Journal of the Geological Survey of Brazil

Predictive lithological mapping through machine learning methods: a case study in the Cinzento Lineament, Carajás Province, Brazil

Iago Sousa Lima Costa[1], Felipe Mattos Tavares[2], Junny Kyley Mastop de Oliveira[3]

[1] CPRM - Geological Survey of Brazil. SBN, Quadra 02, Bloco H, 2° Andar, Distrito Federal, Brasília, Brazil, CEP: 70040-904.
[2] CPRM - Geological Survey of Brazil. Av. Pasteur, 404 - Urca, Rio de Janeiro, Rio de Janeiro, Brazil, CEP: 22290-240.
[3] CPRM - Geological Survey of Brazil. Av. Perimetral, 3645 - Marco, Pará, Belém, Brazil, CEP: 66095-904

Abstract

The Cinzento Lineament (Carajás Mineral Province) represents a complex deformational system with great associated mineral potential, mainly for IOCG deposits. However, the tropical vegetation of the Amazon rainforest considerably limits the number of outcrops available for systematic geological mapping. Therefore, the use of remote data such as airborne geophysics and remote sensing is essential to provide a reliable geological map. The airborne magnetometric data to define lithological units and its boundaries is a challenge, especially in regions with low magnetic latitude and/or remanent magnetization. In this work, we proposed an approach using Magnetization Vector Inversion (MVI) to map the distribution of the magnetic susceptibility, in order to replace techniques such as pole reduction and total gradient. We applied the Random Forest algorithm (supervised Machine Learning algorithm) to recognize patterns in remote data and improve the current mapped lithological units. With 1400 training samples (2.5% of the total samples), we produced two Predictive lithological maps: a first with remote data only and a second with remote data and spatial coordinates. We evaluate the advantages and disadvantages of each Predictive map, and we conclude that both maps need to be analyzed together for the refinement of the current geological map. These predictive maps represent a powerful tool to combine remote data to improve current geological maps, or even generate the first-pass geological map for regions with scarce geological knowledge.

Article Information

Publication type: Research paper
Submitted: 29 January 2019
Accepted: 19 March 2019
Online pub. 27 March 2019
Editor: Adalene M. Silva

Keywords:
Random Forest
Cinzento Lineament
Machine Learn
Airborne Geophysics

*Corresponding author
Iago Sousa Lima Costa
E-mail address:
iago.costa@cprm.gov.br

1. Introduction

The geological mapping in tropical areas with dense rain forest vegetation and thick supergene covers is challenging as rock exposure is very poor, and access to outcrops can be difficult. The use of remote sensing and airborne geophysics in these areas is essential to recognize and delimitate lithological units in the construction of geological maps, which are usually drawn manually by the interpreter. However, these maps are highly dependent on the prior knowledge of the interpreter; in other words, different interpreters can create different maps from the same data. The use of Machine Learning Algorithms (MLA) is a valuable tool to automatically recognize patterns among high-dimensional data to mitigate bias and to speed up interpretations, especially in regions where geophysical and remote-sensing data are widely available. Therefore, MLA represents an efficient method for producing greenfield geological maps or improving existing maps (e.g. Cracknell et al. 2014; Harris and Grunsky 2015; Kuhn et al. 2018; Radford et al. 2018).

The Carajás Mineral Province (CMP), in the southeastern Amazonian Craton, Brazil, is an example of such difficulties. Classic geological mapping at semi-detail scales (1:100.000, 1:50.000) is viable, especially in deforested areas, yet strongly based in remote sensing and airborne geophysics interpretation. Furthermore, vast areas covered by rainforest and with poor access are completely mapped through somehow biased geological-geophysical interpretation. The CMP is one of the largest mineral provinces in the world, with giant iron deposits and also rich in other resources such as copper, gold, PGE, chromium, nickel, manganese, REE, uranium and tin; which reinforces the necessity of delivering precise geological maps to the society, government and mining industry. Therefore, the Geological Survey of Brazil (CPRM) maintains a permanent geological mapping and metallogenic program in the CMP since 2008.

One of the areas recently surveyed by CPRM is the Cinzento Lineament region (Figure 1a), situated in the northernmost part of the Carajás Copper-Gold Belt (Tallarico et al. 2005). The Cinzento Lineament region hosts several IOCG (Iron Oxide – Copper – Gold), VMS (Volcanic Massive Sulfides), and granite-related copper-gold and polymetallic deposits, such as Salobo, Furnas, Pojuca and Paulo Afonso. Remote sensing and geophysical data were strongly used to elaborate the recently published geological map (Figure 1b - Oliveira et al.

ISSN: 2595-1939
https://doi.org/10.29396/jgsb.2019.v2.n1.3
Open access at jgsb.cprm.gov.br

