Qmin: A machine learning-based application for mineral chemistry data processing and analysis

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

Marcos Vinicius Ferreira
Directory of Geology and Mineral Resources, Geological Survey of Brazil https://orcid.org/0000-0001-5213-0825

Iago Sousa Lima Costa
Directory of Geology and Mineral Resources, Geological Survey of Brazil https://orcid.org/0000-0002-3721-8957

Renato Borges Bernardes
Electron Probe Micro-Analyzer Laboratory, Institute of Geosciences, University of Brasilia https://orcid.org/0000-0002-5065-3830

Carlos Eduardo Miranda Mota
Directory of Geoscience Infrastructure, Geological Survey of Brazil https://orcid.org/0000-0002-6652-0493

Federico Alberto Cuadros Jiménez
Institute of Geosciences, University of Brasilia https://orcid.org/0000-0002-2297-9964

Research Article

Keywords: EPMA data processing, Random Forest classifier, Mineral prediction, Mineral formula calculation

Posted Date: June 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-629516/v1

Abstract

Mineral chemistry analysis is a valuable tool in several phases of mineralogy and mineral prospecting studies. This type of analysis can point out relevant information, such as concentration of the chemical element of interest in the analyzed phase and, thus, the predisposition of an area for a given commodity. Due to this, considerable amount of data has been generated, especially with the use of electron probe micro-analyzers (EPMA), either in research for academic purposes or in a typical prospecting campaign in the mineral industry. We have identified an efficiency gap when manually processing and analyzing mineral chemistry data, and thus, we envisage this research niche could benefit from the versatility brought by machine learning algorithms. In this paper, we present Qmin, an application that assists in increasing the efficiency of mineral chemistry data processing and analysis stages through automated routines. Our code benefits from a hierarchical structure of classifiers and regressors trained by a Random Forest algorithm developed on a filtered training database extracted from the GEOROC (Geochemistry of Rocks of the Oceans and Continents) repository, maintained by the Max Planck Institute for Chemistry. To test the robustness of our application, we applied a blind test with more than 11,000 mineral chemistry analyses compiled for diamond prospecting within the scope of the Diamante Brasil Project of the Geological Survey of Brazil. The blind test yielded a balanced classifier accuracy of ca. 99% for the minerals known by Qmin. Therefore, we highlight the potential of machine learning techniques in assisting the processing and analysis of mineral chemistry data.

Highlights

- Qmin is an open-source tool that helps to automate the processing of EPMA data.
- Qmin runs several classifiers to distinguish 17 groups and 100 different minerals.
- We used Shannon Uncertainty to verify the quality of the prediction.
- Balanced accuracy in a blind test achieved 99.44% for known minerals.
- Qmin is written in Python and R and is available at http://apps.cprm.gov.br/qmin/

1. Introduction

Mineral chemistry analysis constitutes a significant part of studies involving different branches of geosciences (e.g., mineralogy, petrology, economic geology). Nowadays, the amount of chemical data produced is enormous, especially by electron probe micro-analyzers (EPMA). Processing, analyzing, and interpreting high-dimensional data, such as those from EPMA, present a tremendous challenge (e.g., Cracknell et al., 2014; Radford et al., 2018). Many datasets, especially in geosciences, represent complex and non-linear physical systems (see Bergen et al., 2019). Therefore, manipulation and interpretation of high-dimensional geoscientific data through basic graphics and spreadsheets are exhaustive and time-consuming tasks that, in turn, may be prone to unsystematic human biases (e.g., the misclassification of a mineral).

Machine learning algorithms (MLA) have emerged as powerful tools to deal with massive datasets and recurring tasks in recent years. Several works have used MLA to solve different geoscientific problems, such as geological mapping (e.g., Costa et al., 2019; Cracknell et al., 2014; Kuhn et al., 2020, 2018; Radford et al., 2018), data-driven mineral prospectivity mapping (e.g., Brandmeier et al., 2020; Carranza and Laborte, 2016, 2015; Prado et al., 2020; Rodriguez-Galiano et al., 2015; Zhang et al., 2021), anomaly detection, among many others (see Dramsch, 2020, and references therein). Specifically, in mineralogy, MLA have been used for mineral identification and classification from rock thin sections images (e.g.,Borges and Aguiar, 2019; Rubo et al., 2019a) or from drill cores (e.g., Koch et al., 2019), and for the calculation of mineral formulas, e.g., for amphiboles (Li et al., 2020). However, to our knowledge, there is no application that tries to deal, in a holistic way, with mineral classification and formula calculation for some of the most common minerals (i.e., the rock-forming minerals) with the use of EPMA data.

In this scope, we provide Qmin – Mineral Chemistry Virtual Assistant – a machine learning-based algorithm focused on processing mineral chemistry data from EPMA analyses. With our application, we aim to automatize and statistically evaluate mineral classification and mineral formula calculation within a single integrated open-access code and, thus, simplify and speed up many post-analytical steps of studies that depend on mineral chemistry results.

1.1 The application

MLAs can be separated into two critical groups: i) algorithms with a defined target (supervised learning) and ii) algorithms that cluster groups of data with similar features in a high-dimensional domain without a pre-defined target (unsupervised learning). One of the most employed MLA in geoscience prediction problems is the Random Forest (RF – Breiman, 2001). The RF combines several independent decision trees to build classification or regression models through bootstrap aggregation.

In this sense, Qmin is a web application, built on top of Python 3 Flask (Grinberg, 2018) and sci-kit-learn library (Pedregosa et al., 2011), that gathers several nested Random Forest models (Breiman, 2001) trained to recognize new entries of EPMA analysis, to classify the mineral group, and to identify the most probable mineral, according to the comparison with a reference dataset.

One of RF's most significant advantages is that this algorithm has a high performance combined with hyperparameters that are easier to be tuned. Also, several geoscientific works have shown that the RF outperformed the other MLA, such as Support Vector Machines, Artificial Neural Networks, Logistic Regression, among others (e.g., Costa et al., 2019; Kuhn et al., 2018; McKay and Harris, 2016; Rodriguez-Galiano et al., 2015). These characteristics make RF widely and effectively used (e.g., Carranza and Laborte, 2015; Costa et al., 2019; Ford, 2019; Hariharan et al., 2017; Harris et al., 2015)

Qmin can evaluate the quality of the analysis by measuring the statistical entropy (Shannon, 1948) of each new data entry. Exploratory data analysis can be done directly on the application, with biplot and triplot graphs.

