Unsupervised drill core pseudo-log generation in raw and filtered data, a case study in the Rio Salitre greenstone belt, São Francisco Craton, Brazil

Guilherme Ferreira da Silva (guilherme.ferreira@cprm.gov.br)
Geological Survey of Brazil (CPRM)  https://orcid.org/0000-0002-3675-7289

João Henrique Larizzatti
Geological Survey of Brazil (CPRM)

Anderson Dourado Rodrigues da Silva
Geological Survey of Brazil (CPRM)

Carina Graciniana Lopes
Geological Survey of Brazil (CPRM)

Evandro Luiz Klein
Geological Survey of Brazil (CPRM)

Kotaro Uchigasaki
Geological Survey of Brazil (CPRM)

Research Article

Keywords: K-means, Model-Based Clustering, Unsupervised Learning, Portable X-Ray Fluorescence

DOI: https://doi.org/10.21203/rs.3.rs-871865/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Unsupervised drill core pseudo-log generation in raw and filtered data, a case study in the Rio Salitre greenstone belt, São Francisco Craton, Brazil

Guilherme Ferreira da Silva (Corresponding author)
Directory of Geology and Mineral Resources, Geological Survey of Brazil – CPRM,
ORCiD https://orcid.org/0000-0002-3675-7289
guilherme.ferreira@cprm.gov.br

João Henrique Larizzatti
Directory of Geology and Mineral Resources, Geological Survey of Brazil – CPRM,
ORCiD https://orcid.org/0000-0002-6778-4132
joao.larizzatti@cprm.gov.br

Anderson Dourado Rodrigues da Silva
Directory of Geology and Mineral Resources, Geological Survey of Brazil – CPRM,
ORCiD https://orcid.org/0000-0001-9841-4225
anderson.rodrigues@cprm.gov.br

Carina Graciniana Lopes
Directory of Geology and Mineral Resources, Geological Survey of Brazil – CPRM,
ORCiD https://orcid.org/0000-0001-8150-3029
carina.lopes@cprm.gov.br

Evandro Luiz Klein
Directory of Geology and Mineral Resources, Geological Survey of Brazil – CPRM,
ORCiD https://orcid.org/0000-0003-4598-9249
evandro.klein@cprm.gov.br

Kotaro Uchigasaki
Directory of Geology and Mineral Resources, Geological Survey of Brazil – CPRM,
ORCiD https://orcid.org/0000-0001-9199-6322
kotaro.uchigasaki@cprm.gov.br
Abstract

We use in situ portable X-Ray Fluorescence data acquired in sawn drill core samples of rocks from the Sabiá prospect, at the Rio Salitre greenstone belt, São Francisco Craton Brazil, for pseudo-log automatic generation through running unsupervised learning models to group distinct lithotypes. We tested the K-means and Model-Based Cluster (MBC) algorithms and compared their performance in the raw and filtered data with a manual macroscopic log description. From the initial 47 available elements, 20 variables were selected for modeling following the criteria of presenting at least 95% of uncensored values. Additionally, we performed a Shapiro-Wilk test that confirmed a non-parametric distribution by verifying the P-value attribute less than the 5% significance level. We also checked if the dataset's distribution was statistically equivalent to the duplicates with the assistance of a Kruskal-Wallis test, which would confirm the representativity power of the measurements at the same 5% significance level. After this step, the pseudo-log models were created based on reduced dimension data, compressed by a centered Principal Component Analysis with data rescaled by its range. Concerning to reduce the high-frequency noise in the selected features, we employed an exponential weighted moving average filter with a window of five samples. By the analysis of the Average Silhouette Width on sample space, the optimum number for K-means was fixed in two, and then the first models were generated for raw and filtered data. From the MBC perspective, the sample space is interpreted as a finite mixture of groups with distinct Gaussian probability distribution. The number of clusters is defined by the analysis of the Bayesian Information Criteria (BIC), where several models are tested, and the one in the first local maximum defines the number of groups and the type of probabilistic model in the simulation. For the data used in this work, the optimum group number for MBC is four, and the probabilistic model type determined by the BIC is elliptical with equal volume, shape, and orientation. Thus, Model-Based Cluster has detected four different cluster groups with almost the same
representativity for the two drill cores' samples. All K-means and MBC models were able to
detect changes in lithotypes not described in the manual log. On the other hand, one lithotype
described by the experts was not detected by this methodology in any attempt. It was needed a
detailed investigation with thin section descriptions to determine the cause of this response.
Finally, compared with the manual log description, it is notable that the models built on filtered
data have better performance than those generated on raw data, and the MBC filtered model
had better performance than the others. Hence, this multivariate approach allied to filtering the
data with a moving average transformation can be a tool of great help during several stages of
mineral exploration, either in the creation of pseudo-log models prior the description of the drill
core samples or in the data validation stage, when it is necessary to standardize several
descriptions made by different professionals.

**KEYWORDS:** K-means, Model-Based Clustering, Unsupervised Learning, Portable X-Ray
Fluorescence
1. Introduction

The "data-rich paradigm" is already a reality in mineral exploration. This scenario can be found in several segments of the mineral industry, such as airborne geophysics, exploratory geochemical surveys, mineral resources, and reserves analyses evaluation, studies of physical properties of the rock, mechanical assays of mine engineering, ore grade control, and environmental monitoring, among many other study branches associated with various stages in the mineral research. Moreover, some government entities (Agencies and Geological Surveys) also provide valuable data for the mineral industry, increasing the available volume of information. In this data-dominated scenario, fast, consistent, and reliable analysis is vital for decision making and resource investment.

One of the mineral industry's main challenges in practically all stages of exploration is to manage and integrate information from large data sets. Log description is a qualitative analysis based on the geologist's experience, and sometimes challenging to match together, mainly when several professionals describe a large volume of samples over a long period. Also, the drill core manual interpretation can be very time-consuming and challenging, especially when integrating into the analysis of multiple variables (Hill et al., 2020).