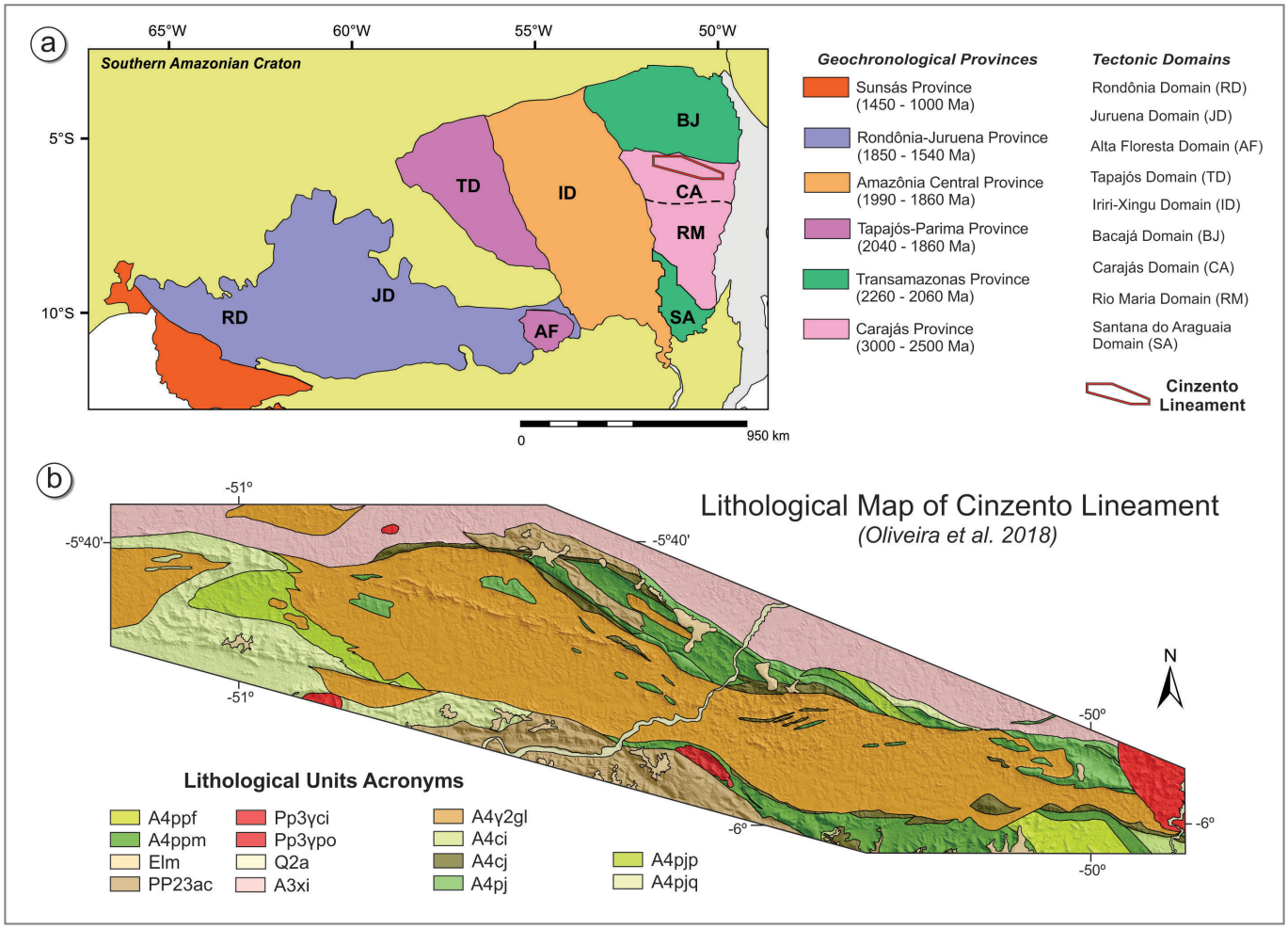

FIGURE 1 – a) Tectonic compartmentalization of the Southern Amazonian craton proposed by Vasquez et al. (2008) highlighting the region of the Cinzento Lineament. b) Lithological Map of Cinzento Lineament (Oliveira et al. 2018).

2018), as about 50% of the area comprises an environmental conservation unit with very limited road access.

In this work, we evaluated several supervised MLA in order to build a Predictive Lithological map of the Cinzento Lineament, based on a dataset that includes airborne magnetometric, radiometric and SRTM (Shuttle Radar Topography Mission) data. We built a predictive map with enough equivalence to the geological chart that was developed by combining traditional surveying tools (fieldwork and manual interpretation of remote sensing and airborne geophysical images).

2. Geological Framework

In general lines, there are three main rock crystallization or depositional ages in the CMP: in the Mesoarchean (3.06-2.83 Ga), Neoarchean (2.76-2.55 Ga) and Paleoproterozoic (2.10-1.88 Ga); all of them represented in the lithostratigraphic framework of the Cinzento Lineament region (Table 1). The Mesoarchean lithologies include orthogneisses, greenstone belts and granitoids developed under an accretionary-collisional system with peak metamorphism at 2.85 Ga (Tavares et al. 2018; Machado et al. 1991).

The Mesoarchean units are the basement of the Neoarchean Itacaiúnas Supergroup (DOCEGEO 1988), which comprises a volcano-sedimentary sequence with mafic and felsic volcanic rocks in its lowermost unit (Parauapebas Formation), followed by thick BIF layers (Carajás Formation) and, at the top, clastic and chemical metasedimentary rocks interbedded with fewer volcanic rocks and BIF (represented by the Igarapé Cigarra and Salobo-Pojuca formations in the study area). Coeval bimodal magmatism is represented by A-type granitic plutons and mafic-ultramafic intrusions, such as the Gelado Metagranite (Barbosa 2004). The Paleoproterozoic sedimentary cover is represented by the clastic Águas Claras Formation, while plutonic units include A-type granitic bodies of the Serra dos Carajás Intrusive Suite, such as the Pojuca and Cigano granites (Dall'Agnol et al. 2005).

The Mesoarchean and Neoarchean units are strongly structured in the WNW-ESE along the study area. For Costa and Siqueira (1990), this feature is the result of a long-term strike-slip fault system with multiple reactivations, namely the Cinzento Lineament. Tavares et al. (2017), however, reinterpreted the tectonic features of the Cinzento Lineament as the result of a complex multistage tectonic evolution, marked by the overlapping of several compressive and extensional phases. For Tavares et al. (2017), the main structures are understood as deep discontinuities related to the basement framework, reactivated as extensional faults during the deposition of the Itacaiúnas Supergroup and the

TABLE 1 – Simplified stratigraphic chart of the Cinzento Lineament region (Oliveira et al. 2018)

Age	Unit Code	Unit Name	Unit Lithotypes
Cenozoic	Q2a	Alluvial deposits	Sands, silts, clays, and gravel
	Elm	Lateritic Cover	Laterites
Paleoproterozoic	PP3γci	Cigano Granite	A-type monzogranites to sienogranites
	PP3γpo	Pojuca Granite	A-type monzogranites to sienogranites
	PP23ac	Águas Claras Formation	Siliciclastic sequence with intercalation of metarenites, metassiltites, and metargilites
Neoarchean	A4γ2gl	Igarapé Gelado Metagranite	A-type metamonzogranites and metagranodiorites
	A4ci	Igarapé Cigarra Formation	Clastic metasedimentary rocks with iron formations
	A4pj	Salobo-Pojuca Formation	Quartzites, shists, BIF
	A4pjp	Salobo-Pojuca Formation	Quartzites, metacherts, muscovite-quartz schist and quartz sericite schist
	A4pjq	Salobo-Pojuca Formation	Al-rich shists
	A4cj	Carajás Formation	BIF
	A4ppf	Felsic Parauapebas Formation	Felsic metavolcanic rocks, meta pyroclastics and chert
	A4ppm	Mafic Parauapebas Formation	Mafic to intermediate metavolcanic rocks
Mesoarchean	A3xi	Xingu Complex	Monzodioritic to tonalitic Orthogneisses

emplacement of the Gelado Metagranite, then again reactivated during the Paleoproterozoic Transamazonian Orogenic Cycle as ductile reverse shear zones, between 2.10 and 2.07 Ga (Tavares et al. 2018). All the tectonic system was later affected by a second Paleoproterozoic orogeny, known as the Sereno Event (1.98-1.95 Ga, see Tavares et al. 2018), of intracontinental character and ductile-brittle behavior. The collapse of the Paleoproterozoic orogenies also produced a later extensional reactivation, in a brittle, fluid-dominant environment, contemporary to the emplacement of the A-type plutons of the Serra dos Carajás Intrusive Suite.