Within Qmin, we developed a tool to determine the empirical mineral formula for each analysis. In the current version of Qmin, this is applicable for some mineral groups, such as Pyroxene, Feldspar, Mica, Garnet, Olivine, and Spinel. In this tool, the mineral formula is calculated explicitly by the charge balance method (Deer et al., 2013). We also developed another tool that calculates the mineral formula of Amphiboles by a probabilistic approach, with a multivariate regression based on the Random Forest Algorithm. However, we emphasize that the latter is experimental.

1.2 Data source

The original data used to train the Qmin algorithm are derived from the GEOROC (Geochemistry of Rocks of the Oceans and Continents) repository, supported by the Max Planck Institute for Chemistry (Sarbas, 2021; Schramm et al., 2006). The GEOROC initiative collects and organizes several standardized spreadsheets with geoscientific data in tables and other supplementary materials. Among the data available on the GEOROC repository are geochemistry of rocks, fluid inclusion data, petrographic descriptions, and mineral chemistry data. The GEOROC repository has more than one million mineral chemistry analyses from 17 different mineral groups and almost 400 different fields to gather metadata, reference, and chemical analysis on different representations.

Workflow

Dealing with a collection with the nature of the GEOROC dataset is a challenge. We aimed at selecting the most consistent analyses of each mineral to train the algorithm to classify new data with as minimum bias as possible. To reach this, we have applied a series of actions that yielded a uniform and workable database that can train a more effective model, which will be much less disturbed by features that could impact the results. Figure 1 summarizes the steps to create the mineral chemistry data classification model. It is the evaluation and deployment of the system. The pre-processing gathers data from an external source (i.e., GEOROC) and proceeds to make the data wrangling to clean and adapt the database for the development of the application. Next, we briefly describe these actions.

2.1 Pre-processing

As the GEOROC analyses are compiled from publications with different objectives, not all original data entries were necessary for our application. We first selected 20 elements from the original GEOROC data that can represent all the variations among the different minerals present in the dataset (SiO_2, S, Al_2O_3, ZrO_2, F, CoO, CaO, Na_2O, MgO, ZnO, Cl, K_2O, FeO_t, TiO_2, CuO, MnO, NiO, P_2O_5, Cr_2O_3, and As).

We then performed a logical data cleaning procedure to eliminate non-descriptive chemical data (see Table 1 for a summary). Where needed, variables were leveled (i.e., concentrations in ppm – parts per million - of elements were converted to wt.% - weight percentage). For data populated with low instance numbers, we replaced missing data using multiple regression imputation methods (Martín-Fernández et al., 2003; Schroeder et al., 2008). Finally, we applied several filters to select the adequate analysis based

on the range of values for the sum of the weight percentage (e,g, filtering for range between 99-101% w.p. for anhydrous minerals).

2.1.1 Mineral name reclassification

To achieve mineral nomenclature consistency and, with that, improve our algorithm output, we removed 17 and reclassified 50 mineral name entries of the original GEOROC database (see Table 1). The main criterium adopted for this step was that the mineral name must have been approved by the IMA (International Mineralogical Association). However, when coherent, exceptions to this rule were allowed to exist. The first exception is in some minerals whose solid solution names are established and can be prominently distinguished with an EPMA analysis (e.g., the plagioclase series). The second exception is when some end-member entry of a solid solution is missing in the original GEOROC database (e.g., apatites, eastonite, polylithionite, and trilithionite) or has a high complexity (e.g., the hornblendes). In these cases, we incorporated the respective mineral as a *sensu lato* entry (e.g., apatite, biotite, lepidolite, and hornblende). This approach is a mineralogically acceptable and a simple and straightforward way to mitigate classification inconsistencies when training the algorithm.

Table 1: Summary of original and processed data after the filtering criteria and mineral name reclassification

Mineral group	GEOROC original number of the analysis	GEOROC original number of mineral names	Filtering criterium	Number of remained analysis (% of the original)	Number of resulting reclassified mineral names
Amphibole	38639	33	missingness of SiO2 (wt%)	33571 (87%)	23
Apatite	12696	3	missingness of PO4 (wt%)	7518 (59%)	1
Carbonate	9189	35	missingness of CaO (wt%)	4283 (47%)	18
Clay Mineral	753	10	missingness of SiO2 (wt%)	588 (78%)	5
Feldspar	174107	18	missingness of SiO2 (wt%)	149351 (86%)	8
Feldspathoid	4332	12	missingness of SiO2 (wt%)	3386 (78%)	8
Garnet	42340	9	missingness of SiO2 (wt%)	37505 (89%)	5
Ilmenite	14894	6	missingness of TiO2 (wt%)	14365 (96%)	1
Mica	35035	18	missingness of SiO2 (wt%)	26448 (75%)	6
Olivine	185404	6	missingness of SiO2 (wt%)	170277 (92%)	3
Perovskite	11022	3	missingness of CaO (wt%)	3363 (31%)	1
Pyroxene	194950	25	missingness of SiO2 (wt%)	170741 (88%)	10
Quartz	5304	3	missingness of SiO2 (wt%)	226 (4%)	1
Spinel	64421	14	missingness of Cr2O3	56672 (88%)	5

			(wt%)		
Sulfide	7004	49	missingness of S (wt%)	3795 (54%)	19
Titanite	5469	1	missingness of CaO (wt%)	1496 (27%)	1
Zircon	265824	2	missingness of ZrO2 (wt%)	1986 (1%)	1

Besides these, we also applied some complementary criteria to reclassify the GEOROC mineral nomenclature. When it was possible, we retrieved an IMA approved name from the original GEOROC entry (e.g., breunnerite, an informal variety of magnesite, was renamed to magnesite, the IMA approved name for the mineral phase). When the mineral has polymorphs (e.g., alabandite), we added all the natural polymorphs to our algorithm-related class (e.g., alabandite/browneite/rambergite). Please refer to Table 1 of the Supplementary Material for a complete list of all the alterations applied to the original GEOROC mineral nomenclature.

2.1.2 Outlier removal

After the previous filtering and reclassification step, some non-compliant analyses on the GEOROC dataset remained. We interpreted these samples as residual outliers and, to avoid confusion on the machine learning training stage (Smiti, 2020), we removed these outliers from the prepared database.

The outliers fall into three categories: i) point or global outliers, as bad analysis (i.e., the sum of elements significantly diverts from 100%); ii) contextual outliers, as atypical samples enriched in certain chemical elements; iii) or classification errors, as the original mineral classification from the original GEOROC database.

We applied the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) unsupervised algorithm (Ester et al., 1996) to solve the outliers problem. The DBSCAN seeks, in the vectorial space, for areas of high density (i.e., inliers) and low density (i.e., outliers). The main idea is that for each sample in the inlier cluster, the neighborhood region of a specific user-defined size must contain at least a minimum number of samples. That is, the density in the neighborhood must exceed a user-defined threshold (Ester et al., 1996; Misra et al., 2020). The DBSCAN's hypermeters are , the minimum number of neighbors, and , the maximum distance where must be found. Basically, the algorithm runs this analysis for each point in the database and those who do not meet the conditions are marked as outliers (Figure 2).