Despite some limitations, portable X-Ray Fluorescence analysis (pXRF) can be useful in numerous scenarios, whether to increase the spatial resolution of laboratory data or to assist in real-time or near-real-time support for operational decisions in drill core management (Fisher et al., 2014). The significant advantages of the pXRF over conventional laboratory analysis are the speed of data collection and the low cost of the analyses (Le Vaillant et al., 2014), which permits the compilation of a high amount of spatially representative information.

Machine Learning Algorithms (MLA) has proved to be a useful tool in managing geological data in several instances. Some of the most notable cases of MLA applied to geosciences focus on a diverse range of topics, such as prediction of hydrothermal alteration, automatic detection
of basement rocks in drill-hole, refining lithological mapping using airborne geophysics, and identifying mineral prospectivity by the integration of regional data (Bérubé et al., 2018; Costa et al., 2019; Hill and Uvarova, 2018; Kuhn et al., 2018; Rodriguez-Galiano et al., 2015; Wedge et al., 2019).

In this work, we present an approach to assist the drill core management with a fast, consistent, and highly reproducible methodology based on open-source code and data already acquired, which sometimes represents a challenge due to survey issues and/or the high number of geochemical features. For this purpose, we employed unsupervised Machine Learning methods in portable X-Ray Fluorescence data of drill core rock samples to generate pseudo-log. The term pseudo-log is used here in the sense of automatic classification of the rock samples stacked in the same order of the drill core depth and not validated by a specialist. We tested and compared the performance of two standard clustering algorithms, K-means, and the MBC, in raw and transformed data, filtered by the EWMA convolution. The clustering models were performed in results obtained from two drill cores near the sulfide-rich Sabiá prospect (Ribeiro et al., 1993), in the Rio Salitre greenstone belt (RSGB), São Francisco Craton, located in Bahia, northeastern Brazil (Figure 1).
The RSGB is located in the São Francisco Craton (SFC), a major geotectonic unit of the South American platform, placed in the southeast and northeast regions of Brazil, and stabilized during the Neoproterozoic Brasiliano cycle of orogenies. (Almeida et al., 1981; Heilbron et al., 2017). The SFC comprises a large number of Archean to Paleoproterozoic TTG-gneisses, granitoids, and greenstone belts, which consist of various sialic units, some of them named Gavião, Serrinha, and Jequié blocks (Alkmim and Martins-Neto, 2012; Alkmim et al., 1993; Barbosa and Barbosa, 2017; Barbosa and Sabaté, 2004; Barbosa, 2012; Cordani et al., 1992; Oliveira et al., 2010; Teixeira et al., 2017, 1996). The Gavião block, north of the São Francisco Craton, comprises a variety of TTG, granitoids, migmatites, and greenstone belt-like
metavolcano-sedimentary associations (e.g., Rio Salitre, Brumado, Boquira, and Riacho de Santana, Barbosa, 2012).

The RSGB is a Paleoproterozoic metavolcano-sedimentary sequence located in the northernmost portion of the Gavião Block (Barbosa and Barbosa, 2017; Garcia, 2017; Teixeira et al., 2017). The Rio Salitre consists of two units, from bottom to top: i) *Baixo Vale do Rio Salitre*: metapelites, metagraywacke, and meta-arkose, interlayered with mafic-ultramafic and felsic metavolcanic rocks and ii) *Sobradinho*: composed of banded iron formations, paragneiss, phyllites, mica schists, metabasic and meta-ultrabasic rocks, calc-silicate rock, and quartzites. These units were deformed and metamorphosed at low greenschist facies metamorphic grade.

The RSGB has potential for base metals (Cu-Pb-Zn) associated with VHMS and orogenic gold type deposits (Barbosa, 2012). Several occurrences of Cu-Pb-Zn have been described in this sequence, highlighting the Sabiá prospect (reserves of 10 million tons of massive sulfide, Ribeiro et al. 1993), presenting pyrite- and pyrrhotite-rich levels in calc-silicate rocks with tremolitization associated with the Sobradinho unit (Angelim, 1997).

1.2. Rock samples

The samples used in this work were collected from two complementary drill cores next to the Sabiá prospect, that intercept all the stratigraphy of the Baixo Vale do Rio Salitre unit. On average, the samples have 15 cm of length, and one sample was taken from about every 2 meters.

The described rocks are mostly fine-grained and show similar color and texture. These features make manual logging, an essential task in mineral exploration, complicated and costly. Despite similarities between the drill cores' lithotypes, we described at least four rocks types: calc-silicate rocks, carbonaceous phyllites with and without mineralization, metapelites, and metamafic rocks (Figure 2).
The mineralized portions, hosted by carbonaceous phyllites, are characterized by centimetric layers of discordant to concordant massive pyrite (Figure 2d), associated with locally disseminated sulfides (pyrite and pyrrhotite) and sulfide remobilization in calcite veinlets.

Figure 2: Representative rock samples from different lithotypes used in this study: a) calc-silicate rock, b) metapelite, c) carbonaceous phyllite, d) mineralized carbonaceous phyllite, and e) metamafic rock.

1.3. X-Ray Fluorescence

The instrument used in this work was a Thermo Scientific Niton XL3t Goldd+ XRF analyzer, with 2W, 50kV Au anode tube, and a geometrically optimized large area drift detector. The
instrument offers three methods of analysis, and the method chosen in this work is named "TestAll Geo," indicated when the concentration of interest elements is unknown. "TestAll Geo" is a hybrid mode able to detect several major, minor, and trace elements: Ag, Al, As, Au, Ba, Bi, Ca, Cd, Ce, Cl, Co, Cr, Cs, Cu, Fe, Hf, Hg, La, Mg, Mn, Mo, Nb, Nd, Ni, P, Pb, Pd, Pr, Rb, Re, S, Sb, Sc, Se, Si, Sn, Sr, Ta, Te, Th, Ti, U, V, W, Y, Zn, Zr. The instrument was coupled on a stationary test stand during the measurements, where the samples were placed. Each measurement took 120s, with 60s of duration for each beam.