3. Data

The airborne geophysical data are a compilation of three projects contracted by the Geological Survey of Brazil: Oeste de Carajás, Rio Maria and Tucuruí. These projects were carried out between 2003 and 2009 and have flight lines with a spacing of 500 m (N-S direction) and control lines with a spacing of 10 km (E-W direction). The airborne magnetic data were interpolated into a grid using the bi-directional method and the radiometric data with the minimum curvature method (Briggs 1974), both with a 125 m cell.

The airborne magnetic data are widely used in geological mapping due to the usually good variation of magnetic susceptibility in rocks. In machine learning applied to predictive lithology, the most commonly used magnetic processing is the Reduced-To-Pole (RTP - Baranov 1957; Gunn 1975) since the RTP anomalies have a clear spatial correlation with the lithological units. However, at low magnetic latitudes (as in the Cinzento Lineament region) the RTP shows poor performance, retaining part of the initial dipoles and creating biased artifacts (Silva 1986). Another approach is to use the Total Gradient (Nabighian 1972) which centralizes the anomaly above the magnetic source. However, the Total Gradient commonly shows lateral dispersions, making it difficult to define the geological boundaries between units sharply.

To overcome these issues, we suggest a new approach using a superficial horizontal slice from the Magnetic Vector Inversion (MVI - Ellis et al. 2012). The MVI directly inverts the magnetization vector accommodating effects of low latitudes and remanent magnetization (Johnson and Aisengart 2014). Figure 2a shows the result of the MVI to a depth of 100 meters.

The radiometric data are dependent on sources up to 30 cm beneath the Earth's surface, which makes these data highly correlated with geological field data. As the airborne radiometric data were obtained at different periods, they were initially compensated using a linear relationship between the data from an intersection area between adjacent projects. The radiometric data show the contents of Potassium (K), Uranium (eU) and Thorium (eTh), which were added to our database. We did not use ratios among the elements to not overparameterize the final model. Figure 2b displays an RGB ternary image with K (red), eTh (green) and eU (blue). Our database also included a digital elevation model derived from the SRTM (Figure 2c). Despite being very useful in other studies (e.g. Kuhn et al. 2018; Cracknell et al. 2014), our preliminary analysis showed a significant bias in Landsat-8 data (Figure 2d). The preliminary results showed a high correlation between lithological units covered by dense vegetation, due to its homogeneous coverage. Moreover, the lithological units appeared highly controlled in the boundaries between dense vegetation and anthropized areas (e.g. mines, tailings dams, cities). Consequently, the Landsat data were not used in predictive models.

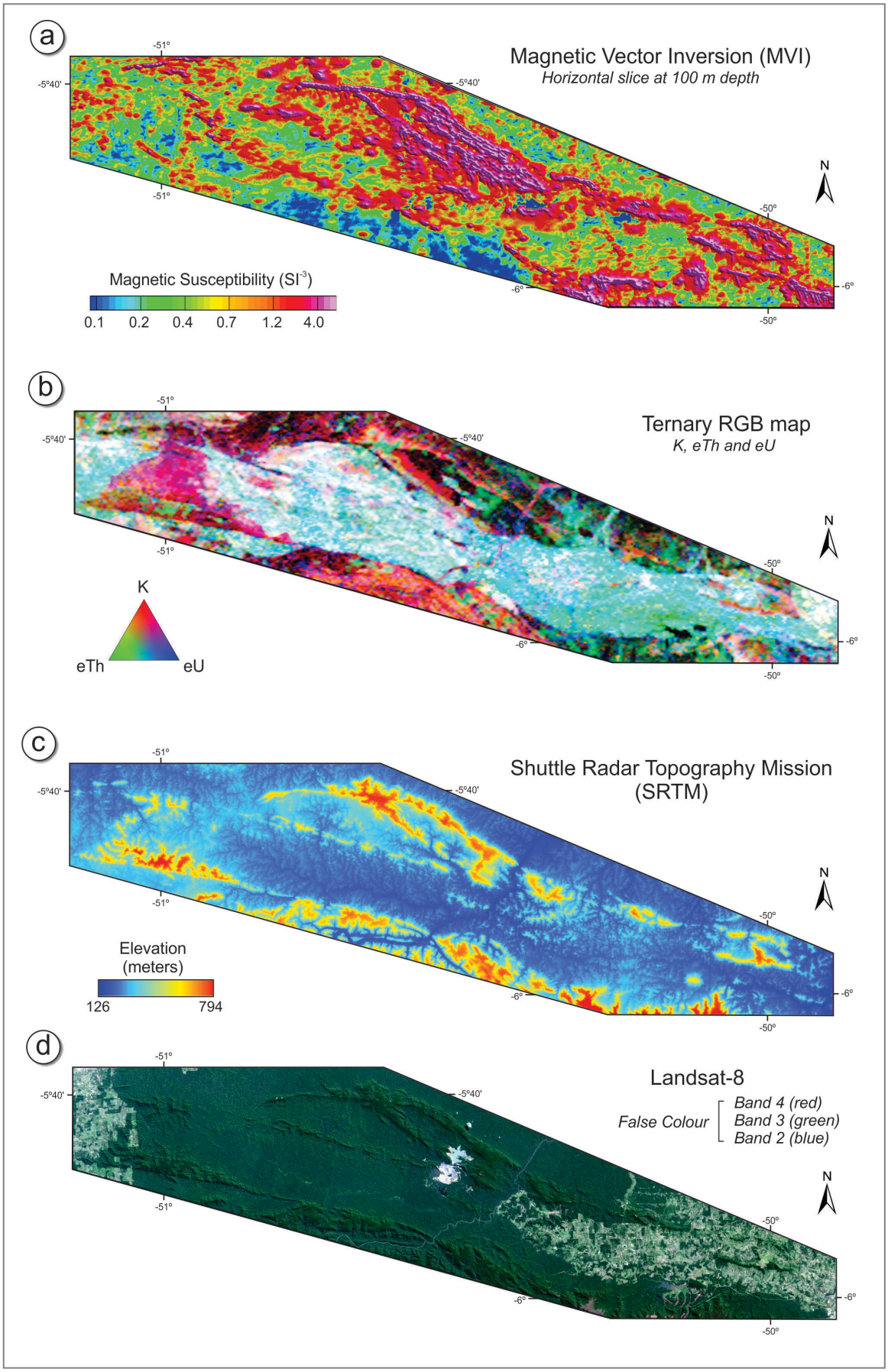

FIGURE 2 – The data used as input in the machine learning algorithms in this study. a) Magnetic Vector Inversion, b) Ternary RGB map (K, eTh and eU), C) Shuttle Radar Topography Mission (SRTM) and d) false color (RGB) of Landsat 8 combining bands 4 (red), 3 (green) and 2 (blue).