The DBSCAN algorithm was run for every one of the 17 mineral groups. The algorithm detected 3,457 outliers from 523,846 samples (i.e. around 0.67%). Despite the algorithm successfully detecting both global and contextual outliers, it was not efficient in removing misclassification outliers. For this case, we

had to apply a visual inspection and manually remove the clear cases within the code lines (e.g., minerals classified as anorthite but with more than 2 wt.% of Na_2O).

2.1.3 Data balancing

Imbalanced data is a common phenomenon in the development of machine learning models, not only in geoscientific problems but in almost all artificial intelligence problems, such as medical diagnosis (e.g., Vijayvargiya et al., 2021). Imbalanced data considerably reduces a model's capacity to perform predictions, especially for the minority class, where the recognition rate decreases considerably (Japkowicz and Stephen, 2002). Therefore, resampling data configures as a mandatory pre-processing step for a successful, high-performance machine learning model.

In this context, and by analyzing the number of instances of each mineral in the filtered data mineral groups, we empirically defined 50 instances as an adequate and sufficient threshold for each mineral class. In this way, if a specific mineral class had more than 50 instances, we randomly undersampled the instances to balance the data for the model training. On the other hand, if a class had less than 50 instances, we oversampled the instances with the synthetic minority oversampling technique (SMOTE, Chawla et al., 2002). This technique creates synthetic samples in a random point between a valid instance and a random neighbor from a determined number of nearest neighbors in the feature space (Figure 3).

2.2 MODEL IMPLEMENTATION

The creation of the model for mineral chemistry classification is divided into two steps (see Figure 1). The first step classifies the minerals into 17 different groups (see Table 1). The second step classifies the mineral specimens within each group. Currently, Qmin can correctly classify 103 different minerals. The classifiers were trained by Random Forest models, a non-parametric technique that grows decision trees to select the best output for the classification task (Breiman, 2001).

2.2.1 Model tunning

To tune the model, we applied a grid search to choose the adequate hyper-parameters of each model. During the random forest training, we searched for the number of trees between 10 and 150, the decision function between Gini and entropy, and the criteria for the maximum features between the square root or the log2 of the number of features.

To assess the best model, we used ten-fold cross-validation. When the best parameters are found in the grid search, we selected them and proceeded to find the parameters for the next one.

2.2.2 Final model validation

We performed two types of quantitative validation in the model implementation: cross-validation and "train-test split" validation.

The cross-validation approach involves taking a random sample of data from a population without no prior data splitting. This technique is usually done to evaluate the quality of the whole dataset without the need to split it into "training" and "testing" data sets. The essence of the cross-validation lies in computing sample statistics on the first set of sample data and then applying it to the second set (Schumacker and Tomek, 2013). This type of validation involves randomly splitting a sample into two halves and then computing it. For the Qmin implementation, we ran the cross-validation at the preliminary and final model evaluation during the pre-processing and model implementation stages (see Figure 1).

To build and evaluate the several classifiers' models, we applied an intermediate "train-test split" validation. This validation is helpful to verify the accuracy of predictions on some randomly taken subsample that the model has not seen previously. For that, we divided the data into two subsamples: one for training, with 70% of the samples, and the other with 30% of the samples, for testing in cross-validation assessments. The final model was implemented based on the training set. The first accuracy was accessed based on the predictions made by the model on the test dataset. Then, the reference values are confronted with the predicted values. Finally, the accuracy and other parameters were calculated (Figure 4).

2.3 Production

2.3.1 Data processing protocol: new data input and output

The new data to be processed and analyzed must be in the form of a flat file (spreadsheet), with the EPMA analyses organized by rows and the chemical elements (features) organized by columns. We designed a web interface to facilitate the input of the data, accepting both CSV and Excel spreadsheet formats.

The program receives the data from EPMA analyses and automatically detects the columns and whether they are expressed in element or oxide weight percentage. If needed, a conversion is done by weight distribution. All these steps simplify the data input by the user, reducing the dataset's manipulation before the input in the virtual assistant.

All data must be numerical (integer, double, or float), and missing values (e.g., non-detected chemical elements) must be previously replaced by zero or other numerical values. If any entry considered in the predicting processing is not presented, the application understands that the referred feature value is equal to zero for all analyses and automatically applies an imputation.

After the data input, the Qmin web application automatically returns to the user a new downloadable data file encoded by a hash-unique value with the following newly added columns: Group classification, Quality control of group prediction, Mineral classification, Quality control of mineral prediction, second most probable mineral classification, Mineral formula (if the calculation is available) and several other columns related to the mineral formula calculation.

2.3.2 Quality control

To measure the quality of the predicted value, we use the Shannon Entropy function (Shannon, 1948), which measures the uncertainty of the prediction in discrete values based on the probability of each guess the model has made. The following equation shows how to calculate the values of Shannon entropy (*H*), where the variable is the probability of each guess the model has made.

See formula 1 in the supplementary files section.

To facilitate for the Qmin final user, we classified the values of the uncertainty (i.e., E = 1 - H) in three categories: High Quality (E>0.7), Medium Quality (0.5 < E <0.7), and Low Quality (E < 0.5).

Blind Test

To test the robustness of Qmin, we did a blind test running the data from the Diamante Brasil Project (Cunha et al., 2017). This project was conducted by the Geological Survey of Brazil and contains more than 22,000 EPMA samples from different minerals. This data was analyzed and manually classified by specialized geologists before our blind test.

Figure 5 summarizes the results using the exploratory graphics that Qmin presents for the user. The application achieved a 99.44% balanced accuracy in the classification across groups and minerals (see Table 2 for a summary).

Some cases of misclassification are, to a certain extent, due to the quality of input data (e.g., the sum of all analyses diverges considerably from 100%). This divergence is particularly prominent for complex hydrated minerals, such as the amphiboles.

Table 2: Statistical summary of resulting Qmin mineral classification of the Diamante Brasil Project data

Group (Mineral)	Instances	Misclassifications	Accuracy
Amphibole	87	13	0.8506
Clinopyroxene	3282	0	1.0000
Othopyrixene	1065	0	1.0000
Spinel	5253	28	0.9947
Mica (Flogopite)	534	4	0.9925
Garnet	9438	52	0.9945
Ilmenite	996	15	0.9849
Kalsilite	10	3	0.7000
Olivine (Monticellite)	40	2	0.9500
Olivine	1204	0	1.0000
Perovskite	213	5	0.9765
Feldspar (Sanidine)	12	0	1.0000

Discussion And Conclusions

We presented a new open-source and free-of-charge application that helps to simplify and speed up EPMA mineral chemistry data processing, analysis, and interpretation by using machine learning techniques. The Qmin application can provide high accuracy predictions among several different mineral groups and specimens, with a mapped uncertainty for each classification. The performances found in this work for the training and test data are above the average reported in several publications in the field (Borges and Aguiar, 2019; Gavish et al., 2018; Koch et al., 2019; Li et al., 2020).