The measurements were performed on quartered and sawed drill core samples, using the "point and shoot" or *in situ* assay mode. The QA/QC adopted procedures followed the suggestions of Fisher et al. (2014) and Piercey (2014) and, as the certified material used in this work, are described in Appendix A.

Aiming to check the representativeness of information, the second measurement of each sample was collected on another spot. Despite some outlier values for all considered elements, the main distribution is maintained in both first and second analyzed spots (Figure 3). The statistical verification of distribution type and equivalence between data from the first and second spot distributions are presented in Appendix B.
2. Data Preparation

2.1. Data management

The bulletins of the X-Ray Fluorescence data were gathered in a single spreadsheet. All procedures of data preparation were performed in the R environment. The data management step was performed using the concepts of "Tidydata" handling, with the "dplyr" package (Wickham, 2014). The graphical analysis and all statistical diagrams shown in this work were done using the package "ggplot2" (Wickham, 2016).

From an initial number of 47 elements, we selected 20 for exploratory data analysis following the proportion of uncensored data. Only variables with at least 95% of valid results were taken. Then, for the multivariate analysis, all missing values were replaced by half of the lower limit of detection, as suggested by Farnham et al. (2002) and Kwak and Kim (2017). No outliers
were removed from the data, as outliers can represent some interesting samples in mineralization.

Each variable was scaled to values between 0 to 1 regarding its range using a Min-Max Feature scaling. This approach matches the ranges into given values but preserves the original data distribution. The normalization is a mandatory step of data preparation in multivariate analysis because of the variance's regularization, an essential parameter to several statistical procedures and verifications (Grunsky, 2010). The Min-Max Feature Scaling is given as:

\[ y(i) = \frac{x(i) - \min(x)}{\max(x) - \min(x)} \]

where \( x(i) = (x_1, ..., x_n) \), and \( y(i) \) is the \( i^{th} \) normalized data.

2.2. Exponentially Weighted Moving Average Filtering

The Exponentially Weighted Moving Average (EWMA) is a filtering process similar to a low-pass convolution, extensively applied in financial market analysis for anomaly detection, forecasting risk analysis, and pattern prediction in general (Lucas and Zhang, 2016; Mitra et al., 2019; Nakano et al., 2017; Ross et al., 2012).

The nature of geological data in drill core samples allows the interpretation that neighbors samples may have a particular correlation. Therefore, the EWMA filtering process is suitable for reducing the high-frequency noise, which can either result from some specificity of a sample (nugget effect) or is a product of accuracy issues of the pXRF analysis.

EWMA filtering is described as follows:

\[ EWMA_{x_i} = \frac{x_i + (1 - \alpha)x_{i-1} + \cdots + (1 - \alpha)^nx_{i-n}}{1 + (1 - \alpha) + \cdots + (1 - \alpha)^n} \]

Where \( EWMA_{x_i} \) is the filtered value for the \( x_i \) sample, \( n \) is the number of neighbors samples considered in the filtering process, and \( \alpha \) is the weighting coefficient that decays exponentially from the centered sample in the function of the distance and is defined as:
\[ \alpha = \frac{2}{n + 1} \]

The EWMA filtering has some advantages, such as a fast performance with low computational cost, strong smoothing properties, with the weights of the neighbor's samples automatically given in function of the number of neighbors. EWMA requires the definition of some hyperparameters, adjusted by the interpreter. The more neighbors' values are considered during the filtering process, the smoother the resulting curve. In most cases, when many neighbors are considered, it also adds an offset or lag incrementation to the curve's shape. Furthermore, it may input loss of information as the first \( n - 1 \) samples are lost in the filtering process.

For data used in this work, the optimum value of neighbors was defined as five, as there was no visible offset added to data, and the algorithm still performed a considerable smoothing in the pattern of the analyzed curve with minimal data loss (Figure 4). For comparison purposes, filtered and raw data sets were treated separated in most of the further analyses.

![Figure 4: concentration of Zr (ppm) of ranked samples for raw data (black line), filtered data for five (red dashed line), fifteen (green dashed line), and twenty-five (blue dashed line) neighbors.](image)

3. **Data Analysis**

   3.1. **Correlation**
The Spearman correlation index ($\rho$) is a rank-based non-parametric correlation of $n$ variables taken by pairs, that measures the capacity of the $i$ rank-ordered values of one variable to predict the rank-ordered values of another. The value of $\rho$ was calculated by the selected elements in the raw and filtered dataset (Figure 5). In the two groups of data, it is possible to identify groups of elements that correlate with each other, both with positive and negative correlations. Ba and Zr correlate with each other and show moderate to strong negative correlation with almost all others. The others groups that show a good correlation is Nd, Pr, Sb, Sn, Cd, Cs, and Te with a positive correlation varying from 0.5 to 0.9 and Cu, Ni, Fe, Cr, V, and Ti with a weak to strong correlation varying from 0.3 to 0.8. By comparing both correlation matrices, it is possible to notice that the elements' relationships tend to be stronger in the filtered data than in the raw data.
Figure 5: Spearman ranked correlogram of the selected elements for filtered (upper triangular matrix) and raw data (lower triangular matrix).

All these groups and associations between elements represent the compositional variation of the main constituent minerals of the analyzed lithotypes. In a traditional assessment, the interpreter tries to link these associations to previously known lithotypes. However, due to the high number of analyzed variables, interpretation can become challenging even for the most experienced professionals.

### 3.2. Dimensionality Reduction
The most common dimensionality reduction method is the Principal Component Analysis (PCA; Grunsky and Arne, 2020). PCA relies on a combination of linear transformations called "basis change" that aims to maximize the data variance on several orthogonal axes, ordered by the first to the last, based on the proportion of explained variance of the dataset. Thus, the first components are often interesting for multivariate analysis, since they typically account for a large proportion of the total variation, as the last components are usually discarded since they may reflect noise rather than the systematic pattern (Forkman et al., 2019). PCA is particularly util before running a cluster analysis because as many clustering methods rely on the "distance concept" (Frey and Dueck, 2007), the PCA's space optimization helps in these processes.