All data (airborne geophysical and SRTM) were reprojected to the coordinate system SIRGAS 2000 - UTM zone 22 south. Furthermore, we performed downsample to cells with 250 meters in all grids. This pre-processing resulted in 55,305 instances in which each one comprises all input data. Since we performed a supervised classification, it was necessary to take a small representative data do train the model (Hastie et al. 2009). In theory, records of field observation are the best data to compose training data. In practice, in a geological survey, the units are not sampled homogeneously, which can bias the model for the classes (or lithological unit) with the most significant number of samples (Japkowicz and Stephen 2002; Cracknell et al. 2014). Alternatively, it is possible to use random samples extracted from previous geological maps such as training data. Therefore, we followed previous works (Cracknell et al. 2014; Kuhn et al. 2018) and used 100 random samples per lithological unit. Another approach could be to define the number of samples by the area of the lithological unit; however, this procedure can bias the predictive model for classes with a large area (Japkowicz and Stephen 2002). Therefore, our training data resulted in 1400 samples (Figure 3) with their classes extracted randomly from the current geological map (Oliveira et al. 2018), which represents 2.5% of the total sample amount (55,305 samples). This rate is close to those of Cracknell et al. (2014) and Kuhn et al. (2018), which used 2.6% and 1.6% of the total sample amount, respectively.

4. Methods

4.1. Evaluation of Machine Learning Algorithms

Several supervised MLA can provide satisfactory predictions from a small number of training data samples. In this work, we perform an initial analysis of five MLA through k-fold Cross-validation Accuracy. This approach splits the instances into several folds and uses one fold at a time as a trainer to predict the remainder. Our work evaluated the Cross-validation Accuracy with ten folds for the MLAs: Naive Bayes, k-Nearest Neighbors (Cover and Hart 1967), Support Vector Machines (Vapnik 1998), Artificial Neural Networks and Random Forest (Breiman 2001). In agreement with Cracknell and Reading (2014), the Random Forest showed the best accuracy for predictions for our database (Table 2).

TABLE 2 – Evaluation of Machine Learning Algorithms through Cross-validation Accuracy (%).

Machine Learning Algorithm	Cross-validation Accuracy (%)
Random Forest	76.9
Neural Network	74.1
Naive Bayes	67.3
k-Nearest Neighbors	64.2
Support Vector Machines	25.9

4.2. Random Forest Algorithm

The Random Forest algorithm (RF) is an ensemble classification method that uses bootstrap aggregation to create multiple decision trees (Breiman 2001). This approach randomly selects 2/3 of the training samples with replacement to generate a data set ("in-bag" data) of the same size as the training data. The in-bag data classify a decision tree through the Gini index (Breiman et al. 1984) that determines the best parting for a given class, while the remaining data ("out-of-bag" data) are used to validate the model. The significant advantage of the RF is that the prediction of a class is a function of the average of the multiple decisions trees, thus improving the predictions and mitigating errors from outliers. In our predictive model, we have reached a suitable accuracy with 100 trees without a maximum limit in depth for each one (Table 3).

Figure 4 shows some of the multiple trees used in this work viewed as Pythagorean trees (Beck et al. 2014). In this representation, the size of the rectangles is proportional to the number of samples, and the most probably class defines the colors.

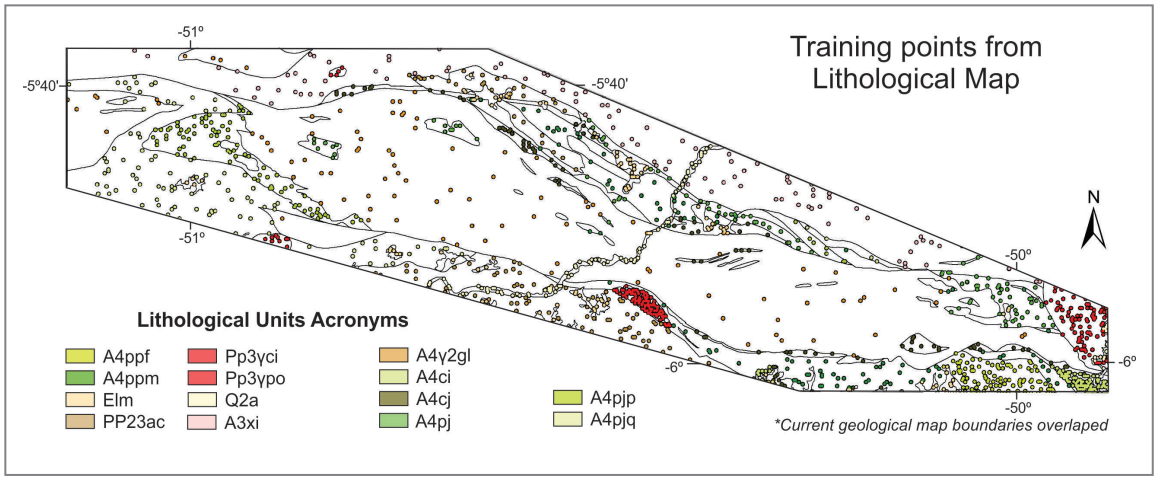

FIGURE 3 – Location of training points used as input in the Machine Learning Algorithms. The training points database is composed of 100 random samples from each class (or lithological unit) resulting in 1400 samples.

The RF also can provide a rank of the input variables concerning the importance in the predictions. Therefore, we perform some accuracy cross-validation tests for each addition of a variable follow the rank (Table 4). The Thorium (eTh) has shown to have the most significant importance in the predictions, which is compatible since it represents the radioelement with less mobility and consequently the most related to the lithological units.

TABLE 3 – Evaluation of Cross-validation Accuracy (%) with the number of trees. The bold text indicates the point where the accuracy stops progressing.

Number of Trees	Cross-validation Accuracy (%)
1	0.65
10	0.79
50	0.825
100	**0.829**
500	0.824
1000	0.826

TABLE 4 – Cross-validation accuracy % for each influence rank in the Random Forest.