Qmin has a great potential for use in the mineral prospection industry, where efficiency is determining. The tool helps simplify the workflow of a considerable part of the post-analytical stage of mineral prospecting campaigns or academic research that uses large volumes of EPMA data. In the study case presented, whose data covers several minerals used in the prospecting for diamonds on a national scale, conventional processing and analysis that could take days was reduced to a few minutes of work on a computer with standard processing capacity. The application is in constant development, and new and improved features are to be implemented in future iterations of the code, especially regarding new mineral entries and an improved amphibole formula calculation.

Finally, we would like to highlight some points of best practices to achieve good performances during the use of Qmin:

• The mineral classification is based on mineral compositions available in the database fed to the models during the training stage. This setting implies that the models cannot recognize any mineral different from those implemented during the training. Thus, the algorithm will classify any new mineral as one already known to it, provided similarity within the created data multivariate space.

• The quality of the predictions and mineral formula are directly associated with the analysis' quality in terms of the analytical balance. Analyses whose sums of chemical elements are far from the total value (i.e. 100%) tend to render inadequate predictions.

• The database used for training, although robust, has inconsistencies such as those existing before any repository that compiles extensive historical series, for example, variation in the precision (i.e. detection limit) of the analytical methods. Inconsistencies have also been identified from discrepancies in the nomenclature pattern or incongruent allocation of minerals within groups (e.g., allocating tellurides along with sulfides or some silicates along with carbonates). These problems affect the models' ability to make good predictions and, if removed, a reduction of uncertainties can be achieved.

Computer Code Availability

- Code name: Qmin – Mineral Chemistry Virtual Assistant
- Code developers: Guilherme Ferreira da Silva, Marcos Vinícius Ferreira, Iago Sousa Costa Lima, Carlos Eduardo Miranda Mota
- Contact details: Serviço Geológico do Brasil – CPRM/SGB, Setor Bancário Norte, Q 2, Bloco H, Edifício Central Brasil, 2º Andar; Brasília – Distrito Federal, Brasil; e-mail: guilherme.ferreira@cprm.gov.br
- Year first available: 2021
- Hardware minimum requirements: 4GB of RAM.
- Required software: Any OS (Linux, macOS, or Windows), Python IDE and libraries scikit learn, and others specified in the readme file.
- Program size: 50 MB
- Programming languages: the app was developed in Python 3.6. The pre-processing step was developed in R 3.6.5
- Details on how to access the source code: the source files of Qmin are available at the GitHub repository: https://github.com/gferrsilva/QMineral_Modeller. The complete application in its current state is hosted on the Geological Survey of Brazil's website (https://apps.cprm.gov.br/qmin/) and can be accessed free of charge.

CRediT authorship contribution statement

GFS: Conceptualization, Data curation, Methodology, Investigation, Software, Writing – original draft, Writing – Review & Editing. **MVF:** Conceptualization, Data curation, Methodology, Investigation, Software,

Writing – original draft. **ISLC:** Conceptualization, Methodology, Software, Writing – original draft. **RBB:** Conceptualization, Data curation, Writing – original draft, Writing – Review & Editing. **CEMM:** Software, Resources, Writing – Review & Editing. **FACJ:** Data curation, Writing – Review & Editing.

Declarations

ACKNOWLEDGMENTS

The Qmin application is registered in the Brazilian National Institute of Industrial Property (INPI), under the number BR512020001784-3. We are thankful to the Geological Survey of Brazil team (SGB/CPRM) for supporting the initiative and providing the necessary conditions for its development, especially to Evandro Luiz Klein, who contributed to the original idea. We are also grateful to professor Nilson Francisquini Botelho (University of Brasilia) for his significant comments and suggestions on an early version of the application.

SUPPLEMENTARY ITEMS

Table of mineral name reclassification, used in all our algorithm instances.

Table of Qmin processed results of Diamante Brasil Project data.

 The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

References

Bergen, K.J., Johnson, P.A., de Hoop, M. V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science (80-.). 363, eaau0323. https://doi.org/10.1126/science.aau0323

Borges, H.P., Aguiar, M.S., 2019. Mineral Classification Using Machine Learning and Images of Microscopic Rock Thin Section, in: Martínez-Villaseñor, L., Batyrshin, I., Marín-Hernández, A. (Eds.), Advances in Soft Computing. Springer, pp. 63–76. https://doi.org/10.1007/978-3-030-33749-0_6

Brandmeier, M., Cabrera Zamora, I.G., Nykänen, V., Middleton, M., 2020. Boosting for Mineral Prospectivity Modeling: A New GIS Toolbox. Nat. Resour. Res. 29, 71–88. https://doi.org/10.1007/s11053-019-09483-8

Breiman, L., 2001. Random forests. Mach. Learn. 56, 5–32.

Carranza, E.J.M., Laborte, A.G., 2016. Data-Driven Predictive Modeling of Mineral Prospectivity Using Random Forests: A Case Study in Catanduanes Island (Philippines). Nat. Resour. Res. 25, 35–50. https://doi.org/10.1007/s11053-015-9268-x

Carranza, E.J.M., Laborte, A.G., 2015. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm. Ore Geol. Rev. 71, 777–787. https://doi.org/10.1016/j.oregeorev.2014.08.010

Chawla, N. V, Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953

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

Cracknell, M., Reading, A., W. McNeill, A., 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. https://doi.org/10.1080/08120099.2014.858081

Cunha, L.M., Cabral Neto, I., Silveira, F.V., Nannini, F., 2017. Apresentação dos resultados do Projeto Diamante brasil, in: Fomentando o Setor Mineral Brasileiro. Ministério de Minas e Energia, Brasília, DF, p. 25.