For the dataset used in this work, PCA ran on filtered indicates that the filtering process effectively assists on the noise reduction, as the first components explain a considerable amount of data variance, compared to PCA ran on raw data (Figure 6). Thus, the five first components of the filtered data explain 86% of data variance, while the same number of components explains 76% of the variance on raw data, making the other 15 components almost negligible in both cases. Taking advantage of dimensionality reduction, we used the PCA transformation in raw and filtered data as input of the Unsupervised Learning models.
Figure 6: PCA of the selected variables: a) Scree plot of the first ten dimensions and the respective percentage of explained variance for raw (grey curve) and filtered data (red curve); b) Individuals and Eigenvectors plot of First and Second Dimensions colored by the $\cos^2$ value for raw and c) for filtered data.

4. Unsupervised Learning

4.1. K-means

K-means is a common unsupervised approach that aims to minimize the dissimilarity between the $i$th sample and its respective centroid (Landau and Chis Ster, 2010; Lloyd, 1982). Many algorithms work based on Euclidean Distance, but most of them also permit to change this parameter to an alternative spatial representation.
The algorithm functions by determining which cluster's centroid is closest to each point. Each point is then assigned to its nearest cluster. Each cluster's centroid is recalculated as the arithmetic mean or geometric center of all its assigned points' locations. This process is repeated until no points can be assigned to a closer cluster, or the stopping condition has been reached (Mouton et al., 2020). The algorithm's standard input is the data, the number of iterations, and the number of clusters that cannot be defined automatically for the algorithm.

The Silhouette Width $s(i)$ is defined based on the relation between the $i^{th}$ sample and the distances from the center of the respective cluster and the center of the closest neighbor cluster not containing this sample (Lengyel and Botta-Dukát, 2019), and can be expressed as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where dissimilarity $b(i)$ between the $i^{th}$ sample and the nearest cluster not containing $i$, and $a(i)$ is the same metric between $i$ and the center of the given cluster. Therefore, the Average Silhouette Width is the sum of the respective values of Silhouette Width divided by the number of samples (Lengyel and Botta-Dukát, 2019)

$$S(x_1, ..., x_n) = \frac{\sum_{i=1}^{n} s(i)}{n}$$

Hence, for a specific $k$ number of clusters, the one better suitable for the data is the value of $k$ that maximizes $S$.

For the data used in this work, the maximum value for the Average Silhouette Width was obtained with $k = 2$, value considered for further models of K-means Clustering (Figure 7)
Figure 7: Estimates of the optimum number of clusters by a) the Elbow method and b) the Average Silhouette Width

4.2. Model-Based Clustering

Unlike K-means, a Model-Based Clustering (MBC) does not require any hyperparameter imputation or prior information about the number of groups or the distributions. In MBC, it is assumed 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, 1998; Lai et al., 2018; Landau and Chis Ster, 2010).

The data partition is determined by the Expectation-Maximization (EM) for maximum likelihood, with initial values from agglomerative Hierarchical Clustering (Fraley and Raftery, 1998). Then, 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 \textit{a priori} (Fraley & Raftery, 1998). 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 considered to be the best suitable model (Boehmke and Greenwell, 2019)

\[
BIC = \ln(n)k - 2\ln(\hat{L})
\]

Where \( n \) is the sample size, \( k \) is the number of parameters estimated by the model and \( \hat{L} \) is the maximized value of the likelihood function (Boehmke and Greenwell, 2019), as follows:

\[
\hat{L} = p(x | \hat{\theta}, M)
\]

Where \( M \) is a probabilistic model, \( \hat{\theta} \) is the parameters that maximize the likelihood function and \( x \) the observational data.

The EM algorithm shows that for the filtered data, the suitable number of clusters is four, given by the first local maximum BIC value for the EEE model. This model adjusts a given number of elliptical clusters of equal volume and direction, and a probabilistic model is calculated based on point density for each cluster (Figure 8).
5. Results

Four clustering models were generated in raw data and filtered data, varying according to the clustering method, K-means, or Model-based Clustering (Figure 9). As the K-means Clustering depends on the predefined $k$ parameter given by the maximum Silhouette Average Width (Figure 9a and b), the generated models indicate only two groups by this method (Figure 9a and 8b). Generally, these two groups split those samples with an
association of Ni-Cu-Cr (negative values of PC1) from those with an association of Ba-Zr (positive values of PC1).

The MBC show some differences in the number of groups between raw and filtered data. Only three clusters were defined on raw data, with a superposition of groups on the space generated by the two first dimensions (Figure 9c). While for the filtered data, four distinct groups with no superposition on PC1-PC2 space were defined (Figure 9d).

Moreover, on the pseudo-log generation (in-depth constrained clustering analysis), the effect of filtering becomes more evident as the high-frequency noise tends to decrease in both K-means and MBC pseudo-logs (Figure 10). For all pseudo-logs generated, the MBC ran on filtered data is closer to the manually described log. Furthermore, there are still notable differences, such as an apparent greater detail in the APA2001 drill core (separating a class for the more sulfide range). Besides, no clustering model was able to distinguish between the rocks described as calc-silicate and metamafic.
Figure 9: Bivariate plot of the first and second dimensions of PCA with each sample's position and the respective cluster classification. K-means for Clustering for a) raw and b) filtered data; MBC for c) raw and d) filtered data.

Figure 10: Pseudo-log of drill cores APA2001 and APA3001 generated by K-means and MBC for raw and filtered data and comparing element variation and manual log description.