Rank	Data	Cross-validation Accuracy
1	eTh	47.2 %
2	SRTM	60.7 %
3	eU	64.4 %
4	MVI	71.8 %
5	K	76.9 %

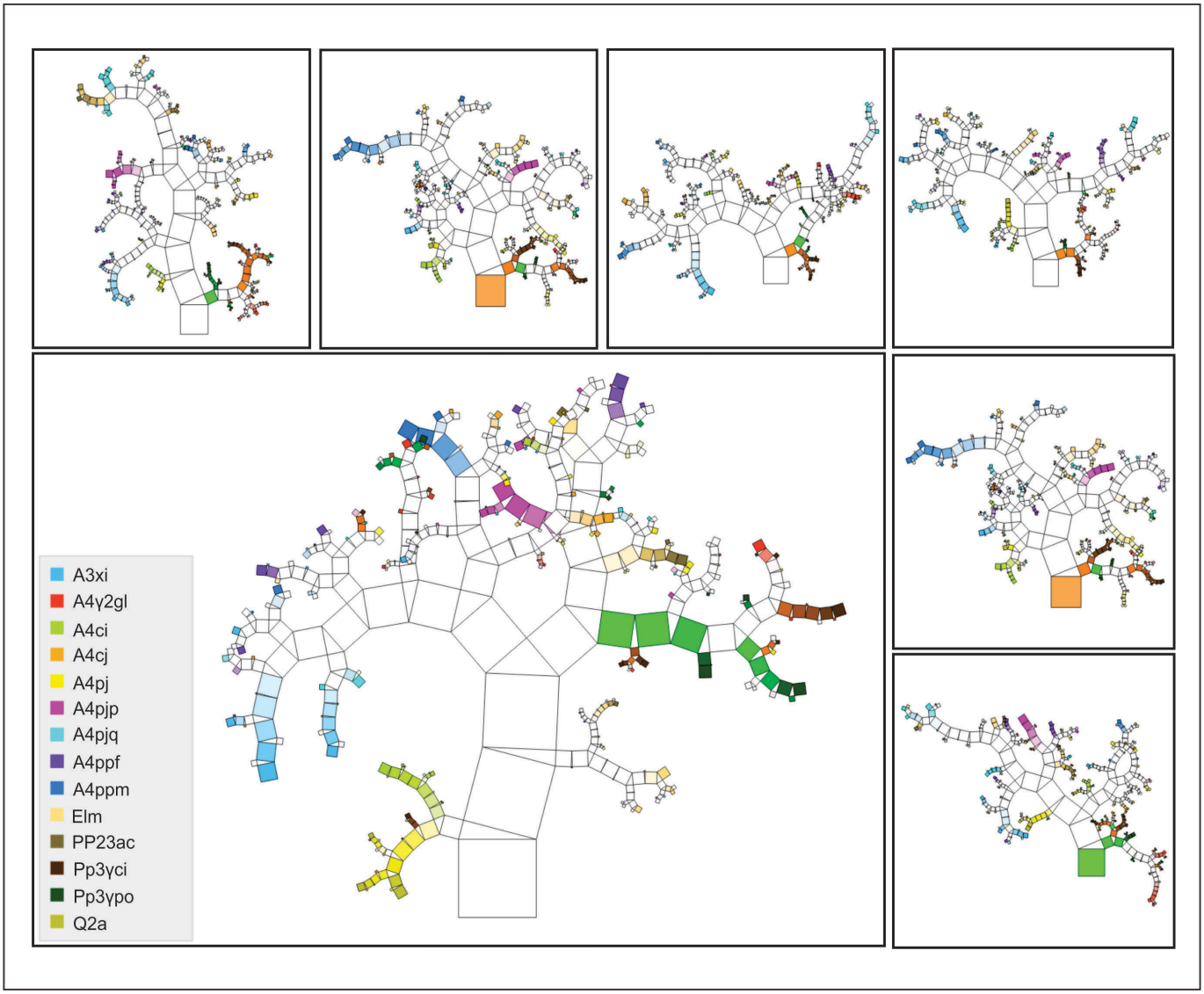

FIGURE 4 – Examples of Pythagorean trees from the multiple trees used in this work. Colors show the most likely classes on each node and the size of the rectangle is proportional to the number of samples in each group.

5 Results

The Random Forest algorithm shows, as a result, a class probability map for each class defined in the training data (e.g. Figure 5). These probabilities can be spatially categorized in the most probable class creating a Predictive map. Figure 6 shows the comparison between the current geological map (Figure 6a), the spatial distribution of the training points (Figure 6b) and the RF Predictive Lithological map with airborne geophysics and SRTM (Figure 6c). The RF Predictive map with remote data predicted 52.7% of the current geological data (correct predictions/total predictions). This mismatch between the geological map and the Predictive map may represent misclassified samples in the training data, inappropriate predictions and/or the degree of new information that can be taken from the remote data to improve the current geological map.

Table 5 shows the predicted percentage for each lithological unit and the second most related. The smaller classes (and fewer samples) tend to have a better recovery due to its lower variation. For example, the Paleoproterozoic Pojuca Granite (PP3γpo) has a predicted percentage of 81.1%; however, the sample amount of this unit (106 samples) is remarkably lower concerning larger units as the Igarapé Gelado Metagranite (A4γ2gl) which has 23,553 samples. In this case, the Igarapé Gelado Metagranite represent a lithological unit with considerable compositional variation, and in some regions has similarities with other units such as Cigano Granite (PP3γci - 11.8%). In the same way, a small part of the occurrence of felsic metavolcanic rocks (A4ppf) presents great similarities with orthogneisses of the Xingu Complex (A3xi - 19.7%). These regions may represent a class error; however, they also may represent lithofacies variations or small contrasting lithologies within the Igarapé Gelado Metagranite (A4γ2gl) or portions of orthogneisses preserved within felsic metavolcanic rocks (A4ppf), respectively.

Despite the satisfactory recovery (52.7%), the Predictive map with remote data (Figure 6c) presents a considerable content of high-frequency noise. Cracknell and Reading (2014) proposed that the use of spatial coordinate constraints could significantly reduce this high-frequency content. Therefore, we inserted constrains containing the spatial coordinates of the instances in order to mitigate noise and improve model recovery. Figure 6d shows the Predictive Lithological map with remote data and spatial correlation. This predictive map recovered 78.7% of the geological map, which represents a compatible recovery in comparison with previous works. For example, Yu et al. (2012) obtained a prediction of 62.2% using Support Vector Machines (SVM) and an 11 x 11 majority filter. Kuhn et al. (2018) and Cracknell et al. (2014) through the Random Forest algorithm predicted 76% and 78%, respectively. Table 6 shows a substantial increase in recovery when spatial constraints are inserted. The predicted percentage ranges from 59.6 to 98.1%, while for the predictive map without spatial correlation it varies between 36.9 and 81.1% (Table 5).