Deer, W.A., Howie, R.A., Zussman, J., 2013. An Introduction to the Rock-Forming Minerals, 3rd ed. Mineralogical Society of Great Britain and Ireland, London. https://doi.org/10.1180/DHZ

Dramsch, J.S., 2020. 70 years of machine learning in geoscience in review, in: Advances in Geophysics. pp. 1–55. https://doi.org/10.1016/bs.agph.2020.08.002

Ester, M., Kriegel, H.-P., Sander, J., Xiaowei, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in: International Conference on Knowledge Discovery and Data Mining. KDD-96, Portland, Oregon, pp. 226–231. https://doi.org/10.1016/B978-044452701-1.00067-3

Ford, A., 2019. Practical Implementation of Random Forest-Based Mineral Potential Mapping for Porphyry Cu – Au Mineralization in the Eastern Lachlan Orogen , NSW , Australia. Nat. Resour. Res. https://doi.org/10.1007/s11053-019-09598-y

Gavish, Y., O'Connell, J., Marsh, C.J., Tarantino, C., Blonda, P., Tomaselli, V., Kunin, W.E., 2018. Comparing the performance of flat and hierarchical Habitat/Land-Cover classification models in a NATURA 2000

site. ISPRS J. Photogramm. Remote Sens. 136, 1–12. https://doi.org/10.1016/j.isprsjprs.2017.12.002

Grinberg, M., 2018. Flask web development: developing web applications witg python.

Hariharan, S., Tirodkar, S., Porwal, A., Bhattacharya, A., Joly, A., 2017. Random Forest-Based Prospectivity Modelling of Greenfield Terrains Using Sparse Deposit Data: An Example from the Tanami Region, Western Australia. Nat. Resour. Res. 26. https://doi.org/10.1007/s11053-017-9335-6

Harris, J.R., Grunsky, E., Behnia, P., Corrigan, D., 2015. Data- and knowledge-driven mineral prospectivity maps for Canada's North. Ore Geol. Rev. 71, 788–803. https://doi.org/10.1016/j.oregeorev.2015.01.004

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study1. Intell. Data Anal. 6, 429–449. https://doi.org/10.3233/IDA-2002-6504

Koch, P.H., Lund, C., Rosenkranz, J., 2019. Automated drill core mineralogical characterization method for texture classification and modal mineralogy estimation for geometallurgy. Miner. Eng. 136, 99–109. https://doi.org/10.1016/j.mineng.2019.03.008

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

Kuhn, S., Cracknell, M.J., Reading, A.M., Sykora, S., 2020. Identification of intrusive lithologies in volcanic terrains in British Columbia by machine learning using random forests: The value of using a soft classifier. GEOPHYSICS 85, B249–B258. https://doi.org/10.1190/geo2019-0461.1

Li, X., Zhang, C., Behrens, H., Holtz, F., 2020. Lithos Calculating amphibole formula from electron microprobe analysis data using a machine learning method based on principal components regression. LITHOS 362–363, 105469. https://doi.org/10.1016/j.lithos.2020.105469

Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2003. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Math. Geol. 35, 253–278. https://doi.org/10.1023/A:1023866030544

McKay, G., Harris, J.R., 2016. Comparison of the Data-Driven Random Forests Model and a Knowledge-Driven Method for Mineral Prospectivity Mapping: A Case Study for Gold Deposits Around the Huritz Group and Nueltin Suite, Nunavut, Canada. Nat. Resour. Res. 25, 125–143. https://doi.org/10.1007/s11053-015-9274-z

Misra, S., Osogba, O., Powers, M., 2020. Unsupervised outlier detection techniques for well logs and geophysical data, Machine Learning for Subsurface Characterization. Elsevier Inc. https://doi.org/10.1016/b978-0-12-817736-5.00001-6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Prado, E.M.G., de 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 Geol. Rev. 124, 103611. https://doi.org/10.1016/j.oregeorev.2020.103611

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. https://doi.org/10.1109/JSTARS.2018.2855207

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

Rubo, R.A., de Carvalho Carneiro, C., Michelon, M.F., Gioria, R. dos S., 2019. Digital petrography: Mineralogy and porosity identification using machine learning algorithms in petrographic thin section images. J. Pet. Sci. Eng. 183. https://doi.org/10.1016/j.petrol.2019.106382

Sarbas, 2021. GEOROC - Geochemistry of Rocks if the Oceans and Cotinents [WWW Document]. URL http://georoc.mpch-mainz.gwdg.de/georoc/

Schramm, B., Jochum, K.P., Sarbas, B., Nohl, U., 2006. GEOROC and GeoReM—Linking the information of two Geological databases. Geochim. Cosmochim. Acta 70, A565. https://doi.org/10.1016/j.gca.2006.06.1045

Schroeder, M., Cornford, D., Farrimond, P., Cornford, C., 2008. Addressing missing data in geochemistry: A non-linear approach. Org. Geochem. 39, 1162–1169. https://doi.org/10.1016/j.orggeochem.2008.02.016

Schumacker, R., Tomek, S., 2013. Understanding Statistics Using R. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-6227-9

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Smiti, A., 2020. A critical overview of outlier detection methods. Comput. Sci. Rev. 38, 100306. https://doi.org/10.1016/j.cosrev.2020.100306

Vijayvargiya, A., Prakash, C., Kumar, R., Bansal, S., João, J.M., 2021. Human knee abnormality detection from imbalanced sEMG data. Biomed. Signal Process. Control 66. https://doi.org/10.1016/j.bspc.2021.102406

Zhang, Shuai, Carranza, E.J.M., Wei, H., Xiao, K., Yang, F., Xiang, J., Zhang, Shihong, Xu, Y., 2021. Data-driven Mineral Prospectivity Mapping by Joint Application of Unsupervised Convolutional Auto-encoder Network and Supervised Convolutional Neural Network. Nat. Resour. Res. 30, 1011–1031. https://doi.org/10.1007/s11053-020-09789-y