6. Conclusions

We performed two types of Clustering Analysis in portable X-Ray Fluorescence data of drill core samples from a VHMS occurrence hosted in an Archean/Paleoproterozoic greenstone belt: K-means and Model-Based Clustering. Another comparison we tested was models developed over raw and Exponentially Weighted Mean Average filtered data.
The application of a filter to noisy data, such as that found in some pXRF assays, helps separate groups of lithotypes. As the clustering models rely on geochemical contrast, the filtering process allied to Principal Component Analysis helped increase the signal to noise compensation and maximize the contrast between the clusters. In general, both models developed using the filtered data showed clustering performance closer to that described by professionals.

Both the Silhouette and elbow methods suggest only two clusters for the K-means, which vaguely resembles separating the metapelites and calc-silicate rock, but this method could not detect any other lithotypes.

On the other hand, the Model-Based Clustering analysis suggested four clusters, resembling more with the manual log description. The model could assess this resemblance even then some key-elements were not available for the analysis, like C, Mg, Ca, and K (some of them not analyzed by the pXRF or not pass in the several data tests ran before the clustering model).

In some cases, the pseudo-log models consistently proposed changes on the rock type not detected in the manual log. The method was able to detect small variations in composition amid large rock packages that went unnoticed in the manual description.

However, one lithotype described as "metamafic rock" at the bottom of the APA3001 drill core could not be detected by any models. A possible explanation is the lack of contrast between the analyzed elements due to the similar mineralogy confirmed by thin-section petrography.

The samples from the core APA2001 have a higher level of difficulty fitting a model because of the thin layers repeating rhythmically, as seen in the manual log description. The core APA3001 is more homogeneous than the former, and even in raw data, the models had a certain degree of convergence.

One issue to be considered is the spatial resolution of the measurements. This matter allied to the geochemical contrast can contribute to the pseudo-log's resemblance to the manual
description, and for small resolution as the one taken in this work, only general discrimination is expected.

After all, the results suggest that the MBC method had a better performance than the K-means, mainly for the filtered data. The use of the methods described in this work, especially the Model-Based Clustering combined with EWMA and PCA, can be applied as an important tool in the mineral exploration industry. It highlights that its application can be of aid to solve problems during the validation of geological models based on drill hole data.

**CRediT author statement**

GFS Conceptualization, Investigation, Data Curation, Methodology, Visualization, Writing - Original Draft, Formal analysis, Writing - Review & Editing. JHL Supervision, Data Curation, Formal analysis, Validation, Writing - Review & Editing. ADRS Investigation, Writing - Original Draft, Validation, Writing - Review & Editing. CGL: Project Administration Investigation, Writing - Original Draft, Validation, Writing - Review & Editing. ELK: Investigation, Writing - Review & Editing. KU: Resources, Data Curation. Writing - Review & Editing.

**Acknowledgments**

The authors would like to thank the Geological Survey of Brazil - CPRM, for providing the necessary resources for this research, with mention to the geologists Valter Rodrigues Sobrinho for his assistance with field logistic and sampling and Francisco Ferreira de Campos for help the pXRF managing. Special thanks to the Companhia Baiana de Pesquisa Mineral – CBPM, for allowing the sample collection for this study. GFS would like to thank Marcos Vinicius Ferreira and Iago Sousa Lima Costa for relevant commentaries about the code and manuscript. The data processing was done using several open-source packages from the Tidyverse Collection in the R environment (Wickham, 2014).

**Funding**
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Data Availability**

Datasets related to this article can be found at [https://github.com/gfersilva/RioSalitre](https://github.com/gfersilva/RioSalitre) hosted at Github repository (Silva, 2020). Any commentaries about the code or further collaborations are welcome.

**Appendix A. Quality Control**

The analytical workflow followed the recommendations of Fisher et al. (2014). At the beginning of every batch analysis, the instrument was configured to read continually for at least 40 minutes to correct the instrumental drift caused by variations on the cathode temperature (Ida, 2004; Thermo-Scientific, 2013).

The reference material RM 180-646 was read between every ten samples (or 20 spot measurement). For elements Zr, Sr, Cu, Ni, Fe, and Ti, was detected a translational bias higher than 2% on the measurements of the RM after the 22nd reading (Figure A.1). We applied a correction factor for the adjustment of the values. The factor is calculated by the relation of the average of the first 22 measures. Then this numeral is divided by the average for the offset values (Table A.1.)

Table A.1.: Average concentration for elements with a translational bias higher than 2% and correction factor

| Element | Zr      | Sr      | Cu      | Ni      | Fe      | Ti      |
|---------|---------|---------|---------|---------|---------|---------|
| \( \bar{X} \) (1-21) | 377.15  | 103.22  | 284.46  | 100.56  | 40507.43| 4141.71 |
| \( \bar{X} \) (22-36) | 343.07  | 101.02  | 277.74  | 90.73   | 39482.14| 3882.05 |
| Correction | 1.099   | 1.022   | 1.024   | 1.108   | 1.026   | 1.067   |
Figure A.1.: Shewhart control chart (Piercey, 2014) for measurements on RM 180-646. Identification of the translational bias for a) Fe and b) Zr, and values corrected c) and d) for the same elements.

Appendix B. Significance Tests

B.1 Verification of Data Distribution

An essential step in exploratory data analysis is the determination of the distribution type. If the data is parametric, the mean and standard deviations are reasonable estimates for the center, and the spread of data and several inference tools can be used to analyze and infer population parameters. Otherwise, data must be treated differently with methods that do not rely on these geometric parameters.

There are nearly 40 tests available for normality verification, but several authors (Razali and Wah, 2011; Saculinggan and Balase, 2013; Yap and Sim, 2011) show that the Shapiro-Wilk test is the more powerful test for most types of distributions and sample size. The Shapiro-Wilk
test (Shapiro and Wilk, 1965) was initially defined for small samples \( n < 50 \), and then it was improved by Royston (1982) that expanded the test for a greater range of values \( 3 \leq n \leq 5000 \).

In this work, this test was performed for each selected element on the database, considering only the first pXRF spot data. For a significance level of 5% and analyzing the P-Value parameter, all the selected elements presented \( P \)-Value \(< 0.01 \) (Table B.1), which means that the respective distribution for each element is not the normal distribution, what can be confirmed in the Q-Q Plot diagram (Figure B.1).