6. Discussions

The Random Forest produced two predictive maps: one with remote data only (Figure 6c), and one with remote data and spatial coordinates (Figure 6d). In the first case (only with remote data), the predictive map shows a considerable amount of noise; however, it also showed potential areas for reevaluation. In the northwest portion, the geological map (Figure 6a) shows an expressive occurrence of the Igarapé Gelado Metagranite (A4γ2gl), however, the predictive map shows that this region has more similarities with the Mafic Parauapebas (A4ppm) or with Salobo-Pojuca (A4pj / A4pjp) formations or even shows that the occurrence of metavolcano-sedimentary rocks in the northwest portion is much more common than in the southeast portion as is shown on the current geological map (Oliveira et al. 2018). Considering that this entire northwestern portion comprises the area of environmental protection and thus was barely investigated with the geological mapping, the current interpretation of this portion of the map is mainly reflected by the interpretation of the remote sensors and airborne geophysical images. Therefore, it is suggested to re-evaluate the possibility to improve the interpretation with the new data obtained with the predictive map of this work. Also in the northwest, the

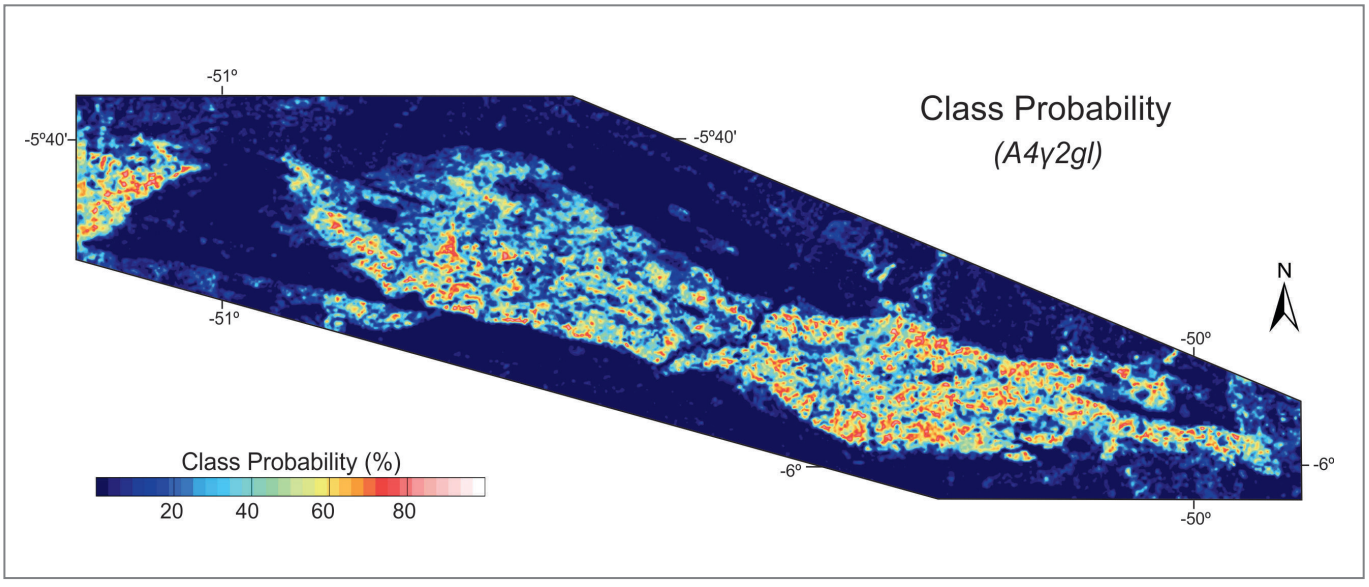

FIGURE 5 – An example of class probability (in percentage) for the lithological unit A4γ2gl from the Predictive Lithological map with remote data only.

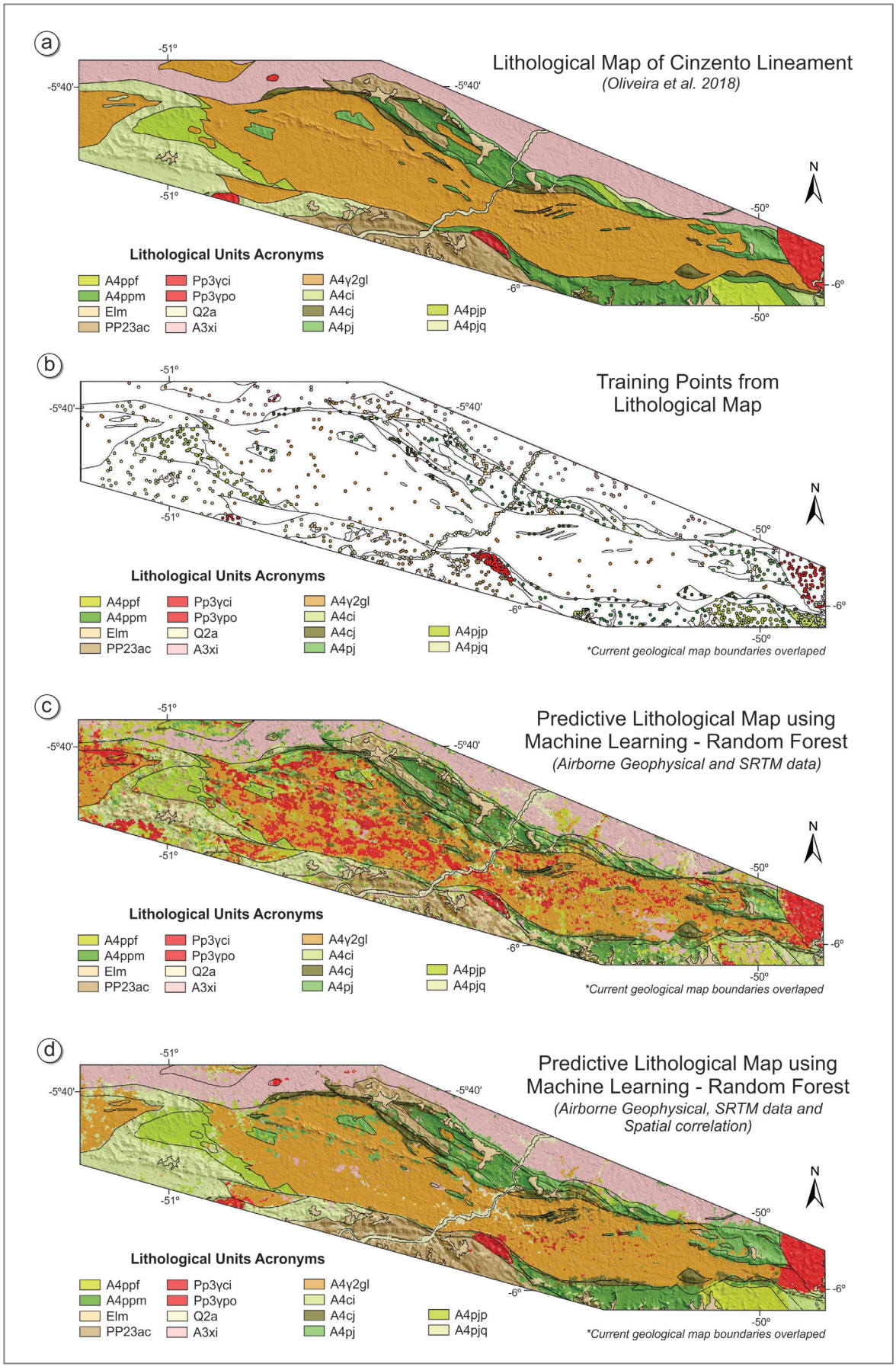

FIGURE 6 – Comparison between the a) Lithological Map of Cinzento Lineament from Oliveira et al. (2018), b) training points, c) Predictive Lithological Map using remote data only, and d) the Predictive Map with remote data and spatial correlation. The limits of the geological map were superimposed on the predictive map to improve the association.