Figures

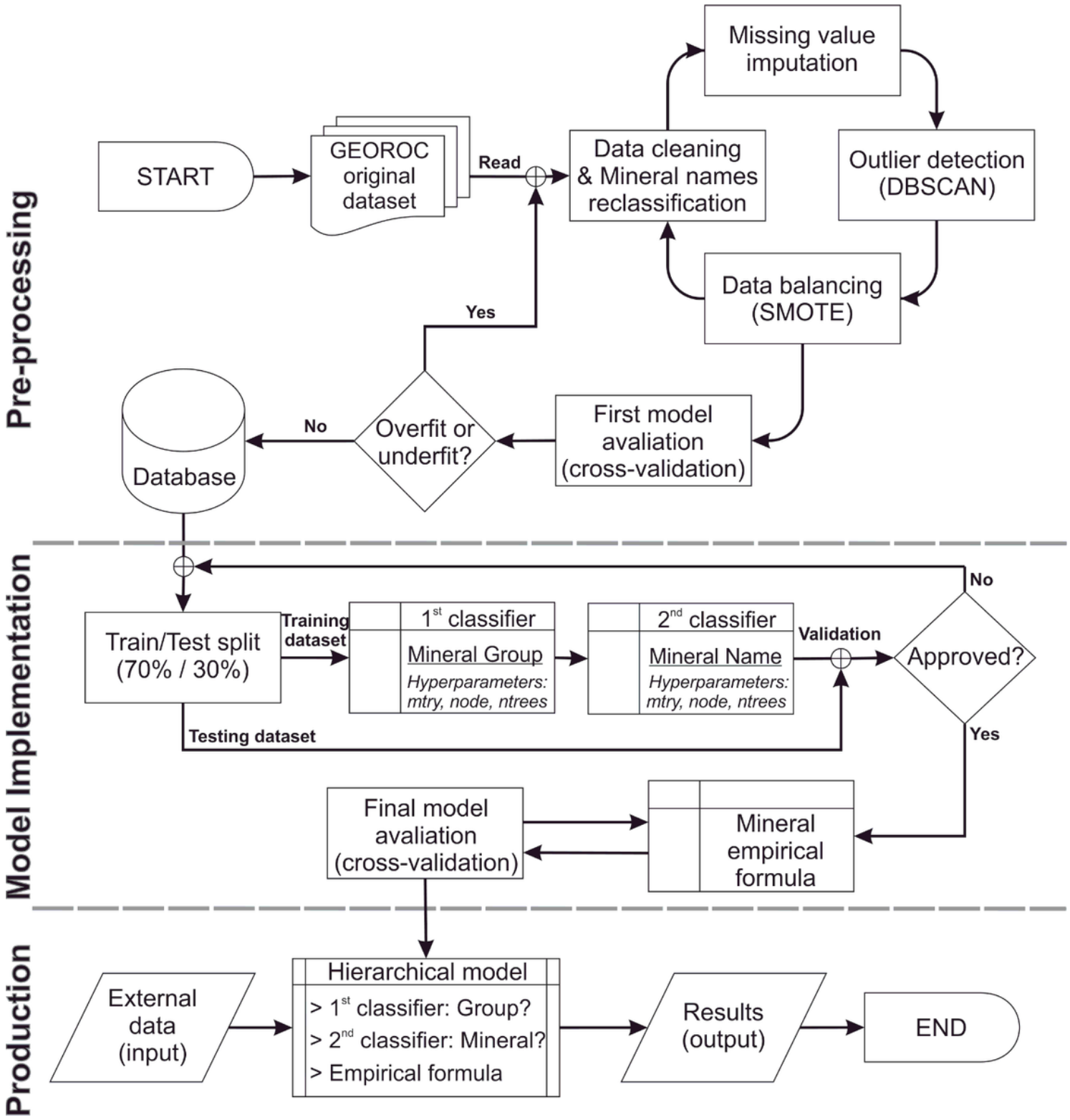

Figure 1

Qmin development and utilization flowchart with the three main stages: Pre-processing, Model Implementation, and Production. Each of these stages is related to chained processes that end with a quality evaluation and in the construction of the basis for the subsequent stages. The Application stage ends with the output results of external entry data.

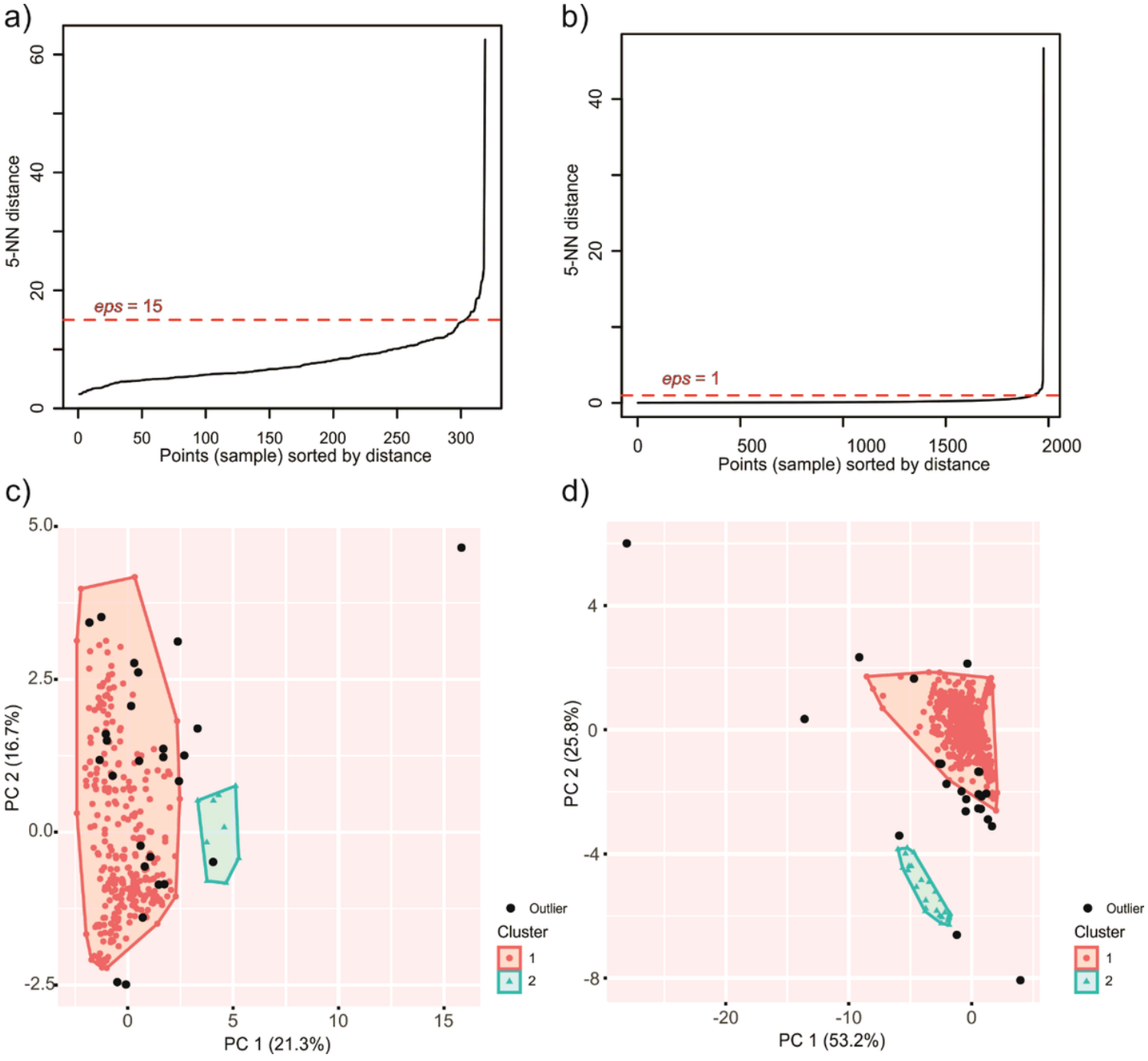

Figure 2

Some examples of the results obtained from running the DBSCAN algorithm. Distance plot of samples ordered according to the k parameter (in both cases k is set to 5 nearest neighbors) and the optimum eps

value (horizontal dashed line) for a) Clay minerals and b) Zircon. Principal Component (PC) plot of the DBSCAN outlier detection and filtering for c) Clay Minerals and d) Zircon.

