigasaki K. 2022a. Unsupervised drill core pseudo-log generation in raw and filtered data, a case study in the Rio Salitre greenstone belt, São Francisco Craton, Brazil. Journal of Geochemical Exploration, 232, 106885. https://doi.org/10.1016/j.gexplo.2021.106885

Silva G.F., Silva A.M., Toledo C.L.B., Chemale Junior F., Klein E.L. 2022b. Predicting mineralization and targeting exploration criteria based on machine-learning in the Serra de Jacobina quartz-pebble-metaconglomerate Au-(U) deposits, São Francisco Craton, Brazil. Journal of South American Earth Sciences, 116, 103815. https://doi.org/10.1016/j.jsames.2022.103815

Silva G.F., Graça M.C. 2018. Gamma-ray attenuation caused by rainforest dispersion compared to Vegetation Index: estimates on the effects in airborne gamma-spectrometry data – example from the State of Rondônia, Amazonia, Brazil. Jounal of Geological Survey of Brazil, 1(1), 1–9. https://doi.org/10.29396/jgsb.2018.v1.n1.1

Teixeira L.M. 2002. Minerais portadores de elementos terras raras em granitos das subprovíncias Tocantins e Paranã-província estanífera de Goiás. PhD Thesis, Universidade de Brasília.

Teixeira L.M., Botelho N.F. 2006. Comportamento geoquímico de Etr durante evolução magmática e alteração hidrotermal de granitos: exemplos da província estanífera de Goiás. Revista Brasileira de Geociências, 36(4), 679–691. Availabe online at: https://ppegeo.igc.usp.br/portal/wp-content/uploads/tainacan-items/15906/46547/9310-11010-1-SM.pdf / (accessed on 31 March 2025).

Torppa J., Nykänen V., Molnár F. 2019. Unsupervised clustering and empirical fuzzy memberships for mineral prospectivity modelling. Ore Geology Reviews, 107, 58–71. https://doi.org/10.1016/j.oregeorev.2019.02.007

Vesanto J., Alhoniemi E. 2000. Clustering of the self-organizing map. IEEE Transactions on Neural Networks, 11(3), 586–600. https://doi.org/10.1109/72.846731

Vieira C.C., Botelho N.F., Garnier J. 2019. Geochemical and mineralogical characteristics of REEY occurrences in the Mocambo Granitic Massif tin-bearing A-type granite, central Brazil, and its potential for ion-adsorption-type REEY mineralization. Ore Geology Reviews, 105, 467–486. https://doi.org/10.1016/j.oregeorev.2019.01.007

Zapata A.M., Botelho N.F. 2018. Mineralogical and geochemical characterization of rare-earth occurrences in the Serra do Mendes massif, Goiás, Brazil. Journal of Geochemical Exploration, 188, 398–412. https://doi.org/10.1016/j.gexplo.2018.02.005