TABLE 5 – The recovery of the predicted samples of the Predictive Lithological Map using remote data only. The last column shows the unit that most relates to the main class.

Lithological Unit	Actual Samples	Predicted Samples	Predicted Percentage	Second Most Related Unit
PP3γpo	106	86	81.1 %	PP3γci (8.5 %)
Elm	899	716	79.6 %	PP23ac (5.9 %)
Q2a	438	331	75.6 %	PP3γci (6.8 %)
A4ci	138	95	68.8 %	A4cj (8.0 %)
A3xi	9,600	6,407	66.7 %	A4ppf (12.5 %)
PP3γci	985	613	62.2 %	A4γ2gl (11.2 %)
A4pjp	1,702	1,000	58.6 %	PP23ac (10.6 %)
PP23ac	3,883	2,161	55.7 %	Elm (11.3 %)
A4γ2gl	23,553	12,066	51.2 %	PP3γci (11.8 %)
A4cj	1,481	684	46.2 %	A4ppm (13.4 %)
A4ppf	834	383	45.9 %	A3xi (19.7 %)
A4pjq	5,530	2,310	41.8 %	A4ppf (12.1 %)
A4pj	1,852	752	40.6 %	A4cj (11.9 %)
A4ppm	4,452	1,641	36.9 %	A4cj (13.3 %)

TABLE 6 – The recovery of the predicted samples of the Predictive Lithological Map using remote data and spatial correlation. The last column shows the unit that most relates to the main class.

Lithological Unit	Actual Samples	Predicted Samples	Predicted Percentage	Second Most Related Unit
A4ppf	834	818	98.1 %	A4cj (0.8 %)
PP3γpo	106	101	95.3 %	PP23ac (2.8 %)
A4ci	138	128	92.8 %	A4ppm (4.8 %)
PP3γci	985	884	89.7 %	A4γ2gl (3.1 %)
Elm	899	785	87.3 %	A4pjq (4.8 %)
Q2a	438	377	86.1 %	A4ci (5.3 %)
A4pjp	1,702	1,461	85.6 %	A4pjq (5.7 %)
A3xi	9,600	8,188	85.3 %	Q2a (2.7 %)
PP23ac	3,883	3,075	79.2 %	Elm (7.5 %)
A4pjq	5,530	4,329	78.3 %	A4pjp (9.5 %)
A4γ2gl	23,553	18,452	78.3 %	A4cj 6.7 %)
A4pj	1,852	1,446	78.1 %	A4γ2gl (5.1 %)
A4cj	1,481	957	64.5 %	A4ppm (9.5 %)
A4ppm	4,452	2,652	59.6 %	A4cj (9.5 %)

geological map shows an intrusive Paleoproterozoic body related to Cigano Granite (PP3γci); however, the predictive map does not recognize this Paleoproterozoic body and proposes a Salobo-Pojuca Formation (A4pj / A4pjp) instead. The predictive map with remote data only showed several new portions of Laterite Cover (Elm) that can be validated by its clear signature in the SRTM (plateaus) and radiometric data (High eTh with low K and eU).

The predictive map with remote data and spatial coordinates has considerably increased the recovery compared to the geological map (78.7 %). This predictive map correlated very well with the geological map and proposed some changes, which can bring a considerable gain in the refinement in the boundaries of the geological units. However, this map brought little information about new lithological units; in other words, the insertion of the spatial correlation may have suppressed, besides noise, several features with possible geological logic. Furthermore, small units such as Pojuca Granite (PP3γpo), which have a dense cluster of sample training, seem to attract solutions that do not fit the unit.

7 Conclusions

The machine learning methods allow bringing a new approach for data interpretation. Although the definition of the parameters can produce bias, this approach allows reproducing the predictive models since the parameters are defined. In this work, the Random Forest Algorithm showed the best performance among several methods of machine learning, in agreement with Cracknell and Reading (2014). The Random Forest make it possible to generate two predictive maps: one with remote data only, and another with remote data and a spatial constraint. The first one showed a recovery of 52.7 % concerning the geological map besides several possible new lithological units, however with a high noise content. The second one obtained a better recovery of 78.7 % with several suggestions in the reevaluation of the limit of several units; however, did not bring information about new lithologies. In fact, both Predictive maps with remote data only and with spatial coordinates have their advantages and disadvantages and need to be used together to enrich the geological map, especially in regions with scarce outcrops such as the Cinzento Lineament region.

Acknowledgments

We thank the Geological Survey of Brazil (CPRM) for providing the airborne geophysical data as well as the necessary infrastructure for the development of this work. Special thanks to the Ph.D. student Steve Kuhn (University of Tasmania) for the valuable discussions on the methodology. We also thank the reviewers for their helpful comments. The python library Orange Data Mining (Demsar et al. 2013) implemented the machine learning algorithms used in this work.

References

Baranov V. 1957. A new method for interpretation of aeromagnetic maps: pseudo-gravimetric anomalies. Geophysics, 22(2), 359-383. https://doi.org/10.1190/1.1438369

Barbosa J.P.O. 2004. Geologia estrutural, geoquímica, petrografia e geocronologia de granitóides da região do Igarapé Gelado, norte da Província Mineral de Carajás. Msc Dissertation, Centro de Geociências, Universidade Federal do Pará, Belém, Pará, 105 p.

Beck F., Burch M., Munz T., Di Silvestro L., Weiskopf D. 2014. Generalized Pythagoras Trees for visualizing hierarchies. In: International Conference on Information Visualization Theory and Applications, IEEE, 5, 17-28. https://doi.org/10.1007/978-3-319-25117-2_8

Breiman L. 2001. Random Forest. Machine Learning, Springer, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman L., Friedman J., Olshen R.A., Stone C.J. 1984. Classification and regression trees, the wadsworth statistics and probability series. Wadsworth International Group, Belmont California, 358 p.