Table B.1: Statistical summary for selected elements with more than 95% measurements above the lower limit of detection and respective P-Value for the Shapiro-Wilk and Kruskal Wallis Statistical Significance Test

| N | Element | Values (ppm) | Statistical Significance: P-Value (\( \alpha = 0.05 \)) |
|---|---------|--------------|------------------------------------------------------|
|   |         | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Shapiro-Wilk | Kruskal-Wallis |
| 1 | Ba      | 448.32 | 1069.58 | 1236.66 | 1333.76 | 1592.95 | 2497.94 | \(< 0.01 \) | 0.38 |
| 2 | Cd      | 10.87  | 60.20  | 71.61  | 73.15  | 84.45  | 158.46 | \(< 0.01 \) | 0.57 |
| 3 | Cl      | 1120.56 | 8656.18 | 9432.14 | 9266.63 | 10027.47 | 21456.72 | \(< 0.01 \) | 0.37 |
| 4 | Cr      | 9.17   | 14.06  | 15.54  | 21.66  | 27.82  | 122.06 | \(< 0.01 \) | 0.84 |
| 5 | Cs      | 144.26 | 483.02 | 537.49 | 534.80 | 589.15 | 885.52 | \(< 0.01 \) | 0.05 |
| 6 | Cu      | 249.05 | 644.71 | 731.96 | 731.35 | 812.98 | 1579.56 | \(< 0.01 \) | 0.70 |
| 7 | Fe      | 1905.72 | 11173.17 | 16594.67 | 18730.61 | 23714.21 | 79021.55 | \(< 0.01 \) | 0.33 |
| 8 | Nd      | 335.07 | 634.00 | 785.79 | 783.22 | 918.23 | 1390.39 | \(< 0.01 \) | 0.79 |
| 9 | Ni      | 98.66  | 1398.96 | 1624.67 | 1614.64 | 1865.83 | 3699.80 | \(< 0.01 \) | 0.71 |
| 10 | P       | 1042.74 | 2064.94 | 2316.29 | 2355.65 | 2646.65 | 6825.10 | \(< 0.01 \) | 0.98 |
| 11 | Pr      | 218.89 | 408.59 | 485.02 | 494.92 | 576.88 | 966.06 | \(< 0.01 \) | 0.99 |
| 12 | S       | 1712.74 | 5801.68 | 6326.13 | 6805.11 | 6969.70 | 38275.17 | \(< 0.01 \) | 0.42 |
| 13 | Sb      | 37.93  | 180.11 | 209.29 | 210.09 | 236.84 | 383.74 | \(< 0.01 \) | 0.89 |
| 14 | Sr      | 12925.20 | 20730.95 | 24491.43 | 27574.16 | 27999.96 | 310304.66 | \(< 0.01 \) | 0.31 |
| 15 | Sn      | 32.08  | 109.56 | 127.64 | 128.42 | 145.10 | 222.73 | \(< 0.01 \) | 0.44 |
| 16 | Te      | 4.21   | 18.43  | 29.12  | 38.90  | 44.56  | 256.44 | \(< 0.01 \) | 0.79 |
| 17 | Ti      | 5.63   | 43.76  | 63.92  | 81.72  | 91.53  | 4211.33 | \(< 0.01 \) | 0.71 |
| 18 | V       | 2.18   | 4.24   | 5.25   | 5.71   | 6.37   | 89.51 | \(< 0.01 \) | 0.76 |
| 19 | Zr      | 10.58  | 53.10  | 73.90  | 97.42  | 122.78 | 427.97 | \(< 0.01 \) | 0.97 |

Due to this work's nature, no transformation on data was performed, aiming to force a normal distribution, and we choose to work the data with non-parametric approaches.
Figure B.1: Q-Q plot for selected elements classified by the distribution of measurements and duplicates with the respective P-Value of the Shapiro-Wilk test of normality.

B.2. Representativity of measurements

The window size of the portable X-Ray Fluorescence can be a struggle point for the analysis, as just a small bit of the rock sample is measured by turn in "point and shoot" mode (Lemièrè, 2018). This fact could lead to taking values that are not representative of the whole sample, even that the rock has small granulation.

Regarding this, we ran a double analysis for each sample in different spots to check its statistical representativity. We tested the distributions' statistical equivalence on the two-spot approach by each selected element with the assistance of the Kruskal-Wallis test (KW). KW assess if the distribution of the measurements and duplicates are statistically equivalent (Figure B.2). If it is acceptable that they have an equivalent distribution, we can accept the hypothesis that the data measurements represent that portion of the samples.
The KW test (Kruskal and Wallis, 1952) is an extension of the Wilcoxon-Mann-Whitney for non-parametrical data and verifies if the same distribution for ranked values originates two or more independent samples. The KW test is performed for each of the selected elements to check if the first pXRF spot data distribution can be equivalent to the second pXRF spot data distribution. The KW value test is defined as:

\[
KW = (N - 1) \frac{\sum_{i=1}^{g} n_i (\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (r_{ij} - \bar{r})^2}
\]

As \( N \) is the total number of analysis for the samples, \( n_i \) is the number of observations of the \( i^{th} \) group, \( r_{ij} \) is the rank of observation \( j \) from group \( i \), \( \bar{r} \) is the average rank of \( r_{ij} \) and \( \bar{r}_i \) is the average rank of the \( i^{th} \) group.

The verification of P-Value assesses the result of the KW for a given \( \alpha \) defined as 5% (Table B.1). Therefore, if P-Value is higher than 0.05 for a given element, it is considered that the distribution of the first pXRF spot can be equivalent to the distribution of the second pXRF spot.

All the 20 selected elements have respective P-Value higher than 5%, and thus are suitable for the modeling. The density distribution for both groups for each element is shown in Figure B.2.
Figure B.2.: Density distribution plot for selected elements classified by the distribution of measurements and duplicates with the respective P-Value of the KW test of equivalence of the two distributions.