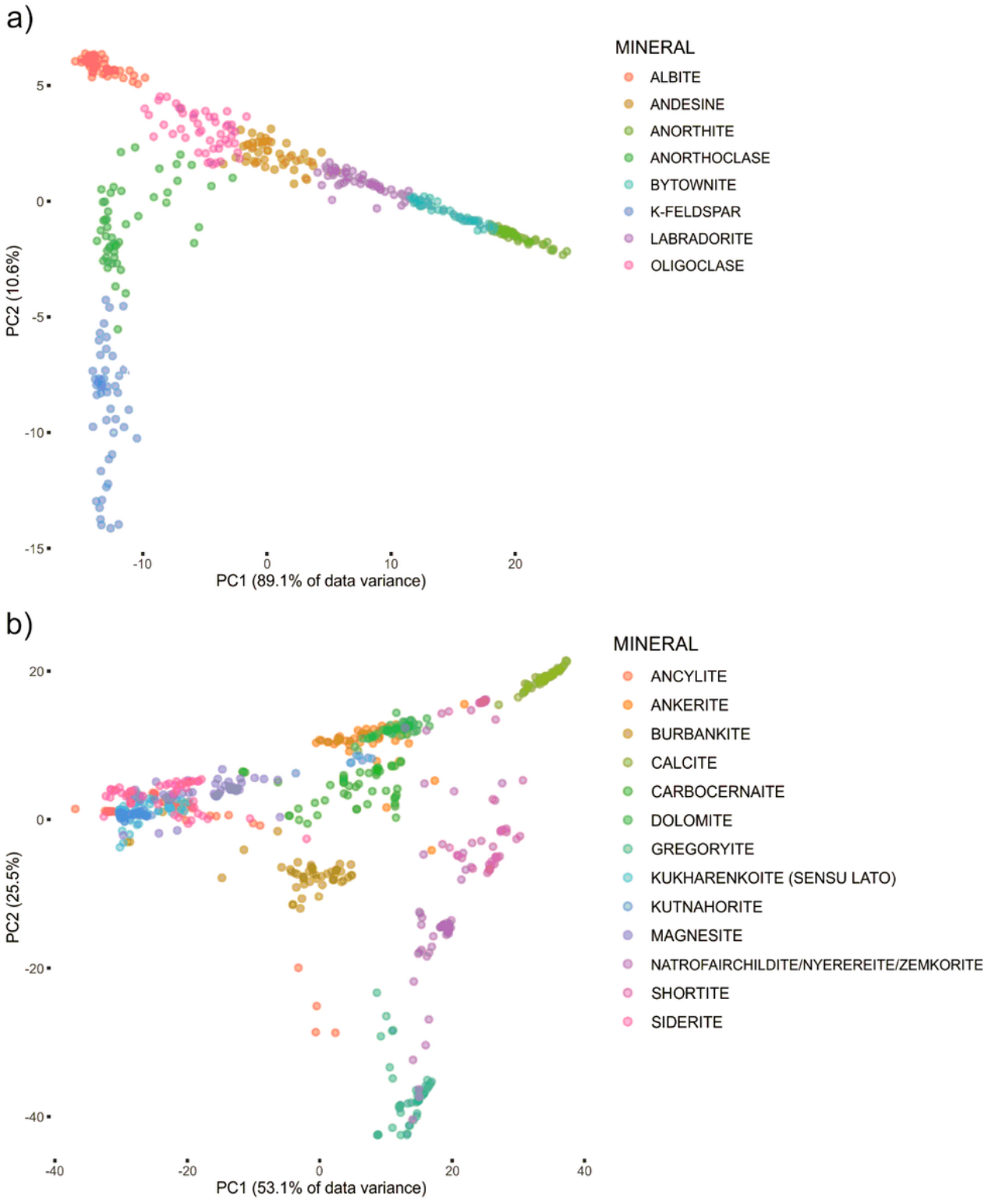

Figure 3

Examples of Principal Component (PC) plot of post-SMOTE balanced samples (i.e., 50 instances per mineral) used for training the classifiers. a) Feldspar group samples and b) Carbonate group samples.