Briggs I.C. 1974. Machine Contouring Using Minimum Curvature. Geophysics, 39(1), 39-48. https://doi.org/10.1190/1.1440410

Costa I.S.L., Oliveira J.K.M., Tavares F.M., de Paula R.R. 2017. Caracterização geofísica de depósitos IOCG ao longo do Lineamento Cinzento, Província Mineral de Carajás. In: Simpósio de Geologia Da Amazônia, 15, Belém, Pará.

Costa J.B.S., Siqueira J.B. 1990. Transtração e transpressão ao longo do lineamento Cinzento (região da serra dos Carajás). Revista Brasileira de Geociências, 20(1-4), 234-238. https://doi.org/10.25249/0375-7536.1990234238

Cover T.M., Hart P.E. 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Cracknell M., Reading, A. 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

Cracknell M., Reading A., McNeill A.W. 2014. Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer-Mt Charter region, Tasmania, using Random Forests (TM) and Self-Organising Maps. Australian Journal of Earth Sciences, 61, 287-304. https://doi.org/10.1080/08120099.2014.858081

Dall'Agnol R., Teixeira N., Ramo O., Moura C., Macambira M., Oliveira D. 2005. Petrogenesis of the Paleoproterozoic rapakivi A-type granites of the Archean Carajás metallogenic province, Brazil. Lithos, 80(1), 101-129. https://doi.org/10.1016/j.lithos.2004.03.058

Demsar J., Curk T., Erjavec A., Gorup C., Hocevar T., Milutinovic M., Mozina M., Polajnar M., Toplak M., Staric A., Stajdohar M., Umek L., Zagar L., Zbontar J., Zitnik M., Zupan B. 2013. Orange: data mining toolbox in Python. Journal of Machine Learning Research, 14, 2349-2353.

Docegeo 1988. Revisão litoestratigráfica da província mineral de Carajás Litoestratigrafia e principais depósitos minerais. In: Congresso Brasileiro de Geologia, Sociedade Brasileira de Geologia, 35, Belém, Brazil.

Ellis R., de Wet B., Macleod I. 2012. Inversion of magnetic data from remanent and induced sources. In: ASEG Conference and Exhibition, 22, Brisbane, Australia. https://doi.org/10.1071/ASEG2012ab117

Gunn P.J. 1975. Linear transformations of gravity and magnetic fields. Geophysical Prospecting, 23(2), 300-312. https://doi.org/10.1111/j.1365-2478.1975.tb01530.x

Harris J., Grunsky E. 2015. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. Computers & Geosciences, 80, 9-25. https://doi.org/10.1016/j.cageo.2015.03.013

Hastie T., Tibshirani R., Friedman J.H. 2009. The elements of statistical learning: data mining, inference and prediction. Second edition. New York, Springer. https://doi.org/10.1007/978-0-387-84858-7

Japkowicz N., Stephen S. 2002. The class imbalance problem: a systematic study. Intelligent Data Analysis, 6(5), 429-449. https://doi.org/10.3233/IDA-2002-6504

Johnson A., Aisengart T. 2014. Interpretation of magnetic data at low magnetic latitudes using magnetization vector inversion. Journal of Geophysics, 35(3), 91-96.

Kuhn S., Cracknell M., M. Reading A. 2018. Lithological mapping using Random Forests applied to geophysical and remote sensing data: a demonstration study from the Eastern Goldfields of Australia. Geophysics, 83(4), 1-37. https://doi.org/10.1190/geo2017-0590.1

Machado N., Lindenmayer D.H., Krough T.E., Lindenmayer Z.G. 1991. U-Pb geochronology of Archean magmatism and basement reactivation in the Carajás area, Amazon Shield, Brazil. Precambrian Research, 49(3), 329-354. https://doi.org/10.1016/0301-9268(91)90040-H

Nabighian M.N. 1972. The analytic signal of two-dimensional magnetic bodies with polygonal cross-section - Its properties and use of

automated anomaly interpretation. Geophysics, 37(3), 507-517. https://doi.org/10.1190/1.1440276

Oliveira J.K.M., Tavares F.M., Costa I.S.L. 2018. Mapa Geológico-Geofísico do Lineamento Cinzento, Escala 1:100.000, Serviço Geológico do Brasil, Belém. Available online at: http://geosgb.cprm.gov.br / (accessed on 13 December 2018).

Radford D.D.G., Cracknell M., Roach M., Cumming G. 2018. Geological Mapping in Western Tasmania Using Radar and Random Forests. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(9), 3075-3087. https://doi.org/10.1109/JSTARS.2018.2855207

Silva J.B.C. 1986. Reduction to pole as an inverse problem and its application to low-latitude anomalies. Geophysics, 51(2), 369-382. https://doi.org/10.1190/1.1442096

Tallarico F.H.B., Figueiredo B.R., Groves D.I., Kositcin N., McNaughton N. J., Fletcher I. R., Rego J. L. 2005. Geology and SHRIMP U-Pb geochronology of the Igarapé Bahia deposit, Carajás copper-gold belt, Brazil: An Archean (2.57 Ga) example of iron-oxide Cu-Au-(U-REE) mineralization. Economic Geology, 100(1), 7-28. https://doi:10.2113/100.1.0007

Tavares F.M., Trouw A.J.R., da Silva C.M.G., Justo A.P., Oliveira J.K.M. 2018. The multistage tectonic evolution of the northeastern Carajás Province, Amazonian Craton, Brazil: revealing complex structural patterns. Journal of South American Earth Sciences, 88, 238-252. https://doi.org/10.1016/j.jsames.2018.08.024

Tavares, F.M., Costa, I.S.L., Oliveira, J.K.M. 2015. Controles estruturais das mineralizações de cobre-ouro do lineamento cinzento, Província Mineral de Carajás. In: Simpósio de Geologia Da Amazônia, 15, Belém, Pará.

Tavares F.M., Oliveira J.K.M., de Paula R.R., Costa I.S.L., Prado E.B. dos S. 2017. O Cinturão norte do cobre da Província Mineral de Carajás: épocas metalogenéticas e controles críticos das mineralizações. In: Simpósio de Geologia Da Amazônia, 15, Belém, Pará.

Vapnik V. 1998. Statistical learning theory. New York, NY Wiley, 736 p.

Vasquez M.L., Rosa-Costa L.T., Silva C.M.G., Klein E. L. 2008. Compartimentação tectônica. In: Vasquez M.L., Rosa-Costa L.T. (ed.). Geologia e recursos minerais do Estado do Pará: Sistema de Informações Geográficas - SIG: texto explicativo dos mapas geológico e tectônico e de recursos minerais do estado do Pará. Belém, Serviço Geológico do Brasil, 39-112.

Yu L., Porwal A., Holden E.-J., Dentith M. 2012. Towards automatic lithological classification from remote sensing data using support vector machines. Computers & Geosciences, 45, 229-239. https://doi.org/10.1016/j.cageo.2011.11.019