REFERENCES

Alkmim, F.F., Martins-Neto, M.A., 2012. Proterozoic first-order sedimentary sequences of the São Francisco craton, eastern Brazil. Mar. Pet. Geol. 33, 127–139. https://doi.org/10.1016/j.marpetgeo.2011.08.011

Alkmin, F.F., Brito Neves, B.B., Alves, J.A.C., 1993. Arcabouço tectônico do Cráton do São Francisco: Uma revisão, in: Dominguez, J.M.L., Misi, A. (Eds.), O Cráton Do São Francisco. SBG - Sociedade Brasileira de Geociências, Salvador, BA, pp. 45–62.

Angelim, L.A.A., 1997. Petrolina, folha SC.24-V-C: estados da Bahia, Pernambuco e Piauí. Escala 1:250.000. Programa Levantamentos Geológicos Básicos do Brasil - PLGB. CPRM - Companhia de Pesquisa de Recursos Minerais, Brasília, DF.

Barbosa, J.S., Barbosa, R.G., 2017. The Paleoproterozoic eastern Bahia orogenic domain., in: The São Francisco Craton and Its Margins, Eastern Brazil. Springer-Verlag, pp. 57–69.
Barbosa, J.S., Sabaté, P., 2004. Archean and Paleoproterozoic crust of the São Francisco Craton, Bahia, Brazil: geodynamic features. Precambrian Res. 133, 1–27. https://doi.org/10.1016/j.precamres.2004.03.001

Barbosa, J.S.F., 2012. Geologia da Bahia: Pesquisa e atualização, 1st ed, Séries e publicações especiais - 13. Companhia Baiana de Pesquisa Mineral, Salvador, BA.

Bérubé, C.L., Olivo, G.R., Chouteau, M., Perrouty, S., Shamsipour, P., Enkin, R.J., Morris, W.A., Feltrin, L., Thiémonge, R., 2018. Predicting rock type and detecting hydrothermal alteration using machine learning and petrophysical properties of the Canadian Malartic ore and host rocks, Pontiac Subprovince, Québec, Canada. Ore Geol. Rev. 96, 130–145. https://doi.org/10.1016/j.oregeorev.2018.04.011

Boehmke, B., Greenwell, B., 2019. Hands-On Machine Learning with R, 1st ed. Chapman and Hall/CRC. https://doi.org/10.1201/9780367816377

Cordani, U., Iyer, S., Taylor, P., Kawashita, K., Sato, K., McReath, I., 1992. Pb/Pb, Rb/Sr, and K/Ar systematics of the Lagoa Real uranium province (south-central Bahia, Brazil) and the Espinhaço cycle (ca. 1.5-1.0 Ga). J. South Am. Earth Sci. 5, 33–46. https://doi.org/10.1016/0895-9811(92)90058-7

Costa, I., Tavares, F., Oliveira, J., 2019. Predictive lithological mapping through machine learning methods: a case study in the Cinzento Lineament, Carajás Province, Brazil. J. Geol. Surv. Brazil 2, 26–36. https://doi.org/10.29396/jgsb.2019.v2.n1.3

de Almeida, F.F.M., Hasui, Y., de Brito Neves, B.B., Fuck, R.A., 1981. Brazilian structural provinces: An introduction. Earth-Science Rev. 17, 1–29. https://doi.org/10.1016/0012-8252(81)90003-9

Farnham, I.M., Singh, A.K., Stetzenbach, K.J., Johannesson, K.H., 2002. Treatment of nondetects in multivariate analysis of groundwater geochemistry data. Chemom. Intell. Lab. Syst. 60, 265–281. https://doi.org/10.1016/S0169-7439(01)00201-5

Fisher, L., Gazley, M.F., Baensch, A., Barnes, S.J., Cleverley, J., Duclaux, G., 2014. Resolution of geochemical and lithostratigraphic complexity: A workflow for application of portable X-ray
fluorescence to mineral exploration. Geochemistry Explor. Environ. Anal. 14, 149–159. https://doi.org/10.1144/geochem2012-158

Forkman, J., Josse, J., Piepho, H., 2019. Hypothesis Tests for Principal Component Analysis When Variables are Standardized. J. Agric. Biol. Environ. Stat. 24, 289–308. https://doi.org/10.1007/s13253-019-00355-5

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

Fraley, C., Raftery, A.E., 1998. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis I. Seattle, WA.

Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. Science (80-. ). 315, 972–976. https://doi.org/10.1126/science.1136800

Garcia, P.M.P., 2017. A Província Cuprífera Do Nordeste Meridional: Evolução Dos Processos E Modelos. UFBA - Universidade Federal da Bahia.

Grunsky, E.C., 2010. The interpretation of geochemical survey data. Geochemistry Explor. Environ. Anal. 10, 27–74. https://doi.org/10.1144/1467-7873/09-210

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 Activities 2019: Minerals, Geoscience BC, Report 2020-01. Geoscience BC, Vancouver, BC, pp. 55–76.

Heilbron, M., Cordani, U.G., Alkmim, F.F., 2017. São Francisco Craton, Eastern Brazil: Tectonic Genealogy of a Miniature Continent. Springer 326. https://doi.org/10.1007/978-3-319-01715-0

Hill, E.J., Pearce, M.A., Stromberg, J.M., 2020. Improving Automated Geological Logging of Drill Holes by Incorporating Multiscale Spatial Methods. Math. Geosci. https://doi.org/10.1007/s11004-020-09859-0

Hill, E.J., Uvarova, Y., 2018. Identifying the nature of lithogeochemical boundaries in drill holes. J. Geochemical Explor. 184, 167–178. https://doi.org/10.1016/j.gexplo.2017.10.023
Ida, H., 2004. X-ray fluorescence analysis with portable instruments. Kyoto University.