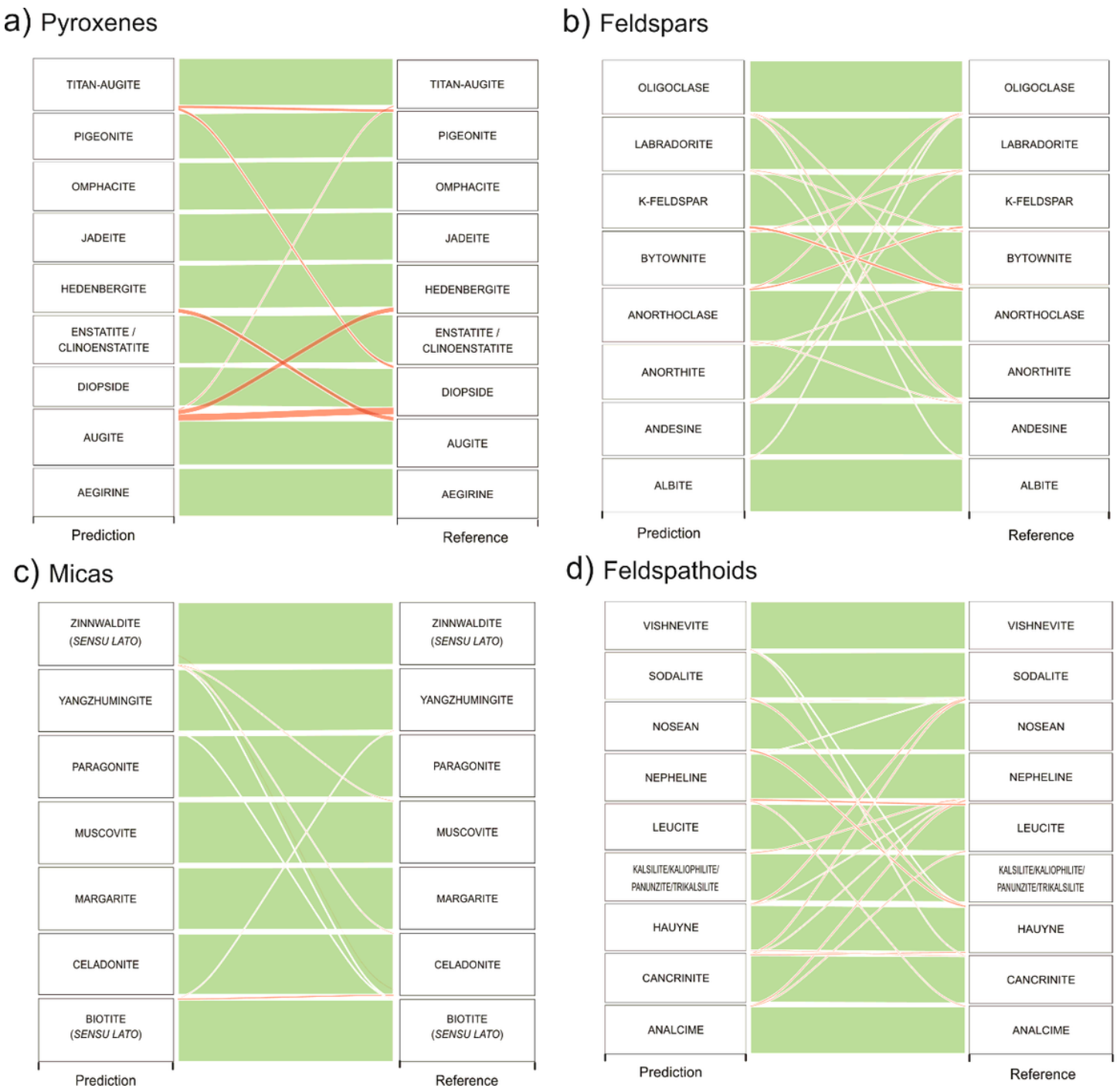

Figure 4

Alluvial diagrams for a) Pyroxenes, b) Feldspars, c) Micas, and d) Feldspathoids showing the relation between prediction and reference classes. Good predictions are represented by green links, while bad (i.e., wrong) predictions are represented as reddish links (e.g., despite most samples being classified correctly, some augite samples were classified as diopside or hedenbergite). The link's width represents the actual proportion of the data sample.

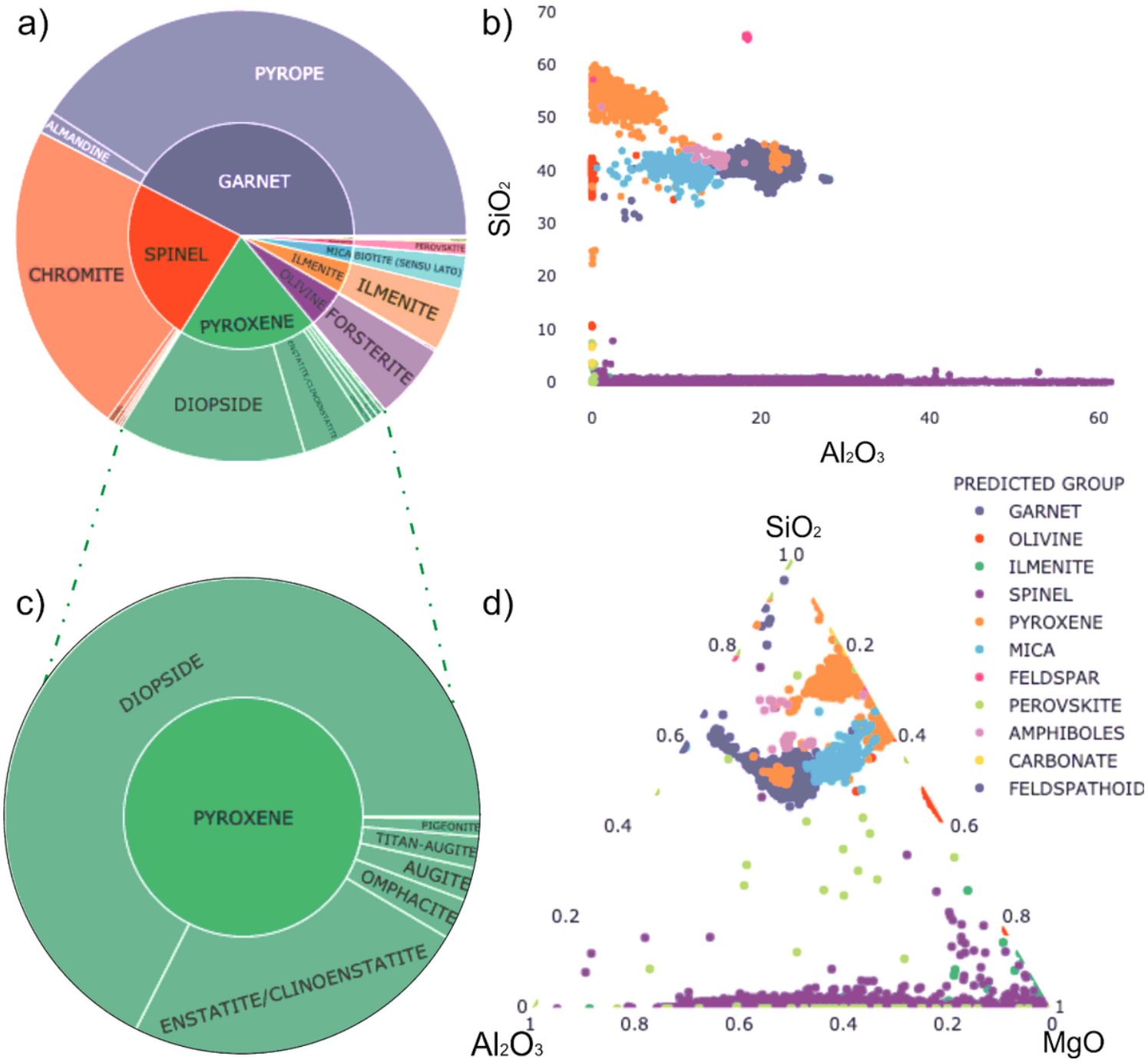

Figure 5

Exploratory graphical analysis of the Diamante Brasil Project data. a) Proportions of the predicted groups and minerals in the blind test; b) Bivariate analysis of SiO2 (%) vs. Al2O3; c) Proportion of predicted minerals in the Pyroxene Group; d) Triplot of SiO2 (%) - Al2O3 (%) - MgO (%). All colors are according to the predicted group, as shown in the key legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- QminGEOROCMINERALDATABASEUPDATE.xlsx
- QminDiamanteBrasilProjectdataoutput.xlsx