Kruskal, W.H., Wallis, W.A., 1952. Use of Ranks in One-Criterion Variance Analysis. J. Am. Stat. Assoc. 47, 583–621. https://doi.org/10.1080/01621459.1952.10483441

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

Kwak, S.K., Kim, J.H., 2017. Statistical data preparation: Management of missing values and outliers. Korean J. Anesthesiol. https://doi.org/10.4097/kjae.2017.70.4.407

Lai, K., Twine, N., O’Brien, A., Guo, Y., Bauer, D., 2018. Artificial intelligence and machine learning in bioinformatics. Encycl. Bioinforma. Comput. Biol. ABC Bioinforma. 1–3, 272–286. https://doi.org/https://doi.org/10.1016/B978-0-12-809633-8.20325-7

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

Le Vaillant, M., Barnes, S.J., Fisher, L., Fiorentini, M.L., Caruso, S., 2014. Use and calibration of portable X-Ray fluorescence analysers: Application to lithogeochemical exploration for komatiite-hosted nickel sulphide deposits. Geochemistry Explor. Environ. Anal. 14, 199–209. https://doi.org/10.1144/geochem2012-166

Lemière, B., 2018. A review of pXRF (field portable X-ray fluorescence) applications for applied geochemistry. J. Geochemical Explor. 188, 350–363. https://doi.org/10.1016/j.gexplo.2018.02.006

Lengyel, A., Botta-Dukát, Z., 2019. Silhouette width using generalized mean—A flexible method for assessing clustering efficiency. Ecol. Evol. 9, 13231–13243. https://doi.org/10.1002/ece3.5774

Lloyd, S.P., 1982. Least Squares Quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137. https://doi.org/10.1109/TIT.1982.1056489

Lucas, A., Zhang, X., 2016. Score-driven exponentially weighted moving averages and Value-at-Risk forecasting. Int. J. Forecast. 32, 293–302. https://doi.org/10.1016/j.ijforecast.2015.09.003

Mitra, A., Lee, K.B., Chakraborti, S., 2019. An adaptive exponentially weighted moving average-type
control chart to monitor the process mean. Eur. J. Oper. Res. 279, 902–911. https://doi.org/10.1016/j.ejor.2019.07.002

Mouton, J.P., Ferreira, M., Helberg, A.S.J., 2020. A comparison of clustering algorithms for automatic modulation classification. Expert Syst. Appl. 113317. https://doi.org/10.1016/j.eswa.2020.113317

Nakano, M., Takahashi, A., Takahashi, S., 2017. Generalized exponential moving average (EMA) model with particle filtering and anomaly detection. Expert Syst. Appl. 73, 187–200. https://doi.org/10.1016/j.eswa.2016.12.034

Oliveira, E.P., Windley, B.F., Araújo, M.N.C., 2010. The Neoproterozoic Sergipano orogenic belt, NE Brazil: A complete plate tectonic cycle in western Gondwana. Precambrian Res. 181, 64–84. https://doi.org/10.1016/j.precamres.2010.05.014

Piercey, S.J., 2014. Modern analytical facilities 2. A review of quality assurance and quality control (qa/qc) procedures for lithogeochemical data. Geosci. Canada 41, 75–88. https://doi.org/10.12789/geocanj.2014.41.035

Razali, N.M., Wah, Y.B., 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J. Stat. Model. Anal. 2, 21–33.

Ribeiro, A.F., Garrido, I.A., Brito, R.S.C., Nonato, I.F., 1993. Geologia e potencialidade para mineralizações de ouro e sulfetos da Faixa Rio Salitre, Juazeiro – Bahia. Série Arquivos Abertos. CBPM - Companhia Baiana de Pesquisa Mineral, Salvador, BA.

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 Geol. Rev. 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Ross, G.J., Adams, N.M., Tasoulis, D.K., Hand, D.J., 2012. Exponentially weighted moving average charts for detecting concept drift. Pattern Recognit. Lett. 33, 191–198. https://doi.org/10.1016/j.patrec.2011.08.019

Royston, J.P., 1982. An Extension of Shapiro and Wilk’s W Test for Normality to Large Samples. Appl. Stat. 31, 115–124. https://doi.org/10.2307/2347973
Saculinggan, M., Balase, E.A., 2013. Empirical power comparison of goodness of fit tests for normality in the presence of outliers. J. Phys. Conf. Ser. 435. https://doi.org/10.1088/1742-6596/435/1/012041

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611. https://doi.org/10.1093/biomet/52.3-4.591

[dataset] Silva, G.F. da, 2020. Model-Based Clustering for unsupervised drill core pseudo-log generation - Repository [WWW Document]. Github. URL https://github.com/gferrsilva/RioSalitre (accessed 9.16.20).

Teixeira, W., Carneiro, M.A., Noce, C.M., Machado, N., Sato, K., Taylor, P.N., 1996. Pb, Sr and Nd isotope constraints on the Archaean evolution of gneissic-granitoid complexes in the southern São Francisco Craton, Brazil. Precambrian Res. 78, 151–164. https://doi.org/10.1016/0301-9268(95)00075-5

Teixeira, W., Oliveira, E.P., Marques, L.S., 2017. Nature and evolution of the Archean crust of the São Francisco Craton, in: Heilbron, M., Alkmin, F.F., Cordani, U.G. (Eds.), The São Francisco Craton and Its Margins, Eastern Brazil. Springer-Verlag, pp. 29–56.

Thermo-Scientific, 2013. Mining and exploration: Solutions from early-stage discovery through mineral processing. San Jose, California.

Wedge, D., Hartley, O., McMickan, A., Green, T., Holden, E.J., 2019. Machine learning assisted geological interpretation of drillhole data: Examples from the Pilbara Region, Western Australia. Ore Geol. Rev. 114, 103118. https://doi.org/10.1016/j.oregeorev.2019.103118

Wickham, H., 2016. ggplot2, Use R! Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-24277-4

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

Yap, B.W., Sim, C.H., 2011. Comparisons of various types of normality tests. J. Stat. Comput. Simul. 81, 2141–2155. https://doi.org/10.1080/00949655.2010.520163